

Statistics for Social and Behavioral Sciences

Maria Kateri
Irina Moustaki *Editors*

Trends and Challenges in Categorical Data Analysis

Statistical Modelling and Interpretation

 Springer

Statistics for Social and Behavioral Sciences

Statistics for Social and Behavioral Sciences (SSBS) includes monographs and advanced textbooks relating to education, psychology, sociology, political science, public policy, and law.

Maria Kateri • Iriini Moustaki
Editors

Trends and Challenges in Categorical Data Analysis

Statistical Modelling and Interpretation

 Springer

Editors

Maria Kateri
Department of Mathematics
RWTH Aachen University
Aachen, Germany

Irini Moustaki
Department of Statistics
London School of Econ. & Polit. Science
London, UK

ISSN 2199-7357

ISSN 2199-7365 (electronic)

Statistics for Social and Behavioral Sciences

ISBN 978-3-031-31185-7

ISBN 978-3-031-31186-4 (eBook)

<https://doi.org/10.1007/978-3-031-31186-4>

Mathematics Subject Classification: 62H, 62J

© Springer Nature Switzerland AG 2023

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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

Preface

The analysis of categorical data has led to the development of a whole new set of methods, tools, and theory. The methodological developments have been followed by commercial and open source software, which has facilitated the spread and use of the methods in many substantive areas of application. The book aims to bring together and provide a comprehensive review of a selected list of topics connected to recent advances in statistical modelling and interpretation of categorical data. The focus is on cross-sectional as well as time-dependent data.

We consider research questions of both symmetrical and regression-type nature, such as studying and modelling the association of a number of categorical variables, as well as regression-type analysis of a categorical response variable explained by a number of observed covariates.

Categorical data predominate in social surveys and their analysis, from descriptive and exploratory to statistical modelling, require special treatments that take into account the nature and information included in these data. Traditionally, categorical data analysis (CDA) methodology has focused on two- and three-way contingency tables, while for higher dimensional tables, it is usually commented that they are analysed analogously. Many of today's real applications involve high-dimensional and complex data with many more than three variables. Categorical data methods have been extended to handle multivariate data of higher dimensions, addressing issues of sparseness, model estimation, fit, and model selection. More specifically, binary and ordinal response models have become the focus of attention in areas of supervised machine learning. Graphical models and networks involving categorical data have applications, in social sciences, biology, and natural language processing, among others. Developments and problems in data science necessitate special treatment for different types of categorical data and impose new challenges on CDA.

To tackle problems in contemporary applications of categorical data, a thoughtful revisiting of traditional methods of CDA is required.

Serving this goal, the current volume covers nine distinct topics, underlining, when necessary, their inter-relationships and helping the reader to place methods and tools for categorical data into a general framework. It reviews association models for multi-way contingency tables and their connection to item response

theory models and graphical models, marginal models, regression type models with categorical responses, and/or categorical covariates including simple measures of interpretation, time series models for count and binary data, models for binary panel data, as well as methodology for bias correction and Bayesian inference.

The volume is intended for statisticians, data scientists, graduate students of statistics, but also computer scientists or researchers with a strong interest in methods and tools used for the analysis of categorical data. The chapters include applications from economics, education, psychiatry, medicine, and finance, but the applicability of the methods discussed go beyond those areas.

The volume is organised into three parts. Part I (Chaps. 1–4) focuses on modelling multivariate (multiple response variables) categorical data through their joint and marginal distributions. Chapter 1 reviews classical association models and establishes the connection with item response theory models and graphical models that provide multiple insights into the data problem. A computationally feasible composite likelihood estimation method and testing framework are proposed. Real data examples from massively open online courses (MOOC) and from the Depression, Anxiety and Stress Scale (DASS) are included, as well as information on the R packages `logmulti` and `pleLMA`. Graphical models are discussed in more detail in Chap. 2, which covers undirected graphical log-linear models, directed graphical models, and graphical chain models for modelling complex multivariate associations. The infant survival data, presented in other seminal books on categorical data, are used to illustrate the various graphical models. Graphical models already covered in Chaps. 1 and 2 are shown to be connected to the class of marginal models presented in Chap. 3. In this chapter, a thorough overview of marginal models is provided. Marginal models are helpful for testing hypotheses about relations among correlated categorical marginal distributions. The content of this chapter is motivated with examples from repeated measurements/panel data, missing data, and graphical data in which marginal distributions of higher-dimensional joint distributions play an important role. Potential estimation methods are thoroughly discussed. Information on three available R packages (`cmm`, `mph.fit`, and `hmmm`) for marginal modelling is provided. The chapter concludes with a list of further theoretical and methodological developments in the area of marginal modelling and extensions for the future. Chapter 4 offers a Bayesian treatment of multivariate categorical data with emphasis on estimation, choices of priors, and model selection. The explored tools are applied to two-way contingency tables from three medical areas of research, namely risk for coronary heart disease, lymphoma and chemotherapy, and toxemia in pregnancy.

Part II (Chaps. 5–7) focuses on regression type models for binary and ordinal responses. Chapter 5 proposes probability-based effect measures that provide a simpler interpretation of regression coefficients of logistic and probit models with linear and non-linear predictors, which are missing from the traditional literature on binary and ordinal regression. The proposed measures are used to compute effective measures for a class of generalised linear models with logit, log, and identity link functions, fitted to data from an Italian survey on employment status and a generalised additive model fitted to the horseshoe crab data. R code is provided

for replicating the analysis. Chapter 6 proposes mean and median bias reduction in adjacent-categories logit models with proportional odds and mean bias reduction in models with non-proportional odds. The methodology is illustrated using real examples, and the R code is provided to replicate all the numerical and graphical results. Chapter 7 gives an overview of regularised estimation methods for generalised additive models with ordinal covariates, considering predictor selection and merging of predictor categories with the effect of reducing the number of parameters and easing interpretability. The proposed method is compared to existing classical methods and is applied to a real data set from the International Classification of Functioning, Disability and Health study on chronic widespread pain. Information on R packages that perform the different types of analysis discussed in the chapter is provided.

Part III (Chaps. 8 and 9) discusses models for discrete time-dependent data. Chapter 8 presents an overview of a unified framework of ARMA-type models widely used for continuous time series for binary and count data, with emphasis on associated stochastic properties and likelihood-based inferential tools. The methodology is applied to two real data sets: the daily number of deaths from COVID-19 in Italy, for which a Poisson and a negative binomial distribution is assumed for the data; and a binary series of log-returns for the weekly closing prices of Johnson & Johnson. The code for replicating the analysis in the chapter is provided. Finally, Chap. 9 reviews the formulation and estimation of fixed-effects type models for binary panel data. In particular, the chapter reviews and illustrates, through an extensive simulation study, estimation methods for dealing with the inconsistency of the maximum likelihood estimator due to incidental parameters, embedding in a unified framework the target-corrected and conditional maximum likelihood estimators, including a pseudo conditional maximum likelihood estimator. The methodology is applied to data on female labour force participation from the US Panel Study of Income Dynamics. The chapter also includes a review of packages available to estimate the models discussed.

Each chapter makes its own methodological and distinct contribution to the modelling of categorical data and can be read independently. In some cases, connections are made among the topics covered in the edited volume, but these connections or overlaps do not imply that the reader needs to read the chapters in any particular order. The division of the book in three parts is also indicative and does not provide a strict separation of the contributions.

The seed for this volume was sown during the workshop “Challenges for Categorical Data Analysis” (CCDA2018) held in Aachen in 2018. We would like to thank all the participants of this workshop for the inspiring discussions and for motivating our book project. We specially thank Eva Hiripi, Senior Editor at Springer, for her continuous support and guidance in the process of preparing the volume.

Finally, we are grateful to the friends and colleagues who contributed chapters to this volume. Without their engagement and impressive work, this project would not have been possible.

Aachen, Germany
London, UK
February, 2023

Maria Kateri
Irina Moustaki

Contents

1 Log-Linear and Log-Multiplicative Association Models for Categorical Data	1
Carolyn J. Anderson, Maria Kateri, and Irini Moustaki	
2 Graphical Models for Categorical Data	43
Peter W. F. Smith	
3 Marginal Models: An Overview	67
Tamás Rudas and Wicher Bergsma	
4 Bayesian Inference for Multivariate Categorical Data	117
Jonathan J. Forster and Mark E. Grigsby	
5 Simple Ways to Interpret Effects in Modeling Binary Data	155
Alan Agresti, Claudia Tarantola, and Roberta Varriale	
6 Mean and Median Bias Reduction: A Concise Review and Application to Adjacent-Categories Logit Models	177
Ioannis Kosmidis	
7 Regularization and Predictor Selection for Ordinal and Categorical Data	199
Jan Gertheiss and Gerhard Tutz	
8 An Overview of ARMA-Like Models for Count and Binary Data	233
Mirko Armillotta, Alessandra Luati, and Monia Lupporelli	
9 Advances in Maximum Likelihood Estimation of Fixed-Effects Binary Panel Data Models	275
Francesco Valentini, Claudia Pigini, and Francesco Bartolucci	

Contributors

Alan Agresti Department of Statistics, University of Florida, Gainesville, FL, USA

Carolyn J. Anderson Department of Educational Psychology, College of Education, University of Illinois at Urbana-Champaign, Champaign, IL, USA

Mirko Armillotta Department of Statistical Sciences, University of Bologna, Bologna, Italy

Francesco Bartolucci Department of Economics, University of Perugia, Perugia, Italy

Wicher Bergsma Department of Statistics, London School of Economics and Political Science, London, UK

Jonathan J. Forster Department of Statistics, University of Warwick, Coventry, UK

Jan Gertheiss Helmut Schmidt University, Hamburg, Germany

Mark E. Grigsby Proctor and Gamble, The Heights Weybridge, Surrey, UK

Maria Kateri Institute of Statistics, RWTH Aachen University, Aachen, Germany

Ioannis Kosmidis Department of Statistics, University of Warwick, Coventry, UK

Alessandra Luati Department of Statistical Sciences, University of Bologna, Bologna, Italy

Monia Lupparelli Department of Statistics, Computer Science, Applications, University of Florence, Florence, Italy

Irini Moustaki Department of Statistics, London School of Economics and Political Science, London, UK

Claudia Pigini Department of Economics and Social Sciences, Marche Polytechnic University, Ancona, Italy

Tamás Rudas Department of Statistics, Faculty of Social Sciences, Eötvös Loránd University, Budapest, Hungary

Peter W. F. Smith Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, UK

Claudia Tarantola Department of Economics and Management, University of Pavia, Pavia, Italy

Gerhard Tutz Ludwig Maximilians University, Munich, Germany

Francesco Valentini Department of Economics and Social Sciences, Marche Polytechnic University, Ancona, Italy

Roberta Varriale Istat, Rome, Italy

Chapter 1

Log-Linear and Log-Multiplicative Association Models for Categorical Data



Carolyn J. Anderson, Maria Kateri, and Irimi Moustaki

1.1 Introduction

Log-linear models are useful for determining whether dependencies exist between categorical variables; however, when there are interactions, the nature of the association needs to be described. Unfortunately, the descriptions can be challenging especially when the categorical variables have a large number of categories and/or the table is high-dimensional. To fully capture the dependency structure would require computing all possible conditional odds ratios (ORs), which in the case of large tables is often not very enlightening. Association models (AMs) provide a solution to this problem by imposing special structures on the interactions between categorical variables thus leading to more parsimonious models that facilitate insightful interpretation of interactions. A central characteristic of all AMs is that interactions are represented by multiplicative terms.

Basically, AMs have a special multiplicative structure imposed on some or all interaction terms of a standard log-linear model. The parameters of the multiplicative terms have high interpretative value and reduce the number of parameters needed to describe the nature and strength of interactions. In some AMs, the

C. J. Anderson (✉)

Department of Educational Psychology, College of Education, University of Illinois at Urbana-Champaign, Champaign, IL, USA
e-mail: cja@illinois.edu

M. Kateri

Institute of Statistics, RWTH Aachen University, Aachen, Germany
e-mail: maria.kateri@rwth-aachen.de

I. Moustaki

Department of Statistics, London School of Economics and Political Science, London, UK
e-mail: i.moustaki@lse.ac.uk

© Springer Nature Switzerland AG 2023

M. Kateri, I. Moustaki (eds.), *Trends and Challenges in Categorical Data Analysis*,
Statistics for Social and Behavioral Sciences,
https://doi.org/10.1007/978-3-031-31186-4_1

model remains log-linear, while others are log-multiplicative, i.e. non-linear in their parameters. ORs, which play a predominant role in log-linear model (and AMs) analysis and interpretation, are functions of the parameters introduced in an AM, and plots of these parameters give pictures of the features and structure of the associations.

In addition to providing visual plots representing associations between variables, the models themselves have graphical representations. The graphics greatly aid in communication because they represent scientific content and in some cases underlying processes. To differentiate between models and for clarity, we advocate that models should be presented both graphically and algebraically. Many of the AMs that we discuss have the same basic graphical representation. The algebraic representations, when used without their graphical representations, tend to cloud the relationships between models, but the algebraic form provides details that may be lacking in the graphical representation.

AMs for the analysis of categorical variables have been derived from numerous frameworks. They provide useful structural representations of interactions among variables allowing a special treatment for ordinal variables. AMs have been developed either directly for specific modeling purposes (e.g. contingency table analysis) or have been arisen through a theorized underlying process (e.g. item response theory (IRT)). They have been proposed over different fields and sub-fields, often independently, which has led to a fractured literature on the subject. It is evident that AMs offer a powerful and flexible platform for diverse areas of applications. The class of models that we generically refer to as AMs consists of many models with different names but of the same general form. These include, among others, linear by linear models (LL), row models (R), column models (C), uniform models (U), and M -dimensional row-column AMs ($RC(M)$) [30, 32, 33], generalized additive effects and multiplicative interaction models used to study plant genetics [23], graphical latent variable models for categorical data [4], IRT models [5, 6, 38, 52], Ising model [47], generalized Newton's law of gravity [18], network psychometrics [52], fused graphical models [16], formative response models, distance-based models [18–20, 61], conditional multinomial models [3, 5, 37], and discretized multivariate normal distributions [9, 31, 60, 67, 68]. It is worth mentioning that there are efforts to build bridges between different fields and further explore their utility, such as connecting IRT to log-linear models [44, 45], and to log-multiplicative interactions [5, 6, 52], and others.

AMs are closely linked to log-linear models. For this reason, we start Sect. 1.2 with a brief presentation of log-linear models upon which we build the family of AMs for two-way tables (i.e., LL , R , C , and $RC(M)$ models). Many of the basic features of these AMs for two-way tables carry over to models for more variables and more complex situations. We subsequently review statistical graphical representations of log-linear and AMs and use this as a step toward high-dimensional generalizations of the $RC(M)$ model. Subsequently, we present high-dimensional models in detail, including estimation and the equivalence with IRT models. Some discussion on testing and model selection under the pseudo-likelihood framework is given. To illustrate the use and benefits afforded by AMs,

two examples are given: (i) the analysis of a (16×6) table by models for two-way tables, and (ii) responses to 42 four category items from three correlated scales by models for high-dimensional tables. Lastly, we follow with a discussion that reflects on the material presented in the chapter and provides future research directions.

1.2 Preliminaries

Throughout this chapter, we assume that we have I items (or variables), $\mathbf{Y} = (Y_1, \dots, Y_I)'$, measured on n subjects. Let $\mathbf{Y} = (Y_1, \dots, Y_I)'$ be a random response vector where $Y_i \in \mathcal{C}_i = \{1, \dots, J_i\}$, and let $\mathbf{y}_s = \{y_{1s}, \dots, y_{Is}\}$ be observed responses for subject $s \in \{1, \dots, n\}$, i.e. $y_{is} = j_i \in \mathcal{C}_i$, for $i = 1, \dots, I$. Furthermore, assume that there exists a set of M latent variables, $\Theta = \{\theta_1, \dots, \theta_M\}$ and $\theta = (\theta_1, \dots, \theta_M)'$ is a realization of them. We restrict to models with $M \leq I$ while more complex models with $M > I$ are possible.

In a contingency table representation, data form an I -dimensional table, produced by cross-classifying the subjects' responses on all items, having cell entries n_{j_1, \dots, j_I} , the frequencies of subjects with responses $\mathbf{y} = (j_1, \dots, j_I)'$, where $j_i \in \mathcal{C}_i$, $i = 1, \dots, I$. In this setup, the subject index s is suppressed, but will be needed later in the chapter. Obviously, $\sum_{j_1, \dots, j_I} n_{j_1, \dots, j_I} = n$ and the underlying distribution, depending on the study design, can be a multinomial $\mathcal{M}(n, \boldsymbol{\pi})$ with probability table $\boldsymbol{\pi} = \{\pi_{\mathbf{y}}\} = \{\pi_{j_1, \dots, j_I}\}$, or independent Poisson distributions $\mathcal{P}(m_{\mathbf{y}})$ in every cell, where $m_{\mathbf{y}}$ is the predicted or expected cell frequency. Given the sample size n , the expected cell frequencies equal $m_{\mathbf{y}} = m_{j_1, \dots, j_I} = n\pi_{j_1, \dots, j_I}$.

1.2.1 Hierarchical Log-linear Models

Contingency tables are traditionally analyzed by hierarchical log-linear models, expressed in terms of cell probabilities or expected cell frequencies.¹ Here we shall model the cell probabilities. In the case of many items, the corresponding contingency table is high-dimensional and is often extremely sparse, which causes inferential and estimation problems. Usually lower order interactions (even only two factor interactions) are sufficient to model the response patterns and the corresponding marginal tables are not sparse. The two-way marginal tables are sufficient statistics for estimating two-factor interactions. Thus, the response probabilities

¹ Note that "hierarchical" refers to different models (e.g., linear regression, multi-level models, log-linear models) In this chapter "hierarchical" refers to models where all lower order terms that comprise an interaction are included in the model.

$P(\mathbf{y})$ can be modeled, for example, by a log-linear model with all two-factor interactions,

$$\log P(\mathbf{y}) = \log(\pi_{\mathbf{y}}) = \lambda + \sum_{i=1}^I \lambda_{j_i}^{[i]} + \sum_{\substack{i,k \\ i < k}}^I \lambda_{j_i j_k}^{[ik]}, \quad j_i \in \mathcal{C}_i, \quad j_k \in \mathcal{C}_k, \quad (1.1)$$

where λ ensures that probabilities sum to 1, $\lambda_{j_i}^{[i]}$ is the marginal (main) effect term for the category j_i of the i -th item, and $\lambda_{j_i j_k}^{[ik]}$ is the interaction term between the levels j_i and j_k of the i -th and k -th items, respectively.

Identification constraints are required on parameters in (1.1) to obtain parameter estimates. Common constraints are setting the first category to zero, i.e.,

$$\lambda_1^{[i]} = \lambda_{11}^{[ik]} = \lambda_{1j_k}^{[ik]} = \lambda_{j_i 1}^{[ik]} = 0, \quad \text{for all possible values of } i, k, j_i, j_k. \quad (1.2)$$

Alternative constraints set the last category to zero or set the sum over categories equal to zero.

In some applications, we are only interested in the relationship between variables; however, in other modeling applications, we make a distinction between response and explanatory variables. Regardless of the situation, the model for tables is the same. For example, when modeling response behavior, explanatory variables, such as demographic ones, may be present that may be categorical or on an interval scale. In such cases, log-linear models of type (1.1) can be employed that incorporate the main effects for the explanatory variables and interactions between explanatory variables and that response variable.

The simplest case of having just two items reduces (1.1) to

$$\log P(\mathbf{y}) = \log(\pi_{\mathbf{y}}) = \lambda + \lambda_{j_1}^{[1]} + \lambda_{j_2}^{[2]} + \lambda_{j_1 j_2}^{[12]}, \quad j_1 \in \mathcal{C}_1, \quad j_2 \in \mathcal{C}_2. \quad (1.3)$$

In the log-linear modeling framework, log-linear models with interactions may have difficulty dealing with sparse tables that include zero cell frequencies. Using (1.1) as an example, if a cell (j_i, j_k) of the $[ik]$ marginal table has a zero frequency, the corresponding parameter $\lambda_{j_i j_k}^{[ik]}$ for that cell cannot be estimated, since this zero marginal cell corresponds to its sufficient statistic. Necessary and sufficient conditions for the existence of the maximum likelihood estimates (MLE) of the log-linear model parameters, with a focus on the role of sampling zeros in the observed table, are provided by Fienberg and Rinaldo [26]. Fitted values (MLE) for (1.1) can be obtained using iterative proportional fitting, but we cannot fully describe the interaction because some local odds ratios are not estimable. This is not the case for unsaturated AMs. For example, the $RC(M)$ model described in the next section encounters no problems if there are sampling zeros in a (marginal) table. Only the univariate marginals need to be non-zero. The ability of AMs to deal with sparse tables becomes especially important when we have high-dimensional tables.

1.3 Association Models for Two-Way Tables

Model (1.3) is saturated (i.e. has 0 degrees of freedom). For a $J_1 \times J_2$ table, in the classical log-linear modeling framework, there are no models in between the saturated model and that of independence, which has $(J_1 - 1)(J_2 - 1)$ degrees of freedom. A class of non-saturated models is derived by imposing a structure or restrictions on the interaction parameters of a log-linear model which requires fewer parameters. Fewer parameters leads to more parsimonious models that fill the gap between the two extreme models (independence and saturated) and at the same time, offer sound interpretation. These models are known as dependency models or AMs (often called Goodman's AMs) and are based on the concept of assigning scores or estimating scale values for the categories of the classification variables (items).

For a two-dimensional table, association models are of the form

$$\log P(\mathbf{y}) = \log(\pi_{\mathbf{y}}) = \lambda + \lambda_{j_1}^{[1]} + \lambda_{j_2}^{[2]} + \sigma^2 v_{1j_1} v_{2j_2}, \quad (1.4)$$

for $j_1 \in \mathcal{C}_1$, $j_2 \in \mathcal{C}_2$, where $\mathbf{v}_1 = (v_{11}, \dots, v_{1J_1})'$ and $\mathbf{v}_2 = (v_{21}, \dots, v_{2J_2})'$ are scores corresponding to the rows and columns of the contingency table, respectively, and σ^2 is an intrinsic association parameter. Notice that in the literature on association models, the association parameter is usually denoted by ϕ and, for row and column scores that are monotone in the same direction, the sign of ϕ indicates the direction of the underlying association. The model is invariant under linear transformation of the row and column scores and the direction of the scores are generally set such that σ^2 is positive. An important point is that σ^2 reflects the strength of the association and the row and column scores reflect the structure.

The row and column scores, \mathbf{v}_1 and \mathbf{v}_2 , respectively, can be fixed (known) or parameters to be estimated. Typically, the scores of a nominal variable are parameters while those of an ordinal can be fixed or parameters, depending on whether the distances between successive categories are known or not. The simplest association model that considers both of them fixed, has just one parameter more than the independence model and is known as the *linear by linear (LL)* model. If additionally the scores are equidistant for successive row and column categories, then under this specific *LL* model all local odds ratios, which are odds ratios between adjacent rows and columns, are equal. This is called as the *uniform (U)* association model. When the row scores are fixed and column scores are estimated, the model is called the column effect (*C*) model. The row effect (*R*) model is defined analogously. Models *LL*, *U*, *C* and *R* are all log-linear. When both row and column scores are parameters to be estimated, Model (1.4) becomes the *multiplicative row-column effect (RC)* model and no longer has a log-linear structure.

The main effects parameters of Model (1.4) satisfy the corresponding identifiability constraints in (1.2) while the scores, whenever they are parameters, satisfy

$$\sum_{j_1=1}^{J_1} v_{1j_1} = \sum_{j_2=1}^{J_2} v_{2j_2} = 0 \quad \text{and} \quad \sum_{j_1=1}^{J_1} v_{1j_1}^2 = \sum_{j_2=1}^{J_2} v_{2j_2}^2 = 1. \quad (1.5)$$

Since Model (1.4) is invariant under linear transformations of the scores, for comparability, and also in the case of fixed or known scores, scores are transformed to fulfill (1.5). The intrinsic association parameter in (1.4) is redundant and can be set $\sigma^2 = 1$, abandoning the second set of constraints in (1.5), as given by Goodman [30].

An extension of the RC model is the multidimensional row-column or $RC(M)$ association model, which includes multiple sets of scores for each item. It is defined as

$$\log P(\mathbf{y}) = \log(\pi_{\mathbf{y}}) = \lambda + \lambda_{j_1}^{[1]} + \lambda_{j_2}^{[2]} + \sum_{m=1}^M \sigma_m^2 v_{1j_1m} v_{2j_2m}, \quad (1.6)$$

for $M \in \{1, \dots, M^*\}$, $M^* = \min(J_1, J_2) - 1$, where scores and association parameters are assigned to each dimension m , with $\sigma_1^2 \geq \dots \geq \sigma_M^2 \geq 0$, reflecting that the strength of association accounted for each dimension m is decreasing in m . Constraints (1.5) hold for the scores on every dimension, and additionally, scores on different dimensions are orthogonal to each other, i.e.,

$$\sum_{j_1=1}^{J_1} v_{1j_1m} v_{1j_1m'} = \sum_{j_2=1}^{J_2} v_{2j_2m} v_{2j_2m'} = 0, \quad \text{for all } m \neq m'. \quad (1.7)$$

Constraints (1.5) and (1.7) are the most commonly used ones, but are not the only possible ones. When scores are treated as parameters, Model (1.6) has $(J_1 - M - 1)(J_2 - M - 1)$ degrees of freedom (df). Note that $RC(1) = RC$ and $RC(M^*)$ is an equivalent expression of the saturated log-linear model given in (1.3).

The AMs for two-way tables presented in this section can be extended in a straightforward manner to tables of higher dimensions and we will point out how the models for high-dimensional tables are the same and different from the $RC(M)$ association models. Before considering the high dimensional case, we discuss estimation and present an example for a 2-way table.

1.3.1 Estimation and Goodness-of-Fit of AMs

Maximum likelihood estimation of AMs is the most commonly used method to fit the models to data and we focus our attention on ways to do this in \mathbf{R} [62]. Models that are log-linear can be fitted through packages for generalized linear models (GLM), in particular the `glm` function. Models that are non-linear in their parameters, like the $RC(M)$ model introduced above, require special packages for their implementation, such as the `gnm` package of Turner and Firth [63] or the VGAM of Yee [69]. The implementation of association models via `gnm` is extensively illustrated in Section 6.6 of Kateri [42], while functions for fitting specific AMs are

provided in the web appendix of [42]. Here, we fit AMs using maximum likelihood estimation as implemented in the R ([62], version 4.0.0) package `logmulti` [13], which is a wrapper for the more general `gnm` package.

Goodness-of-fit (GoF) of AMs can be tested by the standard GoF tests for contingency table models, i.e., the likelihood ratio statistic (G^2) or the Pearson's X^2 . Since the values of the G^2 and X^2 statistics are strongly influenced by the sample size, we consider two additional statistics that give the practical significance and a more intuitive sense of GoF. The value of G^2 from independence can be thought of as a measure of the amount of dependency in the data. The percent of association accounted for by a model equals

$$\frac{(G_{ind}^2 - G_{model}^2)}{G_{ind}^2} \times 100,$$

where G_{ind}^2 and G_{model}^2 are the likelihood ratio test statistics for the model of independence and the model of interest. A second index, the dissimilarity index (D), equals the proportion of the data that would have to be moved from one cell to another for the model to fit perfectly. The dissimilarity index can be computed using frequencies or proportions; namely,

$$D = \frac{\sum_i |n_i - \hat{m}_i|}{2n} = \frac{\sum_i |p_i - \hat{p}_i|}{2},$$

where the sum is over all cells, n_i is observed frequency, \hat{m}_i is the estimated expected frequency, p_i is the proportion of data in cell i , and \hat{p}_i is the estimated probability of being in cell i . The rule of thumb is that a $D \leq 0.03$ is a good fitting model. We should note that D does not perform well for large tables, because, to achieve perfect fit, observations that would need to move to an adjacent cell have the same weight as those that would need to be moved many cells away.

1.3.2 Example: Who Takes Which MOOCs

The data in Table 1.1 come from a study examining engagement in massively open online courses (MOOCs) with the goal of determining who is being served by taking which course [12]. The data come from MOOCs covering six different disciplines where all MOOCs except one were offered multiple times. In total, there are 16 course offerings. The topics of the MOOCs were computer science (CS1, CS2), education (Educ1, Educ2), organic chemistry (Chem1, Chem2), business administration on subsistence (Bus1, Bus2, Bus3), environmental science (Env1–Env6), and animal and veterinary science (Animal). The students' ages were collected on a category scale of six age groups.

Table 1.1 Who takes which MOOC: a cross-classification of MOOC courses by age groups of students who take the courses

Course	Age groups					
	18–24	25–29	30–39	40–49	50–59	≥60
Animal	33	43	64	30	30	17
Bus1	59	101	100	68	49	28
Bus2	45	57	68	38	21	31
Bus3	20	47	62	32	28	35
Chem1	164	149	174	86	69	48
Chem2	71	52	54	27	18	13
CS1	1472	1472	2068	1110	580	254
CS2	198	199	342	199	124	46
Educ1	13	34	114	117	91	46
Educ2	10	20	77	81	65	37
Env1	92	216	313	154	139	117
Env2	126	265	342	197	176	147
Env3	89	155	217	143	149	114
Env4	90	163	216	99	77	63
Env5	111	175	206	134	109	111
Env6	42	78	119	60	62	72

Table 1.2 Goodness-of-fit statistics for models fitted to the MOOC data in Table 1.1

Model	df	G^2	p	Percent of association	Dissimilarity index
Independence (Poisson)	74	1098.3	<0.01	0.00%	0.09
Independence (Negative binomial)	74	101.46	0.02	90.76%	0.10
R (equidistant scores)	60	291.73	<0.01	73.44%	.16
R (midpoint scores)	60	313.13	<0.01	71.49%	.16
$RC(1)$	56	249.40	<0.01	77.29%	0.04
$RC(2)$	39	52.17	0.08	95.26%	0.02

The models that are relevant to this data set are (1.4) and (1.6) with $J_1 = 16$ and $J_2 = 6$. The $RC(M)$ models were fitted to the data using maximum likelihood estimation as implemented in the R package `logmult`.² Goodness-of-fit statistics for six models fitted to the data are reported in Table 1.2. For each model, we report df , G^2 , p -value, and the two additional statistics discussed in Sect. 1.3.1.

The MOOCs and age groups show a significant relationship ($G^2_{ind} = 1098.3$, $df = 75$, $p \leq 0.01$). Since G^2 may be significant due to large frequencies or extra heterogeneity between students within each of the combinations of MOOCs by age group, we also fitted a model of independence using the Negative Binomial

² The independence model and the R association models were fitted using `glm` and the independence model with a Negative binomial distribution was fitted using `glm.nb` in the MASS package.

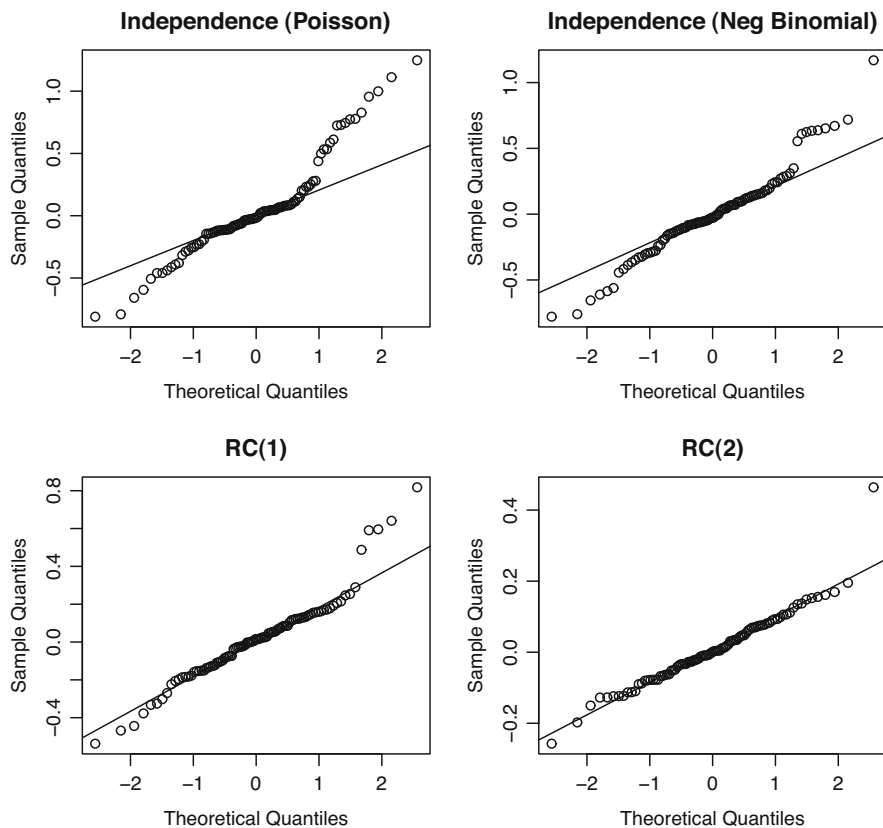


Fig. 1.1 QQplot of standardized residuals from models fitted to the MOOC data

distribution. This independence model also showed significant dependency ($G^2_{ind} = 101.46$, $df = 75$, $p = 0.02$). Furthermore, D is relatively large for both of these two models. The top two plots in Fig. 1.1 are the qqplot's of standardized residuals from the two independence models, and they show considerable departure from normality for smaller and larger frequencies, which gives us further evidence against independence. Examining a table of ($16 \times 6 =$) 96 residuals does not lead to insight into the relationship between age and MOOCs.

The simplest association model that can be fitted to data is the R model, where we can reasonably assign scores only to the column variable (i.e., age). The row variable is nominal and the associated scores have to be parameters. The two natural options for known scores would be either equidistant for successive categories or the midpoints of the corresponding age intervals. Neither of these two models provide satisfactory representations of the association in the data. The $RC(2)$ association model fits better than any of the simpler models ($G^2 = 52.17$, $df = 39$, $p = 0.08$, the percent association = 95%, $D = 0.02$). Furthermore, the bottom right qq-plot in

Fig. 1.1 shows that the standardized residuals from the $RC(2)$ model are very close to normal with the exemption of an outlying case. The parameter estimates from the $RC(1)$ and $RC(2)$ association models are plotted in Fig. 1.2 where the category scale values for the MOOCs and age groups are weighted by the square root of the association parameter (i.e., $\hat{v}_{i_j1}\sqrt{\hat{\sigma}_1^2}$ and $\hat{v}_{k_jk2}\sqrt{\hat{\sigma}_2^2}$).

Even though the $RC(2)$ model is our best model, for the purpose of illustration, the scale value plots for both the $RC(1)$ and $RC(2)$ models are given in Fig. 1.2. For both models, the scale values for the courses contrast STEM (Science, Technology, Engineering and Mathematics) and non-STEM courses; that is, at one extreme are the chemistry and computer science courses and at the other extreme the education courses. On the first dimension of both models, the scale values for student age groups are monotonically ordered with respect to age; however, they are not equally spaced. The age groups 25–29, 30–39 and 40–49 are relatively close in value in the $RC(1)$ model but less so in the $RC(2)$ model. From the $RC(2)$ graph we can say that students aged 18–24 have higher odds of taking STEM courses than the odds for any of the other age groups. Conversely, the students aged 50–59 have higher odds of taking the education courses than any of the students in other age groups. Different offerings of the same course tend to have similar scale values, especially Educ1 & Educ2. The odds of taking one or the other of these courses (regardless of age groups) is close to 1.

As illustrated in this example, the scale values from the $RC(1)$ and $RC(2)$ association models need not be the same. Also, for a given model, the scale values may be reflected (i.e., multiplied by -1) and this is illustrated in the scale values plots. For the $RC(1)$ model on dimension one, the ages go from low to high, but for the $RC(2)$ model go from high to low. The scale values for courses are also reflected in the $RC(2)$ model compared to the $RC(1)$ model, which leads to the same interpretations for the models.

1.4 Graphical Models

Log-linear models for categorical data have graphical representations that are visual representations of theory or scientific information, and they can be used to determine whether tables can be collapsed over items without impacting associations [22, 48]. Graphs also aid us in generalizing the $RC(M)$ association models to higher dimensions. Graphical models for log-linear models are introduced in this section, followed by graphs for $RC(M)$ association models. Lastly, we add more variables to the graphs to represent situations where we have moderate- to very high-dimensional tables (i.e., large numbers of items).

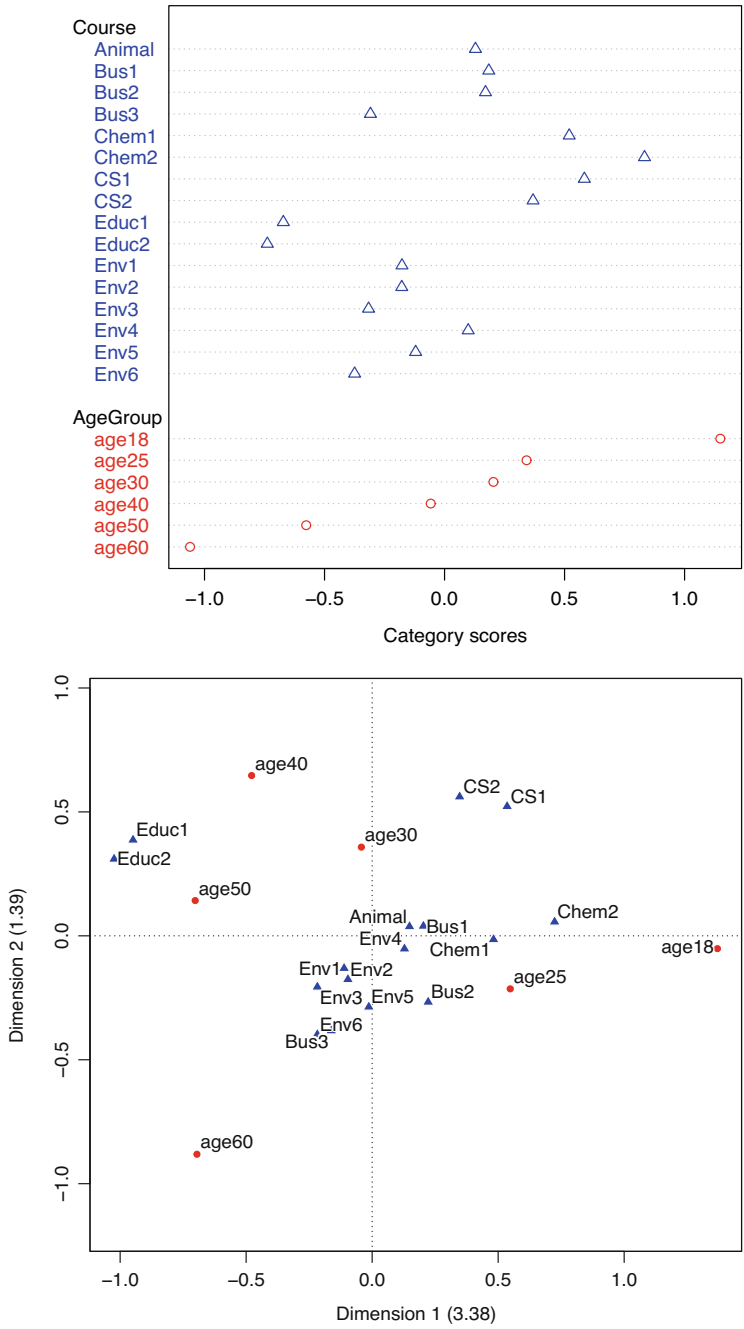


Fig. 1.2 Plot of estimated scale values from $RC(1)$ association model (top) and $RC(2)$ association model (bottom) fitted to MOOC data

1.4.1 Graphs for Log-linear Models

A graph consists of nodes, which for us are variables or items, and edges or lines connecting nodes indicating possible (non-directional) dependency between variables. For example, consider a three-dimensional $J_1 \times J_2 \times J_3$ contingency table, cross-classifying the categorical variables Y_1, Y_2, Y_3 .

Figure 1.3 contains four simple graphs showing the relationship between Y_1, Y_2 and Y_3 . In this chapter, discrete variables are represented by boxes. The absence of a line connecting two variables indicates that the two variables are independent conditional on the rest of the graph. The graph in Fig. 1.3a does not contain any edges and this graph represents complete independence. The presence of a line between two variables only indicates that they *may* be dependent conditional on the rest of the graph. Figure 1.3b represents a log-linear model of joint independence between Y_2 and Y_1 & Y_3 , and Fig. 1.3c represents a log-linear model of conditional independence between Y_1 and Y_2 given Y_3 .

Graphical models (a), (b) and (c) are collapsible over variables. For example, in (b), we can collapse the data over Y_2 and this does not change the dependency between Y_1 and Y_3 ; that is, we can simply analyze the marginal relationship between Y_1 and Y_3 . For model (c), conditional independence of Y_1 and Y_2 given Y_3 , we can collapse over Y_2 to study the relationship between Y_1 and Y_3 , and collapse over Y_1 to study the relationship between Y_2 and Y_3 . Any model for categorical variables that has some form of (conditional) independence can be collapsed over some set of variables (items); however, this is not the case for graph (d).

Figure 1.3d is a model of conditional dependence; that is, none of the variables are independent conditional on the rest of the graph. This graph is both a representation of a log-linear model with all 2-way interactions between pairs of variables and

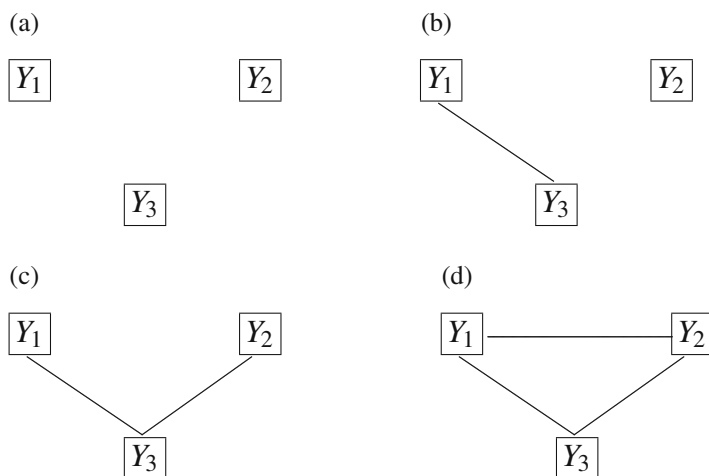


Fig. 1.3 Graphical models corresponding to log-linear models of (a) complete independence, (b) joint independence, (c) conditional independence, and (d) 3-way interaction

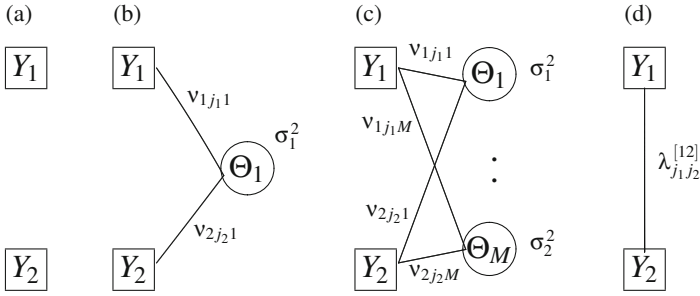


Fig. 1.4 Graphs for log-linear models for 2-way tables where (a) is a log-linear model of independence, (b) is the RC(1) association model, (c) is the RC(M) association model, and (d) is a saturated log-linear model

a model with all 2-way interactions and a 3-way interaction (i.e., a saturated model). For every model there is a unique graph, but every graph with edges (dependencies) can represent multiple models. This yields an ambiguity regarding the complexity of the interaction structure. In this chapter, we use graphs to represent theory and take the most complex model implied by a graph. For example, Fig. 1.3d, which is a complete graph,³ represents the log-linear model with all 2-way interactions and a 3-way interaction.

We can obtain graphical representations for our models such that there is more of a one-to-one correspondence between graphs and models. Consider the simpler case of 2 categorical variables. In Fig. 1.4, the graphs (a) and (d) represent log-linear models of complete independence and dependence, the latter being a saturated log-linear model. As commented on in Sect. 1.3, in this case we have $(J_1 - 1)(J_2 - 1)$ degrees of freedom with which to represent the dependency; however, we may not need all of these degrees of freedom. There are models in between independence and dependence, which we discussed in Sect. 1.3. We introduce an unobserved continuous variable to our graphs, which are represented by the circles in the graphs in Fig. 1.4b and c. The categorical variables are now conditionally independent given the latent continuous variable(s). Consider Graph (b) in Fig. 1.4. If we collapse over the continuous variable, we will produce an association between the categorical variables [48]. The model for observed data is one of dependence. Graph (b) is a representation of the *LL*, *U*, *R*, *C*, and *RC(1)* models. The differences depend on whether the scale values are set equal to specific values or are estimated. For the *LL* model, both v_{1j_11} and v_{2j_21} are set equal to specific values, for the *U* models both v_{1j_11} and v_{2j_21} are set to equally spaced scores, for the *R* (or *C*) model one set of scores (e.g., v_{ij_11}) is set to specific values and the other set (e.g. v_{2j_21}) is estimated, and for the *RC(1)* model, both v_{1j_11} and v_{2j_21} are estimated. Graph (c) is a representation of the *RC(M)* association model previously introduced in Sect. 1.3 and such graphs are discussed in more detail below.

³ A complete (sub)graph is one where all variables are directly related to each other.

1.4.2 Graphs of the $RC(M)$ Association Model

Figure 1.4b is a graphical representation of models for two variables corresponding to the U , LL , R , C and $RC(1)$ association models, and Fig. 1.4c is the representation of the $RC(M)$ model. To represent the AMs, we have added a continuous variable that is unobserved or latent. These continuous variables are represented by the circles. Goodman [30] first mentioned that a latent variable may underlie data fitted by an $RC(1)$ model, but he never expanded on this. We provide explicit details about a possible underlying or latent variable model and use this to generalize the $RC(M)$ model to high-dimensional tables.

Models can be “read” from the graphs. As an example, consider the $RC(1)$ association models represented by graph (b) in Fig. 1.4. All models for data include a parameter to ensure probabilities sum to 1 (i.e., λ), and include marginal effect terms for each categorical variable (i.e., $\lambda_{j_i}^{[i]}$ and $\lambda_{j_k}^{[k]}$). For the interaction, the lines connecting unobserved continuous and observed discrete variables are labeled with the category scale values, and the latent variable Θ_1 is labeled with σ_1^2 . The observed interaction between Y_1 and Y_2 equals the product of parameters on the path between Y_1 and Y_2 ; that is, $v_{1j_1} \sigma_1^2 v_{2j_2}$. Likewise, the interaction between Y_1 and Y_2 represented by Fig. 1.4c is $\sum_m \sigma_m^2 v_{1j_1m} v_{2j_2m}$.

The AMs in Fig. 1.4b and c are models of conditional independence: the (observed) categorical variables are independent given values on the unobserved continuous variables. The number of Θ_m s corresponds to the dimensionality of the $RC(M)$ model, which should not be confused with the dimension of a cross-classification (i.e., the number of variables). Since the Θ_m s are continuous, if we collapse over the Θ_m s, we may observe a dependency between the categorical variables. On the contrary, we cannot collapse over one categorical variable to study the relationship between the other categorical variable and the continuous variable. According to theory on graphical models, the graphs for the $RC(M)$ models are not collapsible [22, 48]; however, this is a property that does not hold in a strong sense, as will be shown in Sect. 1.5 in the context of higher-dimensional models.

To derive the algebraic model from the graph, we need two assumptions in addition to conditional independence. First, the observed data \mathbf{y} come from a multinomial distribution, which as mentioned above is a common assumption for tables of frequencies. This assumption is not restrictive, because for inferential purposes, the three standard sampling schemes for contingency tables (multinomial, product multinomial (i.e. independent multinomials in each row or column), and independent Poisson in each cell) are equivalent. We must also assume that the latent variables follow a (multivariate) normal distribution where the mean and variance are conditional on the response patterns (i.e., cells of the table); that is,

$$\boldsymbol{\theta} \mid \mathbf{y} \sim MVN(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y).$$

Justification for the assumption of a conditional Gaussian distribution for $\boldsymbol{\theta}$ can be found in Chang [14, 15] and [46]. The association parameters of the $RC(M)$ model are the elements of $\boldsymbol{\Sigma}_y$. Typically, we assume a homogeneous conditional

covariance matrix, i.e., $\Sigma_{\mathbf{y}} = \Sigma$. Previously, we discussed the orthogonality identification constraint on the \mathbf{v}_{im} s for $M > 1$, which requires that Σ is a diagonal matrix, $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_M^2)$. The conditional mean of θ_m is the sum of the category scale values that are directly related to θ_m weighted by σ_m^2 ; that is,

$$\boldsymbol{\mu}_{\mathbf{y}} = \left(\sigma_1^2 \sum_i v_{i j_1}, \sigma_2^2 \sum_i v_{i j_2}, \dots, \sigma_M^2 \sum_i v_{i j_M} \right)' \quad (1.8)$$

The $RC(M)$ association models do not include values of the θ_m s; however, the models do include parameters that give us the distributional parameters of $\boldsymbol{\theta}|\mathbf{y}$.

1.5 High-Dimensional Tables

High-dimensional tables are common, especially when considering questions on surveys, items on psychological scales, or items on educational tests. Two problems faced with analyzing high dimensional tables are the large numbers of (i) 2-way interactions, and (ii) cells. For example, with 20 five category items, there are $20(19)/2 = 190$ different 2-way interactions and $5^{20} = 9.536743e + 13$ cells in the cross-classification of the items. To deal with the problem of large numbers of interactions, we generalize the AMs to large numbers of variables. For the second problem where the table is large and data are sparse, we use pseudo-likelihood estimation. In this section, we tackle both problems and discuss the connection between AMs and IRT models.

To generalize the association models to high-dimensional cross-classifications, we start with graphs and subsequently discuss the algebraic model. We continue to only consider two-way interactions, because item response models using the standard assumption that $f(\boldsymbol{\theta})$ is multivariate normal imply only two-way interactions between items. We will discuss the similarities and differences with respect to the $RC(M)$ association model, as well as explicitly show the correspondence of association model parameters and common IRT models.

1.5.1 Graphs for High-Dimensional Association Models

For high-dimensional tables, we simply add variables to the graphs, as in Fig. 1.5. Figure 1.5 has three examples of possible graphs for 6 items. Graph 1.5a is similar to an $RC(1)$ model, except instead of 2 categorical variables we have 6. In all graphs in Fig. 1.5, the categorical variables are conditionally independent given the unobserved continuous variable(s); however, the latent variables can be dependent. The covariance between latent variables θ_m and $\theta_{m'}$ conditional on the observed

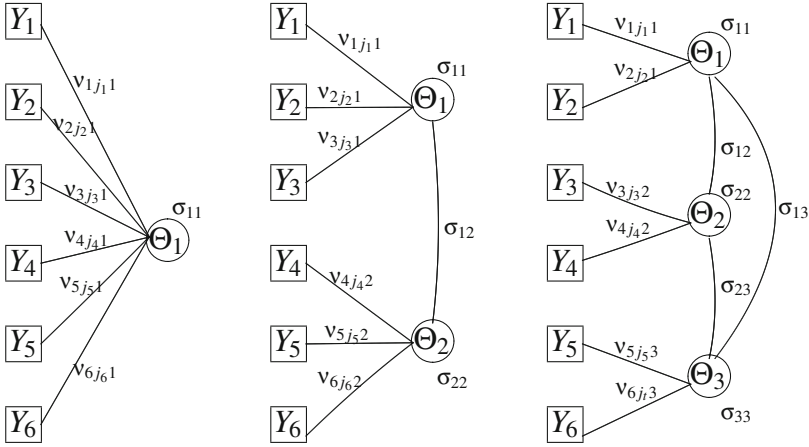


Fig. 1.5 Graphs for log-multiplicative association models for 1, 2, and 3 continuous latent variables (circles) and six observed categorical variables (squares)

variables is equal to $\sigma_{mm'}$. We have changed our notation slightly and are using σ_{mm} rather than σ_m^2 for variances (i.e., association parameters).

If categorical variables are discrete measures of underlying continuous variables, then it would stand to reason that the scale values for the variables are the same over the interactions; that is, the scale values would be homogeneous. For example, in Graph 1.5a the interaction between, say variables Y_i and Y_k , would be represented by $\sigma_{11}v_{ij_1 1}v_{kj_1 1}$ and the interaction between Y_i and Y_ℓ would be $\sigma_{11}v_{ij_1 1}v_{\ell j_\ell 1}$, both of which involve $v_{ij_1 1}$ and σ_{11} .

Just as we replaced two-way interaction parameters in a log-linear model for 2 items by products of association parameters and scale values to get an $RC(M)$ model, we do the same for association models for high-dimensional tables. The interactions between the categorical variables are the products of labels of the paths between them. For example, the interactions between variables Y_1 and Y_6 for the graphs in Fig. 1.5 are

$$\sigma_{11}v_{1j_1 1}v_{6j_6 1} \text{ for Graph (a)}$$

$$\sigma_{12}v_{1j_1 1}v_{6j_6 2} \text{ for Graph (b)}$$

$$\sigma_{13}v_{1j_1 1}v_{6j_6 3} \text{ for Graph (c).}$$

The latter two involve (conditional) covariances between the latent variables.

1.5.2 Algebraic Details and Properties

The most general case, where each item is directly related to each of the latent variables and all latent variables are related to each other, leads to the following complex association model:

$$P(\mathbf{y}) = \exp \left[\lambda + \sum_i \lambda_{J_i}^{[i]} + \sum_i \sum_{k>i} \sum_m \sum_{m' \geq m} \sigma_{mm'} v_{i j_i m} v_{k j_k m'} \right]. \quad (1.9)$$

This model has an intercept, all main effects, and all possible two-factor interactions, where the interactions have a multiplicative structure. For the models to be equivalent to a hierarchical log-linear model of all two-factor interactions would require the number of terms (dimension) for the $RC(M_{ik})$ interaction term of every pair of items Y_i and Y_k , for $i, k = 1, \dots, I$, to equal $M_{ik} = \min(J_i, J_k) - 1$.

A variety of more parsimonious models with a special structure for the associations among the variables of sound interpretation can be obtained by considering smaller values for the rank of the interaction terms or/and homogeneity of scores across interaction terms. Furthermore, higher-order interactions having multiplicative terms among scores for more than two variables are possible. For the case of three-factor interactions and related references we refer to [42, Sections 6.7, 6.8.1].

These complex models require a considerable number of identification constraints; therefore, for the sake of discussion, we restrict our attention to models where each item is related to only one latent variable, which means that all interaction terms are of $RC(1)$ type. The simple structures shown in Fig. 1.5 imply that each item has only one set of $v_{i j_i m}$ s that are not all equal to zero. If Y_i is not related to a Θ_m , then $v_{im} = \mathbf{0}$. For example, in Fig. 1.5b and c, there is no edge between Y_1 and Θ_2 so $v_{1 j_1 2} = 0$, $j_1 = 1, \dots, J_1$. In AM models, we can have additional edges between, say, Y_1 and other Θ s; however, the simple structures discussed here prove to be sufficient for many applications.

Association Model (1.9) can be derived from the same assumptions as the $RC(M)$ models: \mathbf{y} is multinomial, $\boldsymbol{\theta}$ is conditional Gaussian,⁴ and conditional independence of items given $\boldsymbol{\theta}$. One difference is that the conditional covariance matrix does not have to be diagonal [4]. As a result, the means within response patterns also include responses to other variables; namely,

$$E(\theta_m | \mathbf{y}) = \sigma_{mm} \left(\sum_i v_{i j_i m} \right) + \sum_{m' \neq m} \sigma_{mm'} \left(\sum_k v_{k j_k m'} \right). \quad (1.10)$$

For simple structures, items Y_i load on latent variable θ_m and Y_k load on $\theta_{m'}$. Estimates of θ_m are based not only on the items directly related to θ_m , as in (1.8),

⁴ The marginal distribution of $\boldsymbol{\theta}$ is a mixture of Gaussian distributions.

but also those that are indirectly related through $\theta_{m'}$. When $\sigma_{mm'} \neq 0$, measurement can be improved and become more precise by including multiple correlated latent variables [21, 66]. In addition to the derivation based on statistical graphical models, Model (1.9) can be derived from an IRT perspective ([3, 5, 16, 37, 38], conditional specification of models [3, 5]), a theory of ferrimagnetism [47, 52], distance-based models [18–20], and others.

To facilitate the discussion of the models, we use the following two-dimensional model for 4 categorical variables where variables Y_1 and Y_2 are directly related to θ_1 and variables Y_3 and Y_4 are directly related to θ_2 :

$$P(\mathbf{y}) = \exp\left[\lambda + \sum_{i=1}^4 \lambda_{j_i}^{[i]} + \sigma_{11}(v_{1j_11}v_{2j_21}) + \sigma_{22}(v_{3j_32}v_{4j_42}) \right. \\ \left. + \sigma_{12}(v_{1j_11}v_{3j_32} + v_{1j_11}v_{4j_42} + v_{2j_21}v_{3j_32} + v_{2j_21}v_{4j_42})\right]. \quad (1.11)$$

A log-linear model with all 2-way interactions for 4 five-category variables requires $(6 \times 4 \times 4) = 96$ unique parameters to represent the interactions; whereas, the association model with homogeneous scale values across interactions (e.g., (1.11)) requires at most $(4 \times 4) + 1 = 17$ unique parameters.⁵ The difference between the number of parameters of log-linear and AMs increases exponentially for more items and categories per item.

Unlike AMs for two-way tables that require more than two categories per variable, this is not the case for the higher-dimensional models. For example, Model (1.11) and models that correspond to the graphs in Fig. 1.5a, b and c can be fitted to binary variables [4, 5].

The identification constraints on the location of marginal effect terms and the scale values are analogous to the $RC(1)$ model (e.g., $\sum_{j_i} \lambda_{j_i}^{[i]} = 0$ and $\sum_{j_i} v_{ijim} = 0$) and just one scaling constraint is required for each latent variable. For example, in (1.11), possible scaling constraints can be either

$$\sigma_{mm} = 1 \quad \text{for all } m,$$

or

$$\sum_{j_i} (v_{ijim})^2 = 1 \quad \text{for one } i \text{ per } \theta_m \text{ for all } m,$$

but not both, analogous to the $RC(1)$ model. In example (1.11), if $\sigma_{11} = \sigma_{22} = 1$, then we cannot linearly transform the v_{ijim} s without changing the values of the interaction terms. If we fix the variances and rescale v_{1j_11} such that $\sum_{j_i} v_{ijim}^2 = 1$, the interaction between variables does not necessarily remain the same. Placing

⁵ The association model may have even fewer unique parameters depending on whether restrictions are placed on the scale values.

scaling constraints on both of the σ_{mm} and v_{ijm} is a restriction that impacts the goodness-of-fit of the model. For example, if we set variance to $\sigma_{11} = 1$ and re-scale the v_{1j_11} , then the interaction between Y_1 and Y_2 changes,

$$\sigma_{11}v_{1j_11}v_{2j_21} \neq 1(v_{1j_11}/c)v_{2j_21} = (1/c)v_{1j_11}v_{2j_21},$$

where $c = \sqrt{\sum_{j_1} (v_{1j_11})^2}$. To achieve equality, either $\sum_{j_1} v_{1j_1m}^2 \neq 1$ or $\sigma_{11} = 1/c$. Whether the scaling constraint is put on σ_{mm} or the scale values is more a matter of convenience. For example, for estimation of models for our example, we found it more convenient to set $\sigma_{mm} = 1$; however, after the model has been fitted, we can switch to $\sum_{j_i} v_{ijim}^2 = 1$ and adjust σ_{mm} (and the $\sigma_{mm'}$) and other scale values. We did the latter in a simulation study reported below on the collapsibility over items where we needed to separate the effects of the strength and structure of the association.

In the standard AMs framework, (1.11) is an AM having RC(1)-type two-factor interactions and every variable has homogeneous scores (i.e., the same scores across all interaction terms involved)

$$P(\mathbf{y}) = \exp\left[\lambda + \sum_{i=1}^4 \lambda_{j_i}^{[i]} + \phi_{12}v_{1j_11}v_{2j_21} + \phi_{34}v_{3j_32}v_{4j_42} \right. \\ \left. + \phi_{13}v_{1j_11}v_{3j_32} + \phi_{14}v_{1j_11}v_{4j_42} + \phi_{23}v_{2j_21}v_{3j_32} + \phi_{24}v_{2j_21}v_{4j_42}\right], \quad (1.12)$$

with the additional constraint on certain intrinsic association parameters $\phi_{13} = \phi_{14} = \phi_{23} = \phi_{24} = \sigma_{12}$ (notice that $\phi_{12} = \sigma_{11}$ and $\phi_{34} = \sigma_{22}$). Such constraints are unusual for standard AMs, but are found in square tables applications (rows and columns are the same categories) and are linked to latent variables models (e.g., IRT models) later in Sect. 1.5.4. In applications with all variables (items) being measured on the same scale, we find homogeneity constraints on the scores for each variable and dimension (i.e., $v_{ijm} = v_{kjm}$ where $j_i = j_k$) that result in symmetric interaction terms.

To understand the physical interpretation of this constraint, consider (1.12) under the additional assumption that the scores of all variables are known, equidistant for successive categories (i.e. $v_{i(j_i+1)m} - v_{ijim} = c_i$, for all $j_i = 1, \dots, J_i - 1$, with $m = 1$ for $i = 1, 2$ and $m = 2$ for $i = 3, 4$), which means that we assume U -type structures for all interactions. In particular for the (Y_1, Y_2) partial table when $Y_3 = j_3$ and $Y_4 = j_4$, the OR equals

$$\theta_{j_1j_2|j_3,j_4}^{[12]} = \exp\left(\frac{\pi_{j_1,j_2,j_3,j_4}\pi_{j_1+1,j_2+1,j_3,j_4}}{\pi_{j_1,j_2+1,j_3,j_4}\pi_{j_1+1,j_2,j_3,j_4}}\right) \quad (1.13) \\ = \exp(\phi_{12}(v_{1(j_1+1)1} - v_{1j_11})(v_{2(j_2+1)1} - v_{2j_21})) \\ = \exp(\phi_{12}c_1c_2) = \theta^{[12]},$$

while for the other partial tables, the $\theta^{[ik]}$'s, $i, k = 1, \dots, 4$ with $i \neq k$, are defined analogously. Consequently, the conditional local ORs in every partial table (Y_i, Y_k) are all equal to $\theta^{[ik]}$, for all values of j_i and j_k (uniform) but also across all levels of the other items (homogeneous). Thus the underlying model is the homogeneous U model (see [42, Section 6.7]). Notice that due to the sum-to-zero constraints satisfied by the scores, $c_i \neq c_k$ if $J_i \neq J_k$. For the special case of $J_i = J$, $i = 1, \dots, 4$, it holds $c_i = c$ and the additional equality constraint among the ϕ parameters above leads to $\theta^{[13]} = \theta^{[14]} = \theta^{[23]} = \theta^{[24]}$, hence to equality of the corresponding conditional local ORs.

A difference in terms of identification constraints with respect to AMs for two-way tables with M latent variables (i.e. $RC(M)$ models), is that in models of type (1.11) with more than two variables, each variable is directly related to only one of the M latent variables; whereas, under $RC(M)$ each variable is related to all M of them. The orthogonality constraints that are required for the $RC(M)$ are not required for (1.11) and Σ can have non-zero off-diagonals. This is not true for all versions of (1.9); in particular, if every variable is directly related to each and every Θ_M , then an orthogonality constraint is required as well as for underlying bi-factor structures.

An alternative version of (1.9) that has the same identification constraints as (1.11) may have all variables directly related to each of the M latent variables, except one per latent variable. The variables that are related to just one M "anchor" the rotation. Similarly, in a factor analysis/IRT model framework, parameter constraints are imposed to uniquely identify the model parameters. In a factor model with M latent variables, M^2 constraints are required to obtain a unique solution and avoid the rotational indeterminacy issue. Among the constraints are those that set the scale of the latent variable. Similarly to what it has been said above for $RC(M)$ models, the scale of a latent variable is set either by standardizing the latent variable assuming that it has zero mean and unit variance in the population or by forcing its scale to be the same as one of the observed variables. Usually, the variable that best represents the latent variable has its factor loading set equal to one. The selected variable is known as a "reference" variable. Setting the scale of the latent variables to one takes care of M of the required restrictions. The additional ones are imposed on the loading and factor covariance matrices (e.g. in a diagonal factor covariance matrix, certain loadings are set to zero). In exploratory factor analysis, the required restrictions can be imposed on any of the parameters. Those restrictions will produce an arbitrary set of factors which can be then rotated to another set of factors that have better interpretability. In confirmatory factor analysis, the constraints are driven by the investigator's research hypothesis. A useful constraint that eases the interpretation of the factors is to consider that each latent variable has at least one item that loads solely on that factor (i.e. setting specific elements of the loading matrix to zero) [40, 41]. Returning to the AMs framework, anchoring one item (e.g., $\sum_j v_{ijm}^2 = 1$ and $v_{ijm'} = 0$ for $m' \neq m$) that best represents the latent variable is a key to fitting non-simple structure models.

With the $RC(M)$ association model, we cannot collapse over an item and study the relationship between the other item and a latent variable, because then we would have only one observed variable. This is not true in a strong sense for more than 2 items. As mentioned previously, v_{ijim} represents the structure and σ_{mm} represents the strength of the relationship between variables. This leads to a semi-collapsible situation. This is illustrated for the unidimensional model, such as Fig. 1.5a. If we have 50 items but drop or collapse over one of them, the structure of the relationship between an item and the latent variables should not change but the strength of associations increases [6]. To illustrate this property, 100 data sets were simulated for 50 four-category items. A nominal IRT model was used to simulate the data where the slopes (scale values) were drawn from an $N(0, 1.5)$, and the location (marginal effects) were drawn from an $N(0, 2)$. Values of θ for each of $n = 1000$ observations for each of the 100 replications were drawn from an $N(0, 1)$. For cases where a simulated item did not have observations in all categories, a new data set was randomly created. The model was fit to data sets with 50 down to 15 items dropping 5 items at a time, and 15 down to 4 items dropping one item at a time. The scaling identification constraint was placed on the first item (i.e., $\sum_{j=1}^3 (v_{1j1})^2 = 1$) and σ_{11} was estimated.

In Fig. 1.6, the means over replications of association parameters and scale values are plotted by the number of items in the data set. The estimated σ_{11} s are larger for small numbers of items and asymptotes down to 0 for large numbers of items. Recall that σ_{11} is the conditional variance of θ_1 within a response pattern \mathbf{y} (i.e., a cell in a cross-classification of items). When the number of possible patterns is very large (infinity), only one person can fall into a cell of the table, so the variance necessarily equals 0. Also in Fig. 1.6 are the means of the estimated scale values where each line in the figures corresponds to a different item. The scale values plots are only given for three of the four categories because $v_{i4i1} = -\sum_{j=1}^3 v_{iji1}$. The lines are essentially flat. In other words, we can collapse over categorical variables and the structure between variables remains the same. The only thing that changes is the strength of the relationship between observed variables (items), and between items and latent variables. This also holds for multidimensional models [6]. The association between two observed variables, say Y_1 and Y_2 , is represented in the AM by $\sigma_{mm'} v_{1j_1m} v_{2j_2m'}$ (for $m = m$ and $m \neq m'$). As variables are dropped σ_{mm} and $\sigma_{mm'}$ both get larger, but the $v_{ijim'}$ stay essentially the same. Therefore, associations are larger for fewer items but the structure remains the same.

1.5.3 Pseudo-likelihood Estimation

Large numbers of variables result in large, sparse cross-classifications, in which case maximum likelihood estimation becomes computationally infeasible. An alternative is pseudo-likelihood estimation (PLE), which takes a large complex problem and reduces it to a number of simpler and smaller problems [8, 10, 11, 29]. A joint distribution can be specified by a set of conditional distributions [28] where the

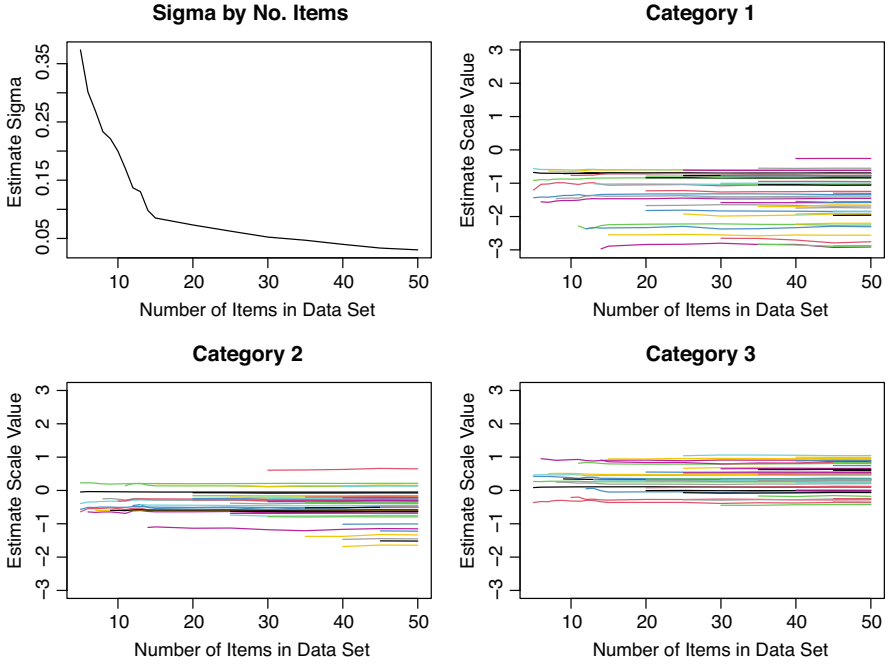


Fig. 1.6 Mean estimated σ_{11} and v_{iji} parameters from fitting a unidimensional nominal model to data simulated from a nominal item response model for $n = 500$ and four category items, with the number of items varying up to 50

conditional distributions are compatible and consistent with the joint distribution and imply a unique model for the joint distribution. Rather than maximize the likelihood of the joint distribution, PLE for AMs maximizes the product of the (log) likelihoods of one variable conditional on the rest. Pseudo-likelihood estimators are asymptotically consistent and normal [8, 29, 51]. The conditional distributions of (1.9) for one item given responses to all other items is a discrete choice model or conditional multinomial logistic regression model. The conditional distribution for item i for individual s is

$$P(Y_{is} = j | \mathbf{y}_{-i,s}) = \frac{\exp[\lambda_{ji}^{[i]} + v_{ijm} \sum_{k \neq i} \sum_{m'} \sigma_{mm'} v_{kjkm'}]}{\sum_{j=1}^J \exp[\lambda_{ji}^{[i]} + v_{ijm} \sum_{k \neq i} \sum_{m'} \sigma_{mm'} v_{kjkm'}]} \quad (1.14)$$

$$= \frac{\exp[\lambda_{ji}^{[i]} + v_{ijm} \tilde{\theta}_{-i,ms}]}{\sum_{j=1}^J \exp[\lambda_{ji}^{[i]} + v_{ijm} \tilde{\theta}_{-i,ms}]} \quad (1.15)$$

$$= \frac{\exp[\lambda_{ji}^{[i]} + \sum_{m'} \sigma_{mm'} \check{\theta}_{im's}]}{\sum_{j=1}^J \exp[\lambda_{ji}^{[i]} + \sum_{m'} \sigma_{mm'} \check{\theta}_{im's}]} \quad (1.16)$$

where $y_{-i,s}$ are the responses by person s to all items except item i , and j_k are the categories chosen by individual s . The predictor variable in (1.15), $\tilde{\theta}_{-i,ms}$, is the weighted sum of person s 's scores on all items except item i ; that is,

$$\tilde{\theta}_{-i,ms} = \sum_{k \neq i} \sum_{m'} \sigma_{mm'} v_{kjkm'}.$$

Note that $\tilde{\theta}_{-i,ms}$ depends on individual s . Using this value for $\tilde{\theta}_{-i,ms}$, we can get estimates of λ_{ij_i} and v_{ijjm} by fitting (1.15) to the data for item i .

There are multiple $\sigma_{mm'}$'s that need to be estimated, one $\check{\theta}_{im's}$ for each $\sigma_{mm'}$. The $\check{\theta}_{im's}$ equal a different weighted sum of persons s 's scale values for $k \neq i$, specifically,

$$\check{\theta}_{im's} = v_{ijm} \sum_{k \neq i} v_{kjkm'},$$

where the sub-script j_k indicates the categories chosen by s on item k . The $\check{\theta}_{im's}$ differ over individuals, items, and categories. The slopes in (1.16) are the same over individuals and items. If we had estimates of $\check{\theta}_{im's}$, estimates of $\lambda_{ij_i}^{[i]}$ and $\sigma_{mm'}$ can be obtained by fitting (1.16) to the data for just item i ; however, the $\sigma_{mm'}$ must be the same over items. We need to also estimate all possible $\sigma_{mm'}$'s. Fitting only (1.16) to one item's data does not yield all possible $\sigma_{mm'}$'s. For example if $m = 1$, then only $\sigma_{1m'}$'s would be estimated but not, say, σ_{23} . To impose the restriction on $\sigma_{mm'}$'s over items and estimate all of them, we vertically concatenate or "stack" the data and fit a single discrete choice model to the stacked data. In the stacked data set, there are blocks of J_i lines for each item and each individual.

Of importance is the recognition that if we have the conditionals in (1.14) for every item, the set of models are compatible and consistent with a joint distribution for all the items. The set actually overdetermines the joint distribution and thus requires restrictions on the parameters. The restrictions are that the terms that represent the interaction of i and k are the same whether i is modeled as a function of k or k as a function of i . These terms equal $v_{ijjm} \sigma_{mm'} v_{kjkm'}$, and since $\sigma_{mm'} = \sigma_{m'm}$, the restriction is met. The set of fully conditional distributions given by (1.14) uniquely imply the AM in (1.9) for the joint distribution of \mathbf{Y} ([1, 3, 5] and references therein). Since the discrete choice models [39] can be considered as a generalization of the stereotype model of Anderson [7], this link of discrete choice models to AM is a natural extension of the connection between the stereotype model and AM with $M = 1$ (s. Section 8.4.4 in [42]).

For unidimensional models, v_{ijjm} and $\lambda_{ij_i}^{[i]}$ are estimated by fitting Model (1.15) to the data for item i using the current estimates of the scale values for all $k \neq i$ and the $\sigma_{mm'}$ parameters to compute the predictor variable $\tilde{\theta}_{-i,ms}$. This is done successively for each item and fitting of the model to item data is iterated until convergence is achieved. For multidimensional models, estimates of $\sigma_{mm'}$'s are obtained by fitting (1.16) to the stacked data set using current estimates of scale values to

compute the $\check{\theta}_{im's}$ values. For unidimensional models, only item parameters are estimated; whereas, for multi-dimensional models, the algorithm iterates between updating v_{ijm} parameters and $\sigma_{mm'}$ parameters. If fixed scores are input (e.g., $v_{ijm} = 0, 1, \dots, (J_i - 1)$), then model (1.16) is only fitted once.

We maximize the pseudo-likelihood function by fitting discrete choice models to data using MLE. In **R**, discrete choice models can be fitted using `mlogit` ([17], `mlogit` [36], `mcllogit` [24]), and others. Due to the data manipulation required and iterative nature of the PLE algorithm, PLE for log-multiplicative association models has been implemented in the **R** package `pleLMA` (Anderson, 2021). The package `mlogit` is used in `pleLMA`, because it is efficient and can handle large data sets. Alternative packages, `IssingSampling` [25] and `plRasch` [2], both implement pseudo-likelihood estimation but they are more limited especially in terms of models for multicategory data and the estimation of category scale values. The `pleLMA` package can be found on CRAN and includes a detailed vignette on the usage of the package.

PLE for estimation of AMs has been extensively studied for small problems and yields estimates for both v_{ijm} and λ_{ij} that are nearly identical to MLE values. Paek [57, 58] simulated data from (M)IRT models for different numbers of categories (3, 4, 5), different numbers of items (4, 6, 20, 50), 1- to 4-dimensional models, and different sample sizes (200, 500, 1000). For small numbers of items and unidimensional models, she found correlations between parameter estimates from MLE and PLE equal to .999 to 1.000, and for multi-dimensional models most correlations were greater than .980. For larger problems where MLE was not possible, data were simulated from an (M)IRT model and results were compared. Paek [57, 58] found that PLE estimates recovered the parameters used to simulate the data, were unbiased, and had small root mean squared errors. This was true for different numbers of categories, different numbers of items, 1- to 4-dimensional models, and different sample sizes.

Alternative tools for model assessment are required, because the data for high-dimensional tables is sparse. Additionally, we do not obtain fitted values for response patterns (i.e., cells in the table), because estimating the λ -parameters is computationally and numerically challenging even given estimates of all other parameters. Some alternative methods are described Sect. 1.6 and others are illustrated in the context of our example; however, we conclude this section by briefly describing the connections between the AMs and (M)IRT models.

1.5.4 Connection to IRT Models

The conditional model in (1.15) has the same form as the nominal response model, including all of its special cases (e.g., models in the Rasch family, the two-parameter logistic model, the generalized partial credit model (GPCM)). The mathematical equivalence between AMs and (M)IRT models can be proven formally [1, 6, 47, 52]. From (1.15), the $\tilde{\theta}_{-i,ms}$ is person s 's value on the latent variable based on all

items except i ; however, after fitting a model we would use (1.10) to estimate the mean given a response pattern. The marginal effect parameters are sometimes referred to as “difficulty” or location parameters, and rather than denoted by $\lambda_{ji}^{[i]}$, they are usually represented by b_{ij} . The scale values v_{ijm} are slopes on the latent variables and are “discrimination” parameters, often denoted as a_{im} . The following restrictions on the category scale values lead to common IRT models:

$$\begin{array}{ll}
 \text{Nominal: } v_{ijm} = a_{ijm} & \text{no restrictions} \\
 \text{GPCM: } v_{ijm} = a_{im}x_j & \text{where } x_j = \text{fixed scores} \\
 \text{Rasch: } v_{ijm} = x_j & \text{where } x_j = \text{fixed scores.}
 \end{array} \tag{1.17}$$

The fixed scores, x_j , are typically set to equally spaced values or consecutive integers. Note that the conditional (partial) odds ratios are functions of the association parameters and category scale values. The conditional odds ratio for items i and k for the nominal model equals

$$\exp[\sigma_{mm'}(v_{ijm} - v_{ij'm})(v_{k\ell m'} - v_{k\ell'm'})].$$

When the x_j s equal consecutive integers, as is typical for GPCM and Rasch models, the local partial OR (1.13) for models in the Rasch family reduce to $\exp(\sigma_{mm'})$, since $c_i = 1$. Instead of just one value for local ORs in the 2-way table case, there is one for each latent variable and one for each pair of latent variables for a total of $M(M - 1)/2 + M$ local conditional ORs. Regardless of the number of variables, the number of these ORs depends on the number of latent variables. For the GPCM with consecutive integers, the local conditional ORs equal $\exp(\sigma_{mm'}a_{im}a_{km'})$; that is, (1.13) for items i and k with $c_i = a_{im}$ and $c_k = a_{km'}$.

In the AM framework, there is flexibility in setting the x_j s, which can be set to non-equally spaced values and different values over items. These possibilities yield item response models that deviate from the traditional Rasch and GPCM models. If the ordering of the response options is not clear, the category scale values from the nominal model can reveal the proper ordering and whether the spacing between category scale values is approximately equal. The scale values from nominal models can show whether a GPCM is plausible. Alternatively, models can be constructed where some items follow a GPCM and others a nominal model. There is great flexibility in crafting a model for data.

1.6 Sampling Properties

Let us denote with ω the parameter vector corresponding to the fitted model. For example, for Model (1.11),

$$\omega' = (\lambda, \lambda^{[1]}, \lambda^{[2]}\lambda^{[3]}, \lambda^{[4]}, \nu_{11}, \nu_{21}, \nu_{32}, \nu_{42}, \sigma_{11}, \sigma_{22}, \sigma_{12}),$$

where $\boldsymbol{\lambda}^{[i]}$ is a vector with elements $\lambda_{ji}^{[i]}$, and \mathbf{v}_{im} is a vector with elements v_{ijim} . From the theory of composite likelihood estimators (pseudo-likelihood), it holds that $\sqrt{N}(\hat{\boldsymbol{\omega}}_{PL} - \boldsymbol{\omega}) \xrightarrow{d} \mathcal{N}(0, G^{-1}(\boldsymbol{\omega}))$, where $G(\boldsymbol{\omega})$ is the Godambe information matrix [50, 64], (also known as the sandwich information matrix) given by

$$G(\boldsymbol{\omega}) = H(\boldsymbol{\omega})J^{-1}(\boldsymbol{\omega})H(\boldsymbol{\omega}),$$

where

$$H(\boldsymbol{\omega}) = E \left\{ -\frac{\partial^2}{\partial \boldsymbol{\omega}' \partial \boldsymbol{\omega}} pl(\boldsymbol{\omega}; \mathbf{y}) \right\},$$

$$J(\boldsymbol{\omega}) = Var \left\{ \frac{\partial}{\partial \boldsymbol{\omega}'} pl(\boldsymbol{\omega}; \mathbf{y}) \right\},$$

and $pl(\boldsymbol{\omega}; \mathbf{y})$ is the log pseudo-likelihood function. $H(\boldsymbol{\omega})$ and $J(\boldsymbol{\omega})$ can be estimated by:

$$\hat{H}(\hat{\boldsymbol{\omega}}_{PL}) = -\frac{1}{N} \frac{\partial^2}{\partial \boldsymbol{\omega}' \partial \boldsymbol{\omega}} pl(\boldsymbol{\omega}; (\mathbf{y}_1, \dots, \mathbf{y}_N)) \Big|_{\boldsymbol{\omega}} = \hat{\boldsymbol{\omega}}_{PL} \quad (1.18)$$

and

$$\hat{J}(\hat{\boldsymbol{\omega}}_{PL}) = \frac{1}{N} \sum_{n=1}^N \left(\frac{\partial}{\partial \boldsymbol{\omega}'} pl(\boldsymbol{\omega}; \mathbf{y}_n) \Big|_{\boldsymbol{\omega}=\hat{\boldsymbol{\omega}}_{PL}} \right) \left(\frac{\partial}{\partial \boldsymbol{\omega}'} pl(\boldsymbol{\omega}; \mathbf{y}_n) \Big|_{\boldsymbol{\omega}=\hat{\boldsymbol{\omega}}_{PL}} \right)', \quad (1.19)$$

respectively.

1.7 Evaluation and Testing

The pseudo-likelihood estimation framework used here falls within the composite likelihood (CL) framework which is used for approximating complex full likelihoods. The inference part under CL requires certain modifications and corrections similar to the ones needed for misspecified models [56]. Overall goodness-of-fit test statistics (e.g. likelihood ratio, Wald, and score test) and model selections criteria (e.g. Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC)) can be derived under the CL estimation framework. Adjusted Wald, score, and the likelihood ratio test statistic for overall fit and nested models under the CL framework have been developed for models for multivariate clustered data, time series data, and structural equation models [29, 43, 49, 56, 64]. Moreover, the model selection criteria AIC and the BIC are appropriately adjusted to hold under CL.

1.7.1 Composite Likelihood Ratio Test for Overall Fit

The fit of the model can be assessed by constructing a likelihood ratio test for testing $H_0 : \pi_r = \pi_r(\omega)$ against $H_1 : \pi_r$ subject to $\sum \pi_r = 1$, where ω is a vector of all independent parameters, r runs over all possible response patterns (cells of the contingency table), and π_r is the probability of response pattern r . In particular, $\pi_r(\omega)$ is defined by a model such as (1.6) or (1.11). The maximum of log-likelihood ($\ln L$) under H_0 and multinomial sampling is

$$\ln L_0 = \sum_r n_r \ln \hat{\pi}_r = N \sum_r p_r \ln \hat{\pi}_r, \quad \hat{\pi}_r = \pi_r(\hat{\omega})$$

and the maximum of $\ln PL$ under H_1 (saturated model) is

$$\ln L_1 = \sum_r n_r \ln p_r = N \sum_r p_r \ln p_r,$$

where n_r is the number of times response pattern r occurs in the sample, $p_r = n_r/N$ and N is the sample size. The likelihood ratio (LR) test statistic is

$$\chi_{\text{LR}}^2 = 2 \sum_r n_r (\ln p_r - \ln \hat{\pi}_r) = 2N \sum_r p_r (\ln p_r - \ln \hat{\pi}_r). \quad (1.20)$$

Alternatively, one can use the goodness-of-fit test statistic

$$\chi_{\text{GF}}^2 = \sum_r [(n_r - N\hat{\pi}_r)^2 / (N\hat{\pi}_r)] = N \sum_r (p_r - \hat{\pi}_r)^2 / \hat{\pi}_r. \quad (1.21)$$

Both statistics (1.20) and (1.21) have the same asymptotic distribution under H_0 .

In principle, these tests are possible to use with full information maximum likelihood (FIML). They cannot be used with the pseudo-likelihood approach because this does not maximize an overall likelihood function, so the $\hat{\pi}_r$ are not directly computed. In practice, however, these tests do not work well because in real data there are often many zero and small frequencies n_r which will distort the approximation to the chi-square distribution [59].

Nevertheless, under the pseudo-likelihood estimation framework the pseudo-likelihood ratio test (PLRT) is written as

$$\chi_{\text{PLLT}}^2 = 2 \times (pl(\hat{\omega}; \mathbf{y}) - pl(\tilde{\omega}; \mathbf{y})), \quad (1.22)$$

where $pl(\hat{\omega}; \mathbf{y})$ and $pl(\tilde{\omega}; \mathbf{y})$ are the log pseudo-likelihood values under the alternative and null hypothesis respectively.

It has been shown that the asymptotic distribution of the composite likelihood (pseudo-likelihood) ratio statistic is a weighted sum of χ_1^2 distribution [29, 43, 49,

56, 64]. We leave the development and studying of the performance of PLRT for testing overall fit and nested association models for future research.

1.7.2 Composite Likelihood Model Selection Criteria

Based on the results of [65], the Akaike pseudo-likelihood (PL) information criterion, AIC_{PL} for the CL framework is defined as:

$$AIC_{PL} = -pl(\hat{\omega}_{PL}; \mathbf{y}) + tr(\hat{J}(\hat{\omega}_{PL})\hat{H}^{-1}(\hat{\omega}_{PL})), \quad (1.23)$$

and, based on the results found in [27], the PL Bayesian information criterion, BIC_{PL} , is defined as:

$$BIC_{PL} = -2pl(\hat{\omega}_{PL}; \mathbf{y}) + tr(\hat{J}(\hat{\omega}_{PL})\hat{H}^{-1}(\hat{\omega}_{PL})) \times \log N, \quad (1.24)$$

where $\hat{\omega}_{PL}$ is the pseudo-likelihood estimate under the hypothesized model, and $tr(\hat{J}(\hat{\omega}_{PL})\hat{H}^{-1}(\hat{\omega}_{PL}))$ defines the number of effective parameters. The model with the smallest AIC_{PL} or BIC_{PL} is selected.

1.8 Example

The data used here, the DASS data (retrieved July, 2020 from OpenPsychometrics.org), consist of responses collected during the period of 2017–2019 to 42 items, and of the 38,776 respondents, only a random sample of 1000 were used in this example. The items were presented online to respondents in a random order. The items included in the DASS data are from scales designed to measure depression (d1–d14), anxiety (a1–a13), and stress (s1–s15). For each item, respondents were asked to consider the last week when making their responses using the following categories:

1. Did not apply to me at all
2. Applied to me to some degree, or some of the time
3. Applied to me to a considerable degree, or a good part of the time
4. Applied to me very much, or most of the time

The items are given in the appendix and in the online supplemental material, along with the data and **R** code used to fit the models to the data.

We used pseudo-likelihood estimation in this example with a relatively strong convergence criterion. We deem that a model has converged if the item with the largest change in the maximum likelihood between iterations is less than $1e - 6$, which also yields changes in many parameters on the order of $1e - 10$. The convergence information is given in Table 1.3, along with the number of iterations

Table 1.3 Global summary statistics and convergence information for models fit to the DASS data

Model	M	# of params	MLPL	Fit statistics		Convergence	
				AIC_{PL}	BIC_{PL}	Criterion	#iter
Independence	1	126	-56,146	56,272	113,162	0	5
Rasch	1	127	-44,609	44,736	90,095	0	6
GPCM	1	168	-44,240	44,408	89,641	2.5e-07	14
Nominal	1	252	-44,069	44,321	89,879	1.6e-07	14
Rasch	3	132	-42,529	42,661	85,969	4.4e-07	6
GPCM	3	171	-42,258	42,429	85,698	3.6e-07	14
Nominal	3	255	-42,030	42,285	85,822	2.5e-07	15

("#iter"), and the value of the convergence criterion. The Rasch and independence models were only fitted once; therefore, we report the convergence information from the `mlogit` output from the stacked regression. Parameters' estimates were close to the final estimations in approximately 5 iterations and the algorithm achieves convergence in less than or equal to 15 iterations. The x_j values for each item for the Rasch and GPCM models and the starting values for the v_{ijm} parameters for the nominal model were -0.1035098 , -0.03450328 , 0.03450328 , and 0.1035098 ; that is, they sum to zero and are equally spaced.

An independence log-linear model was fitted to the data as a baseline model. One- and three-dimensional models corresponding to Rasch, GPCM, and nominal models were fitted to the data. The ordinal nature of the response scale is explicitly incorporated in the Rasch and GPCM models by fixing the category scores to have the same order as the response scale and be equally spaced for successive categories. Category scale values in the nominal model are estimated and their order and spacing is not restricted. Table 1.3 contains basic summary statistics for each model, including the number of unique parameters estimated ('# of params), the maximum of the log of the pseudo-likelihood (MLPL) function, and pseudo-likelihood information criteria, AIC_{pl} and BIC_{pl} (smaller is better). As expected, the unidimensional models fit considerably worse than the three-dimensional models and will not be considered further. Among the three-dimensional models, the Rasch model is not selected whether using the AIC (which tends to select more complex models) or the BIC (which tends to select simpler models). The $M = 3$ nominal model has the smallest AIC_{pl} and the GPCM has the smallest BIC_{pl} . We will further study the results of the three-dimensional nominal and GPCM models.

1.8.1 Measures of Item Fit for the DASS Data

With high-dimensional tables, measures of fit such as D and the percent of association are not useful. The dissimilarity index requires computing fitted values for each possible response pattern (cell of a table). As we see from Fig. 1.6, as

the number of items increases, the strength of associations decreases, which makes using the percent of association accounted for by a model problematic. As a result, in the case of a large number of variables, alternative methods for evaluating the fit need to be employed.

The analyses in this section are a combination of statistics and graphics at the item level. Table 1.4 presents the maximum of the likelihoods for each item from fitting models to each item in the PLE algorithm. These are given for the nominal model and GPCM along with the differences between the models' values. These differences (i.e., Δ or -2Δ) do not meet the regularity conditions for these to be chi-square distributed because the values of the predictor variables are different for the GPCM and nominal models (i.e., different data). However, Δ still provides information regarding which models are better fitting particular items. The sum over items of the maximum likelihoods in Table 1.4 equals a model's MLPL. Table 1.5 further summarizes the item fit statistic and contains the proportion of items within a scale that fall within ranges of the maximum likelihood values. From Tables 1.3 and 1.5, in general the items from the nominal model have larger values than items fit by the GPCM (larger is better). Furthermore, the depression items tend to be fitted better than the items from the anxiety scale, and the items from the stress scale are the worst fit. Based on the difference in Table 1.4, some items appear to be fitted equally well by the GPCM and nominal model. In particular, the Δ s for d14, a4, and a9 equal 1.85, 1.37, and 1.77, respectively; however, items d6, d7, and a10 all have the largest Δ values, which suggests that the nominal model should be used (at least for these items).

The difference between the GPCM and nominal models is that the former has linear restrictions on the scale values. To determine whether this restriction is reasonable, we first examine statistics and then graphics. For the nominal model, a measure of how strongly an item is related to the latent variable that it is directly related to is η_i [4]

$$\eta_{im} = \sqrt{\sum_j v_{ijm}^2}$$

When the location identification constraint is $\sum_j v_{ijm} = 0$, η_{im} is proportional to the standard deviation of v_{ijm} s. Alternatively, we can fit the GPCM model to the data and examine the \hat{a}_{im} parameters, which, when v_{ijm} are equally spaced, will be highly correlated with η_{im} s. In our example, $r(\eta_{im}, \hat{a}_{im}) = 0.996$, which suggests that the v_{ijm} may be equally spaced and the x_j s used to fit the GPCM model are reasonable for the data. Computing η_{im} or a_{im} only requires fitting one model. The η_{im} s and \hat{a}_{im} s are given in Table 1.6. Whether using η_{im} or \hat{a}_{im} , the items that are most strongly related to their respective latent traits are d4, d7, and a10, which indicate that both models are identifying the same items and are highly related to the latent variable and therefore to each other. These statistics indicate the magnitude of association between items within a scale. For example, among the depression items, the relationship between d4 ("I felt sad and depressed") and d7 ("I

Table 1.4 Item statistics: Values of maximums of the likelihoods (components of the PLML) for each item and the nominal and GPCM, and the differences (Δ) between the log-likelihoods

Item	GPCM	Nominal	Δ	Item	GPCM	Nominal	Δ	Item	GPCM	Nominal	Δ
d1	-964.52	-957.28	7.23	a1	-1197.71	-1193.60	4.12	s1	-983.09	-974.35	8.74
d2	-1062.83	-1058.89	3.94	a2	-982.74	-974.77	7.97	s2	-1051.49	-1048.81	2.68
d3	-910.01	-905.11	4.90	a3	-973.90	-964.58	9.32	s3	-1041.61	-1036.94	4.67
d4	-844.14	-839.26	4.88	a4	-1010.04	-1008.67	1.37	s4	-1016.15	-1013.58	2.57
d5	-976.96	-969.36	7.60	a5	-964.33	-956.77	7.56	s5	-1002.18	-995.89	6.28
d6	-891.23	-878.79	12.44	a6	-1097.10	-1092.13	4.98	s6	-1163.45	-1161.15	2.30
d7	-806.19	-795.27	10.93	a7	-1014.66	-1008.84	5.82	s7	-1141.78	-1136.96	4.82
d8	-1009.82	-1001.83	7.99	a8	-815.02	-809.09	5.92	s8	-1020.17	-1014.21	5.96
d9	-969.42	-966.66	2.76	a9	-1118.39	-1116.62	1.77	s9	-1037.74	-1032.84	4.90
d10	-1028.59	-1024.36	4.23	a10	-968.84	-956.47	12.37	s10	-1046.11	-1042.29	3.82
d11	-884.02	-877.86	6.16	a11	-985.39	-977.32	8.07	s11	-1090.11	-1086.84	3.27
d12	-944.48	-936.61	7.87	a12	-1090.40	-1087.02	3.38	s12	-1103.96	-1101.73	2.23
d13	-872.29	-865.68	6.61	a13	-1012.65	-1006.15	6.50	s13	-1010.43	-1008.02	2.41
d14	-1068.67	-1066.82	1.85					s14	-1072.32	-1069.73	2.59
								s15	-1013.22	-1011.09	2.12

Table 1.5 Summary of proportion of items fitted in terms of ranges of the values of the items’ maximum of the log-likelihoods

Model	Scale	Range of log-likelihoods			
		> -799	-800 to -899	-900 to -999	< -1000
Nominal	Depression	0.07	0.29	0.36	0.29
	Anxiety	0.00	0.08	0.38	0.54
	Stress	0.00	0.00	0.13	0.87
GPCM	Depression	0.00	0.36	0.36	0.29
	Anxiety	0.00	0.08	0.38	0.54
	Stress	0.00	0.00	0.07	0.93

Table 1.6 Item statistics:
The slopes \hat{a}_{im} from the GPCM and the η_{im} statistics from the nominal models, which reflect the strength of the relationship between the items and the latent variables, as well as between items themselves

Item	a_{i1}	η_{i1}	Item	a_{i2}	η_{i2}	Item	a_{i3}	η_{i3}
d1	4.77	0.77	a1	2.14	0.34	s1	5.03	0.83
d2	3.59	0.57	a2	4.21	0.68	s2	4.20	0.66
d3	5.71	0.89	a3	4.17	0.64	s3	4.21	0.66
d4	6.60	1.05	a4	5.67	0.88	s4	4.70	0.74
d5	4.63	0.74	a5	3.65	0.61	s5	4.46	0.73
d6	5.62	0.91	a6	2.41	0.37	s6	2.62	0.41
d7	7.45	1.19	a7	5.24	0.81	s7	2.96	0.47
d8	4.29	0.69	a8	3.67	0.63	s8	4.58	0.72
d9	4.73	0.74	a9	3.28	0.51	s9	4.21	0.68
d10	4.00	0.63	a10	6.00	0.94	s10	4.24	0.68
d11	5.80	0.89	a11	5.48	0.85	s11	3.66	0.57
d12	5.03	0.78	a12	4.10	0.64	s12	3.37	0.52
d13	5.56	0.86	a13	3.79	0.58	s13	4.88	0.76
d14	3.32	0.54				s14	3.72	0.59
						s15	4.35	0.67

felt that life wasn’t worthwhile”) is larger than that from any other two items, and the smallest is between a1 (“I was aware of dryness of my mouth”) and a6 (“I felt scared without any good reason”). Comparing across scales, it appears that the items on the depression scale are more highly related to the depression trait and between each other (mean $\hat{a}_i = 5.08$), followed by anxiety items related to the anxiety trait (mean $\hat{a} = 4.14$). The least strongly related to the latent trait and among each other are the stress items (mean $\hat{a} = 4.08$).

To further investigate whether the GPCM or nominal models are better for particular items, we examine the scale value estimates from the nominal model to see if they are indeed linear with respect to equally spaced numbers. Estimated scale values from the nominal model (solid circles and lines) can be plotted against integers with linear regression drawn (dashed lines) in the same plot. Examples for four items are given in Fig. 1.7. The categories for all items are clearly ordinal and increase with values of the integers. The values for aggression item a9 (upper left) are coincident with the regression line, which from Table 1.4 has $\Delta = 1.77$. The

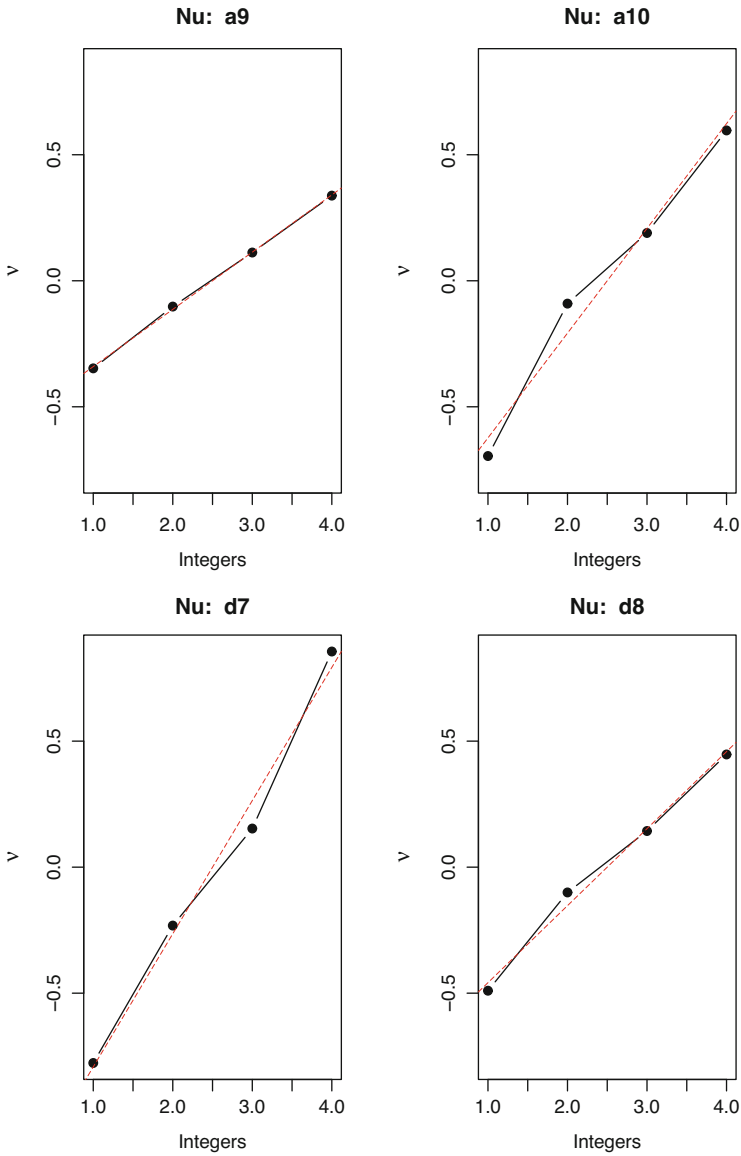


Fig. 1.7 Estimated scale values v_{ijm} (solid points and black lines) from the nominal response model for two aggression items (top) and two depression items (bottom) plotted against integers with linear regression lines (dashed lines)

scale values for the other items a10, d7, and d8, deviate from their regression lines and had Δ values of 12.37, 10.93, and 7.99, respectively. Scale values for item a9 and possibly item d8 might be satisfactorily modeled using equally spaced category scores as in the GPCM. Another aspect to consider is slope of the lines, which would correspond to a_{im} in a GPCM model. Among these four items, a9 (smallest slope) appears to be more weakly related to its the latent variable; whereas, item d7 (the steepest slope) is the most strongly related to its latent variable. These results further confirm our conclusions based on statistics in Table 1.6.

The last analyses look at the correspondence between data and fitted values (probabilities). In logistic regression with continuous predictors and in IRT, the continuous values can be collapsed into groups or bins. Estimates of θ_m were computed using (1.10) and then grouped into 10 categories. The observed proportions who select a category within a group and the fitted probabilities for the group were plotted against the mean of the continuous $\hat{\theta}_m$ for the groups. Two examples of such plots are given in Fig. 1.8 where the data are points and lines are fitted values from the nominal model (one line per category). Item d7 has the largest log-likelihood value in Table 1.4 and the largest η_{im} and \hat{a}_{im} in Table 1.6. This appears to be the item fitted best according to our statistics and there is a close correspondence between the fitted probabilities (lines) and the observed proportions (points). Item s6 has the smallest log-likelihood and one of the smallest η_i and \hat{a}_i , which indicate that this item has the worst fit under the model. The model for item s6 underpredicts the first (squares) category and last (diamonds) category, which further confirms that this item is not fitted well by the nominal model. Item s6 is not fitted well by the nominal model and will not fare any better under a GPCM.

Computing $\hat{\theta}_m$ using (1.10) makes use of responses to all items where items were weighted by the conditional covariances. Since we set $\sigma_{mm} = 1$ for identification, we actually estimated conditional correlation matrices between traits within response patterns. These estimated conditional correlation matrices from the nominal and GPCM models are very similar,

$$\hat{\Sigma}_{nom} = \begin{pmatrix} 1.000 & 0.038 & 0.094 \\ 0.038 & 1.000 & 0.290 \\ 0.094 & 0.290 & 1.000 \end{pmatrix} \quad \text{and} \quad \hat{\Sigma}_{gpc} = \begin{pmatrix} 1.000 & 0.047 & 0.099 \\ 0.047 & 1.000 & 0.299 \\ 0.099 & 0.299 & 1.000 \end{pmatrix},$$

where the subscript *nom* is for the nominal model and *gpc* is for the GPCM. The conditional correlations between depression and anxiety and between depression and stress appear relatively small.

Small conditional correlations do not imply that the marginal correlations are small. The marginal correlation matrices between $\hat{\theta}_m$ from the nominal and GPCM model are

$$\hat{R}_{nom} = \begin{pmatrix} 1.000 & 0.774 & 0.827 \\ 0.774 & 1.000 & 0.954 \\ 0.827 & 0.954 & 1.000 \end{pmatrix} \quad \text{and} \quad \hat{R}_{gpc} = \begin{pmatrix} 1.000 & 0.781 & 0.830 \\ 0.781 & 1.000 & 0.957 \\ 0.830 & 0.957 & 1.000 \end{pmatrix}.$$

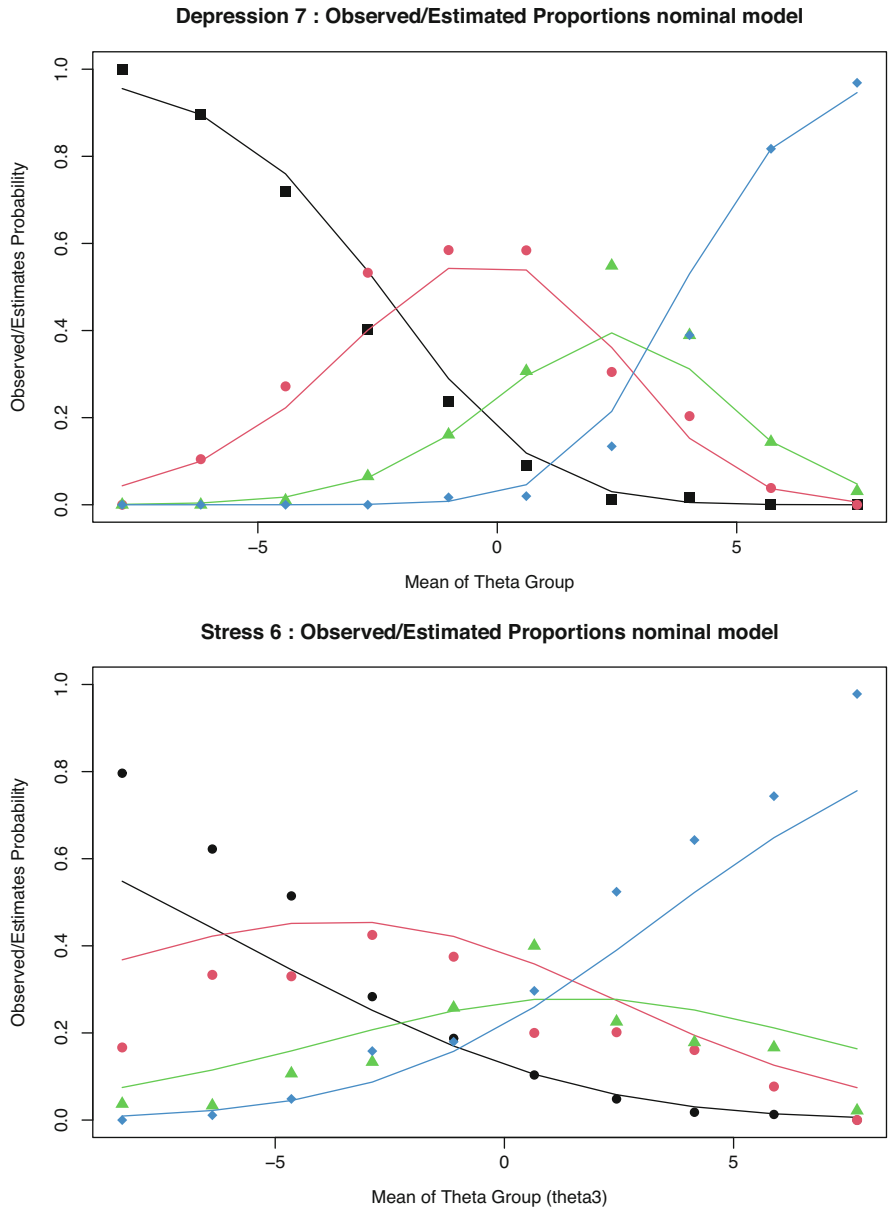


Fig. 1.8 For depression item d7 (top) and stress item s6 (bottom), observed proportions (points) and fitted probabilities (lines) are plotted against the mean of θ_m where estimate of $\hat{\theta}_m$ has been collapsed in to groups. Symbols for categories are 1 = squares, 2 = dots, 3 = triangles, and 4 = diamonds

Table 1.7 Alternative models for DASS data with $M = 2$ where stress and anxiety are one scale, and $M = 3$ where $\sigma_{12} = 0$

Model	M	#param	MLPL	Fit statistics		Convergence	
				AIC_{PL}	BIC_{PL}	criteria	#iter
Rasch	2	129	-42,840	42,969	86,571	4.4e-07	6
GPCM	2	169	-42,490	42,659	86,147	3.2e-07	13
Nominal	2	253	-42,265	42,518	86,278	5.1e-07	13
Rasch	3	131	-42,536	42,667	85,983	4.2e-07	6
GPCM	3	170	-42,268	42,438	85,711	3.2e-07	14
Nominal	3	254	-42,036	42,290	85,827	6.7e-07	15

The order in terms of magnitude of the marginal and conditional correlations have the same pattern (i.e., largest is for anxiety and stress, and the smallest is for depression and anxiety), but are considerably larger than the conditional values.

The large marginal correlations between anxiety and stress suggest that perhaps these are not distinct constructs and a two dimensional model maybe sufficient. The data were reanalyzed using a two-dimensional Rasch, GPCM, and nominal model, which each has 2 fewer parameters. The statistics for these models are reported in Table 1.7, but the new models fitted the data worse than our original three-dimensional models (i.e., original models have smaller AIC_{PL} and BIC_{PL}). These results occurred because the conditional correlations (i.e., .290 and .299) are relatively small. It is important to point out that we do not set the marginal correlations, but rather the conditional correlations (or covariances). If a conditional correlation is close to 1, then a two-dimensional model might be better than a three-dimensional one.

In conclusion, our analysis confirmed our conjecture that the items represent three correlated constructs, as well as the excepted ordering of the category scores. For some items, the relative spacing between items is roughly equal but not for all, which suggests that the nominal model is the best model. We also detected some items that did not fit the data very well (e.g., s6). The items on the depression scale are more closely related to the latent variable of depression, and thus they are also more closely related to each other. The stress items have weaker association with the stress latent variable and also have weaker associations between the stress items themselves. Due to small values of $\hat{\sigma}_{mm'}$ for depression and anxiety and for depression and stress, the estimated value of depression depends mostly on the responses to the depression items. On the other hand, the larger value of $\hat{\sigma}_{mm'}$ for stress and anxiety indicate that each provide more information in the estimating of values on the stress and anxiety constructs.

1.9 Conclusion/Discussion

When there are interactions between categorical variables, the AMs presented in this chapter are just one way to describe the nature and strength of associations. Other possibilities not covered in this chapter, and often missing in the literature on AMs, include (multiple) correspondence analysis [34, 35], optimal scaling [53], and dual scaling [54, 55], which are all scaling methods that in the case of 2-way tables are all essentially the same and yield very similar results. For a history of these methods see [53]. These scaling methods are data analytic techniques without distributional assumptions and statistical tests of model goodness-of-fit to the data. Other related methods that are statistical models are canonical correlation models [32] and latent class models. For 2-way tables, latent class models with 2 latent classes and the correlation models yield similar results.

The AMs discussed here provide useful representations of interactions between categorical variables; however, they have also been derived from an underlying theoretical model (e.g., IRT). Although we focus on models with an underlying simple structure, the log-multiplicative AMs afford more complex structures, including models where items load on multiple latent variables in a more exploratory analysis and bi-factor structures. We can also add covariates to the AMs. Pseudo-likelihood estimation can be used for these more complex structures.

The data analysis examples illustrate how measures of fit, such as item log-likelihood differences, transformed scores, and fitted proportions, can be used to check item misfit and the strength of an item in measuring a latent variable. Furthermore, AMs provide information about the arbitrary selected scores when choosing the response categories of an item. This is also what IRT modeling tries to achieve by estimating discrimination coefficients for each item or each score in the nominal case. Goodness-of-fit tests and model selection criteria can be developed under the pseudo-likelihood estimation framework presented here to test overall fit and select among nested and non-nested models.

The connection between IRT and association models provides multiple insights on the same data analysis problem, i.e. on how to model and interpret associations depending on the aim of our analysis. The availability of statistical software and the extension to high-dimensional tables for multi-category variables is a very useful tool for data analysts who want to have the flexibility of choosing and estimating a suitable model for high-dimensional data.

Appendix: DASS Data

For each item, respondents were asked to consider the last week and use the rating scale:

1. Did not apply to me at all.
2. Applied to me to some degree, or some of the time.

3. Applied to me to a considerable degree, or a good part of the time.
4. Applied to me very much, or most of the time.

Depression Scale

- d1. I couldn't seem to experience any positive feeling at all.
- d2. I just couldn't seem to get going.
- d3. I felt that I had nothing to look forward to.
- d4. I felt sad and depressed.
- d5. I felt that I had lost interest in just about everything.
- d6. I felt I wasn't worth much as a person.
- d7. I felt that life wasn't worthwhile.
- d8. I couldn't seem to get any enjoyment out of the things I did.
- d9. I felt down-hearted and blue.
- d10. I was unable to become enthusiastic about anything.
- d11. I felt I was pretty worthless.
- d12. I could see nothing in the future to be hopeful about.
- d13. I felt that life was meaningless.
- d14. I found it difficult to work up the initiative to do things.

Anxiety Scale

- a1. I was aware of dryness of my mouth.
- a2. I experienced breathing difficulty (eg, excessively rapid breathing, breathlessness in the absence of physical exertion).
- a3. I had a feeling of shakiness (eg, legs going to give way).
- a4. I felt that I was using a lot of nervous energy.
- a5. I had a feeling of faintness.
- a6. I perspired noticeably (eg, hands sweaty) in the absence of high temperatures or physical exertion.
- a7. I felt scared without any good reason.
- a8. I had difficulty in swallowing.
- a9. I was aware of the action of my heart in the absence of physical exertion (eg, sense of heart rate increase, heart missing a beat).
- a10. I felt I was close to panic.
- a11. I felt terrified.
- a12. I was worried about situations in which I might panic and make a fool of myself.
- a13. I experienced trembling (eg, in the hands).

Stress Scale

- s1. I found myself getting upset by quite trivial things.
- s2. I tended to overreact to situations.
- s3. I found it difficult to relax.
- s4. I found myself in situations that made me so anxious I was most relieved when they ended.
- s5. I found myself getting upset rather easily.

- s6. I found myself getting impatient when I was delayed in any way (eg, elevators, traffic lights, being kept waiting).
- s7. I felt that I was rather touchy.
- s8. I found it hard to wind down.
- s9. I found that I was very irritable.
- s10. I found it hard to calm down after something upset me.
- s11. I feared that I would be thrown off by some trivial but unfamiliar task.
- s12. I found it difficult to tolerate interruptions to what I was doing.
- s13. I was in a state of nervous tension.
- s14. I was intolerant of anything that kept me from getting on with what I was doing.
- s15. I found myself getting agitated.

References

1. Anderson, C.J.: Network multidimensional item response models: Beyond simple structure (2017). Paper, International Meeting of the Psychometric Society, Zurich, Switzerland
2. Anderson, C.J., Li, Z., Vermunt, J.K.: Estimation of models in a Rasch family for polytomous items and multiple latent variables. *J. Stat. Softw.* **20**, 1–36 (2007)
3. Anderson, C.J., Verkuilen, J.V., Peyton, B.: Modeling polytomous item responses using simultaneously estimated multinomial logistic regression models. *J. Educ. Behav. Stat.* **35**, 422–452 (2010)
4. Anderson, C.J., Vermunt, J.K.: Log-multiplicative association models as latent variable models for nominal and/or ordinal data. *Sociol. Methodol.* **30**, 81–121 (2000)
5. Anderson, C.J., Yu, H.T.: Log-multiplicative association models as item response models. *Psychometrika* **72**, 5–23 (2007)
6. Anderson, C.J., Yu, H.T.: Theoretical and empirical properties of log-multiplicative association models as multidimensional nominal item response models (2021). Manuscript
7. Anderson, J.A.: Regression and ordered categorical variables. *J. R. Stat. Soc. B* **46**, 1–30 (1984)
8. Arnold, B.C., Straus, D.: Pseudolikelihood estimation: some examples. *Indian J. Stat.* **53**, 233–243 (1991)
9. Becker, M.: On the bivariate normal distribution and association models for ordinal categorical data. *Stat. Probab. Lett.* **8**, 435–440 (1989)
10. Besag, J.: Spatial interaction and the statistical analysis of lattice systems. *J. R. Stat. Soc. B* **36**, 192–225 (1974)
11. Besag, J.: Statistical analysis of non-lattice data. *journal of the royal statistical society. J. R. Stat. Soc. D (The Statistician)* **24**, 179–195 (1975)
12. Bhat, S., Anderson, C.J., Crues, W., Angrave, L., Shaik, N., Hendricks, G.G.: Know your audience: who is served and their engagement levels in MOOCs (2020). Manuscript
13. Bouchet-Valat, M., Turner, H., Friendly, M., Lemon, J., Csardi, C.: Package ‘logmult’ (2020). R package version 0.7.21
14. Chang, H.H.: The asymptotic posterior normality of the latent trait for polytomous IRT models. *Psychometrika* **58**, 445–463 (1986)
15. Chang, H.H., Stout, W.: The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika* **58**, 37–52 (1003)
16. Chen, Y., Li, X., Liu, J., Ying, Z.: Robust measurement via a fused latent variable and graphical item response theory model. *Psychometrika* **85**, 538–562 (2018)

17. Croissant, Y.: Estimation of random utility models in R: The `mlogit` package. *J. Stat. Softw.* **95**(11), 1–41 (2020)
18. de Rooij, M.: The analysis of change, Newton’s law of gravity and association models. *J. R. Stat. Soc. Stat. Soc A* **171**, 137–157 (2007)
19. de Rooij, M.: Ideal point discriminant analysis revisited with a special emphasis on visualization. *Psychometrika* **74**, 317–330 (2009)
20. de Rooij, M., Heiser, W.: Graphical representations and odds ratios in a distance-association model for the analysis of cross-classified data. *Psychometrika* **70**, 99–122 (2005)
21. de la Torres, J., Song, H., Hong, Y.: Comparison of four methods of IRT subscore. *Appl. Psychol. Meas.* **35**, 296–316 (2001)
22. Edwards, D.: *Introduction to Graphical Models*, 2nd edn. Springer, New York (2000)
23. Eeuwijk, F.V.: Multiplicative interaction in generalized linear models. *Biometrics* **51**, 1017–1032 (1995)
24. Elff, M.: Multinomial logit models, with or without random effects or overdispersion. *J. Stat. Softw.* (2020)
25. Epskamp, S.: Package ‘Ising Sampler’ (2020). R package version 0.2.1
26. Fienberg, S.E., Rinaldo, A.: Maximum likelihood estimation in log-linear models. *Ann. Stat.* **40**, 996–1023 (2012)
27. Gao, X., Song, P.X.K.: Composite likelihood Bayesian information criteria for model selection in high dimensional data. *J. Am. Stat. Assoc.* **105**(492), 1531–1540 (2010)
28. Gelman, A., Speed, T.P.: Characterizing a joint distribution by conditionals. *J. R. Stat. Assoc. B* **55**, 185–188 (1993)
29. Geys, H., Molenberghs, G., Ryan, L.M.: Pseudolikelihood modeling in multivariate outcomes in developmental toxicology. *J. Am. Stat. Assoc.* **94**, 734–745 (1999)
30. Goodman, L.A.: Simple models for the analysis of association in cross-classifications having ordered categories. *J. Am. Stat. Assoc.* **74**(367), 537–552 (1979)
31. Goodman, L.L.: Association models and the bivariate normal for contingency tables with ordered categories. *Biometrika* **68**, 347–355 (1981)
32. Goodman, L.L.: The analysis of cross-classified data having ordered and/or unordered categories: association models, correlation models, and asymmetry models for contingency tables with or without missing entries. *Ann. Stat.* **75**, 1–24 (1985)
33. Goodman, L.L.: Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables. *Int. Stat. Rev.* **54**, 243–270 (1986)
34. Greenacre, M.: *Correspondence Analysis in Practice*. Chapman & Hall/CRC Interdisciplinary Statistics. CRC Press, Boca Raton (2017)
35. Greenacre, M., Blasius, J.: *Multiple Correspondence Analysis and Related Methods*. Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences. CRC Press, Boca Raton (2006)
36. Hasan, A., Wang, Z., Mahani, A.S.: Fast estimation of multinomial logit models: R package `mnlogit`. *J. Stat. Softw.* **75**, 1–24 (2016)
37. Hessen, D.J.: Fitting and testing conditional multinomial partial credit models. *Psychometrika* **77**, 693–709 (2012)
38. Holland, P.W.: The Dutch identity: a new tool for the study of item response models. *Psychometrika* **55**, 5–18 (1990)
39. Johnson, T.R.: Discrete choice models for ordinal response variables: a generalization of the stereotype models. *Psychometrika* **72**, 489–504 (2007)
40. Jöreskog, K.G.: A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika* **34**, 183–202 (1969)
41. Jöreskog, K.G.: Structural equation models in the social sciences: specification, estimation and testing. In: Jöreskog, K.G., Sörbom, D. (eds.) *Advances in Factor Analysis and Structural Equation Models*, pp. 105–127. Abt Books, Cambridge (1979)

42. Kateri, M.: *Contingency Table Analysis: Methods and Implementation Using R*. Birkhäuser/Springer, New York (2014)
43. Katsikatsou, M., Moustaki, I.: Pairwise likelihood ratio tests and model selection criteria for structural equation models with ordinal variables. *Psychometrika* **81**, 1046–1068 (2016)
44. Kelderman, H.: Multidimensional Rasch models for partial-credit scoring. *Appl. Psychol. Meas.* **20**, 1–10 (1996)
45. Kelderman, H., Rijkens, C.P.M.: Loglinear multidimensional IRT models for polytomous scored items. *Psychometrika* **59**, 149–176 (1994)
46. Kornely, M., Kateri, M.: Asymptotic posterior normality of multivariate latent traits in an IRT model. *Psychometrika* **87**(3), 1146–1172 (2022)
47. Kruis, J., Maris, G.: Three representations of the Ising model. *Sci. Rep.* **6**, 1–10 (2016)
48. Lauritzen, S.L.: *Graphical Models*. Oxford, Oxford (1996)
49. Liang, K.Y., Self, S.G.: On the asymptotic behaviour of the pseudolikelihood ratio test statistic. *J. R. Stat. Soc. B* **58**, 785–796 (1996)
50. Lindsay, B.G.: Composite likelihood methods. In: Prabhu, N.U. (ed.) *Statistical Inference from Stochastic Processes*, pp. 221–239. American Mathematical Society, Providence (1988)
51. Liu, Q., Ihler, A.: Distributed parameter estimation via pseudo-likelihood. In: 29th International Conference on Machine Learning (Edinburgh, Scotland, UK, 2012, International Machine Learning Society) (2012)
52. Marsman, M., Borsboom, D., Kruis, J., Epskamp, S., van Bork, R., Waldorp, L., van der Maas, H., Maris, G.: An introduction to network psychometrics: relating Ising network models to item response theory models. *Multivar. Behav. Res.* **53**, 15–35 (2018)
53. Muelman, J., Kooij, A.J.V.D., Heiser, W.: Principal components analysis with nonlinear optimal scaling transformations for ordinal and nominal data. In: Kaplan, D. (ed.) *The SAGE Handbook of Quantitative Methodology for the Social Sciences*, chap. 3, pp. 49–72. SAGE Publications, Thousand Oaks (2004)
54. Nishisato, S.: Dual scaling. In: Kaplan, D. (ed.) *The SAGE Handbook of Quantitative Methodology for the Social Sciences*, chap. 1, pp. 2–24. SAGE Publications, Thousand Oaks (2004)
55. Nishisato, S.: *Elements of Dual Scaling*. Psychology Press (2004)
56. Pace, L., Salvani, A., Sartori, N.: Adjusting composite likelihood ratio statistics. *Stat. Sin.* **21**, 129–148 (2011)
57. Paek, Y.: Pseudo-likelihood estimation of multidimensional item response theory model. Ph.D. Thesis, University of Illinois, Urbana-Champaign (2016)
58. Paek, Y., Anderson, C.J.: Pseudo-likelihood estimation of multidimensional response models: polytomous and dichotomous items. In: van der Ark, A., Wiberg, M., Culpepper, S.A., Douglas, J.A., Wang, W.C. (eds.) *Quantitative Psychology—The 81st Annual Meeting of the Psychometric Society*, pp. 21–30. Springer, New York (2017)
59. Reiser, M., Vandenberg, M.: Validity of the chi-square test in dichotomous variable factor analysis when expected frequencies are small. *Br. J. Math. Stat. Psychol.* **47**, 85–107 (1994)
60. Rom, D., Sarkar, S.K.: Approximating probability integrals of multivariate normal using association models. *J. Stat. Comput. Simul.* **35**(1–2), 109–119 (1990)
61. Takane, Y., Bozdogan, H., Shibayama, T.: Ideal point discriminant analysis. *Psychometrika* **52**, 371–392 (1987)
62. Team, R.C.: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2017).
63. Turner, H., Firth, D.: Generalized nonlinear models in R: an overview of the GNM package (2020). R package version 1.1-1
64. Varin, C., Reid, N., Firth, D.: An overview of composite likelihood methods. *Stat. Sin.* **21**, 5–42 (2011)
65. Varin, C., Vidoni, P.: A note on composite likelihood inference and model selection. *Biometrika* **92**, 519–528 (2005)
66. Wang, W.C., Chen, P.H., Cheng, Y.Y.: Improving measurement precision of test batteries using multidimensional item response models. *Psychol. Methods* **9**, 116–136 (2004)

67. Wang, Y.J.: The probability intergrals of bivarite normal distributions: a contingency table approach. *Biometrika* **74**, 185–190 (1987)
68. Wang, Y.J.: Multivariate normal integrals and contingency tables with ordered categories. *Psychometrika* **62**, 267–284 (1997)
69. Yee, T.W.: *VGAM: Vector Generalized Linear and Additive Models* (2020)

Chapter 2

Graphical Models for Categorical Data



Peter W. F. Smith

2.1 Introduction

Graphical models are parametric statistical models for multivariate random variables. In these models, the relationships between the variables are displayed using a mathematical graph. The use of mathematical graphs in statistics dates back to the path diagrams of Wright [20, 21], but it was not until the seminal paper of Darroch et al. [10] that a way of constructing a graph which has a well-defined probabilistic interpretation was proposed. This graph has since been called the conditional independence graph or independence graph, for short.

The (conditional) independence structure of a p -dimensional random vector \vec{X} , whose density is positive for all points in the sample space, can be displayed using a graph $\mathcal{G} = (V, E)$, where the set of vertices, $V = \{1, \dots, p\}$, contains one vertex for each of the p components of the random vector and an edge (i, j) is not in the edge set E if, and only if, the corresponding elements of \vec{X} , X_i and X_j are conditional independent of the remaining $p - 2$ components. More succinctly, using the notation of Dawid [11],

$$(i, j) \notin E \iff X_i \perp\!\!\!\perp X_j \mid \vec{X}_R,$$

P. W. F. Smith (✉)

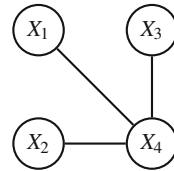
Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, UK

e-mail: P.W.Smith@soton.ac.uk

© Springer Nature Switzerland AG 2023

M. Kateri, I. Moustaki (eds.), *Trends and Challenges in Categorical Data Analysis*, Statistics for Social and Behavioral Sciences, https://doi.org/10.1007/978-3-031-31186-4_2

Fig. 2.1 Example of an independence graph



where in general $\vec{X}_A = \{X_k : k \in A\}$ and here $R = V \setminus (i, j)$. For example, if a 4-dimensional random vector \vec{X} satisfies the following conditional independencies:

$$X_1 \perp\!\!\!\perp X_2 | \{X_3, X_4\}, \quad X_1 \perp\!\!\!\perp X_3 | \{X_2, X_4\} \quad \text{and} \quad X_2 \perp\!\!\!\perp X_3 | \{X_1, X_4\},$$

then \vec{X} has the independence graph presented in Fig. 2.1.

The definition of the independence graph can be extended to included variables with structural zeros, that is variables whose density is zero at some points in the sample space (see [19], page 34). However, this requires some work and in this chapter we will assume that the variables considered satisfy the positivity constraint.

Although any set of jointly distributed random variables has an independence graph, the analyst requires, in order to estimate the graph from some given data assumed to be realisations from \vec{X} , a set of models in which conditional independence can easily be parameterised. For discrete variables considered in this chapter, cross-classifying a contingency table, the log-linear model is used. (See Chap. 1 for further details about log-linear models.) However, there are graphical Gaussian models for when all the elements of \vec{X} are continuous random variables and mixed interaction models for when \vec{X} contains both continuous and discrete random variables.

The core ingredients of graphical modelling are conditional independence and graphs, and these are reviewed in the next two sections. The notation used reflects that this chapter concerns graphical models for discrete random variables. However, the ideas are relevant for graphical Gaussian and mixed interaction models. The Markov properties, which facilitate the interpretation of the association structure displayed in the independence graph, are presented in Sect. 2.4. Section 2.5 discusses graphical log-linear models where all the variables are treated on an equal footing, whereas Sects. 2.6 and 2.7 discuss directed graphical models and graphical chain models, respectively, where directed relations between the variables are considered. The chapter concludes with some suggestions for further reading.

2.2 Independence and Conditional Independence

Two discrete random vectors \vec{X}_A and \vec{X}_B , with sample spaces \mathcal{X}_A and \mathcal{X}_B , respectively, are *independent* if

$$P(\vec{X}_A = \vec{x}_A, \vec{X}_B = \vec{x}_B) = P(\vec{X}_A = \vec{x}_A)P(\vec{X}_B = \vec{x}_B),$$

for all $\vec{x}_A \in \mathcal{X}_A$ and $\vec{x}_B \in \mathcal{X}_B$. This is denoted by $\vec{X}_A \perp\!\!\!\perp \vec{X}_B$ and is equivalent to

$$P(\vec{X}_A = \vec{x}_A | \vec{X}_B = \vec{x}_B) = P(\vec{X}_A = \vec{x}_A)$$

or

$$P(\vec{X}_B = \vec{x}_B | \vec{X}_A = \vec{x}_A) = P(\vec{X}_B = \vec{x}_B),$$

for all $\vec{x}_A \in \mathcal{X}_A$ and $\vec{x}_B \in \mathcal{X}_B$. A useful result when trying to determine if two random vectors are independent is the factorisation criteria for independence [19, Proposition 2.2.1]:

$$\vec{X}_A \perp\!\!\!\perp \vec{X}_B \iff P(\vec{X}_A = \vec{x}_A, \vec{X}_B = \vec{x}_B) = g(\vec{x}_A)h(\vec{x}_B), \quad \text{for all } \vec{x}_A \text{ and } \vec{x}_B. \quad (2.1)$$

Two discrete random variables \vec{X}_B and \vec{X}_C with sample spaces \mathcal{X}_B and \mathcal{X}_C , respectively, are *conditionally independent* given the random vector \vec{X}_A with sample space \mathcal{X}_A if

$$P(\vec{X}_B = \vec{x}_B, \vec{X}_C = \vec{x}_C | \vec{X}_A = \vec{x}_A) = P(\vec{X}_B = \vec{x}_B | \vec{X}_A = \vec{x}_A)P(\vec{X}_C = \vec{x}_C | \vec{X}_A = \vec{x}_A),$$

for all $\vec{x}_B \in \mathcal{X}_B$, $\vec{x}_C \in \mathcal{X}_C$ and $\vec{x}_A \in \mathcal{X}_A$. This is denoted by $\vec{X}_B \perp\!\!\!\perp \vec{X}_C | \vec{X}_A$ and is equivalent to

$$P(\vec{X}_B = \vec{x}_B | \vec{X}_C = \vec{x}_C, \vec{X}_A = \vec{x}_A) = P(\vec{X}_B = \vec{x}_B | \vec{X}_A = \vec{x}_A),$$

$$P(\vec{X}_C = \vec{x}_C | \vec{X}_B = \vec{x}_B, \vec{X}_A = \vec{x}_A) = P(\vec{X}_C = \vec{x}_C | \vec{X}_A = \vec{x}_A)$$

or

$$\begin{aligned} &P(\vec{X}_A = \vec{x}_A, \vec{X}_B = \vec{x}_B, \vec{X}_C = \vec{x}_C) \\ &= \frac{P(\vec{X}_A = \vec{x}_A, \vec{X}_B = \vec{x}_B)P(\vec{X}_A = \vec{x}_A, \vec{X}_C = \vec{x}_C)}{P(\vec{X}_A = \vec{x}_A)}. \end{aligned}$$

A useful result when trying to determine if two random vectors are conditionally independent is the factorisation criteria for conditional independence [19, Proposition 2.2.3]:

$$\vec{X}_B \perp\!\!\!\perp \vec{X}_C | \vec{X}_A \iff P(\vec{X}_A = \vec{x}_A, \vec{X}_B = \vec{x}_B, \vec{X}_C = \vec{x}_C) = g(\vec{x}_B, \vec{x}_A)h(\vec{x}_C, \vec{x}_A), \quad (2.2)$$

for all \vec{x}_B and \vec{x}_C , and all \vec{x}_A with $P(\vec{X}_A = \vec{x}_A) > 0$.

2.3 Relevant Graph Theory

Most of the small amount of graph theory required for this chapter is presented in this section and the ideas illustrated using the graphs presented in Fig. 2.2. See [19, Section 3.1] or [15, Chapter 2] for further details concerning the graph theory relevant to graphical modelling.

A *graph* is denoted by $\mathcal{G} = (V, E)$, where V is the set of vertices and E is the edges set of the graph. Initially, only undirected graphs are considered, that is $(i, j) \in E \iff (j, i) \in E$. However, directed graphs will be introduced in Sects. 2.6 and 2.7 when directed graphical models and graphical chain models, respectively, are considered.

Vertices i and j are *neighbours/adjacent* in \mathcal{G} if $(i, j) \in E$. For example, in Fig. 2.2a vertices 1 and 2 are neighbours, but vertices 1 and 4 are not.

A *path* is a sequence of distinct vertices i_1, i_2, \dots, i_m for which the edges (i_l, i_{l+1}) , $l = 1, \dots, m - 1$ are in E . For example, in Fig. 2.2a vertices 1, 2, 4, 3 form a path.

A *cycle* is path where $i_1 = i_m$. For example, in Fig. 2.2a vertices 1, 2, 4, 3, 1 form a cycle.

A cycle is *chordless* if only successive vertices are neighbours. For example, in Fig. 2.2a vertices 1, 2, 3, 1 form a chordless cycle, but vertices 1, 2, 4, 3, 1 do not, since 2 and 3 are neighbours.

A subset of vertices A *separates* vertices $i \notin A$ and $j \notin A$ if every path joining i and j contains at least one vertex in A . For example, in Fig. 2.2a the subset $A = \{2, 3\}$ separates vertices 1 and 5.

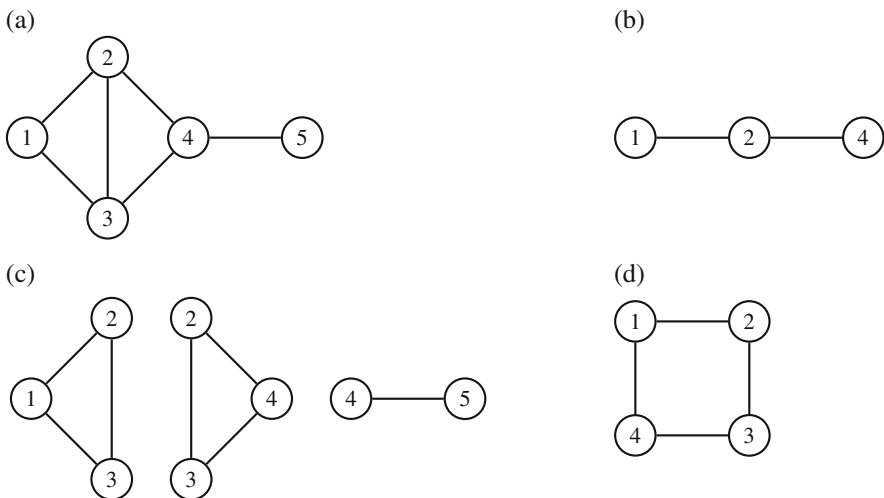


Fig. 2.2 Examples of (a) a graph, (b) a subgraph, (c) the cliques of the graph and (d) an irreducible graph

A subset of vertices A *separates* subsets of vertices B and C (A , B and C disjoint) if for every pair of vertices $i \in B$, $j \in C$ it separates i and j . For example, in Fig. 2.2a the subset $A = \{2, 3\}$ separates subsets $B = \{1\}$ and $C = \{4, 5\}$.

The *boundary* of a vertex i , $bd(i)$, is the set of all neighbours of i . For example, in Fig. 2.2a $bd(3) = \{1, 2, 4\}$.

The *subgraph* of $\mathcal{G} = (V, E)$ induced by a subset of vertices $W \subseteq V$ is the graph $\mathcal{G}_W = (W, F)$, where $(i, j) \in F$ if $i, j \in W$ and $(i, j) \in E$. For example, the subgraph of the graph in Fig. 2.2a induced by $W = \{1, 2, 4\}$ is given in Fig. 2.2b.

A subset of vertices $W \subseteq V$ induces a *complete* subgraph of $\mathcal{G} = (V, E)$ if $(i, j) \in E$ for all $i, j \in W$.

The *cliques* of a graph $\mathcal{G} = (V, E)$ are its maximally complete subgraphs, maximal with respect to inclusion of another vertex from V . A clique can be identified from its vertex set alone since, by definition, it is complete. For example, the subgraph of the graph in Fig. 2.2a induced by $W = \{1, 2\}$ is complete, but not maximally complete, since the subgraph induced by $W' = W \cup \{3\}$ is also complete. The latter subgraph is a clique, since the subgraphs induced by $W' \cup \{4\}$ and $W' \cup \{5\}$ are not complete.

If A , B and C are disjoint subsets of V such that B and C are non-empty, $A \cup B \cup C = V$, A separates B from C in $\mathcal{G} = (V, E)$ and A is complete, then the subgraphs induced by $A \cup B$ and $A \cup C$ form a *reduction* of \mathcal{G} , and \mathcal{G} is said to have been reduced. A graph or subgraph that cannot be reduced is called *irreducible*. Hence, the cliques of a graph are irreducible. For example, since for the graph in Fig. 2.2a, $A = \{2, 3\}$ induces a complete subgraph, the subgraphs induced by $A \cup B = \{1, 2, 3\}$ and $A \cup C = \{2, 3, 4, 5\}$ form a reduction.

A graph is *decomposable* if it is complete or if there exists a reduction of \mathcal{G} into decomposable subgraphs. Equivalently, a graph is decomposable if it can be recursively reduced to its cliques. For example, the graph in Fig. 2.2a is decomposable, since it can be recursively reduced to its cliques, which are given in Fig. 2.2c. The graph in Fig. 2.2d cannot be reduced and therefore is irreducible. It is not decomposable, since it is not a single clique.

A graph is *triangulated* if it contains no chordless cycles and a graph is decomposable if, and only if, it is triangulated. For example, the graph in Fig. 2.2a is triangulated, but the graph in Fig. 2.2d is not, since it is a chordless four cycle.

2.4 Markov Properties

A random vector \vec{X} with sample space \mathcal{X} exhibits, relative to the graph $\mathcal{G} = (V, E)$:

- the *pairwise Markov property* if $(i, j) \notin E \iff X_i \perp\!\!\!\perp X_j \mid \vec{X}_R$;
- the *global Markov property* if A , B and C are disjoint subsets of V , and A separates B and C implies $\vec{X}_B \perp\!\!\!\perp \vec{X}_C \mid \vec{X}_A$;
- the *local Markov property* if, for any $i \in V$, $X_i \perp\!\!\!\perp \vec{X}_{V \setminus \{i, bd(i)\}} \mid \vec{X}_{bd(i)}$.

Now \vec{X} exhibiting the global Markov property relative to \mathcal{G} implies \vec{X} exhibits the local Markov property relative to \mathcal{G} , and \vec{X} exhibiting the local Markov property relative to \mathcal{G} implies \vec{X} exhibits the pairwise Markov property relative to \mathcal{G} . Furthermore, if $P(\vec{X} = \vec{x}) > 0$ for all $\vec{x} \in \mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_p$, where \mathcal{X}_i is the sample space of X_i , then \vec{X} exhibiting the pairwise Markov property relative to \mathcal{G} implies \vec{X} exhibits the global Markov property relative to \mathcal{G} . The proof of the equivalence of the three Markov properties under this positive constraint can be found in [19, Theorem 3.4.1] and [15, Section 3.2.1].

Assuming the variables satisfy the positivity constraint, then since, by definition, \vec{X} exhibits the pairwise Markov property relative to its independence graph, it also exhibits the global and local Markov properties relative to its independence graph. Hence, the graph can be used to ascertain other conditional independence statements about \vec{X} . For example, if the graph in Fig. 2.2a is the independence graph for $\vec{X} = (X_1, \dots, X_5)^T$, then by definition

$$X_1 \perp\!\!\!\perp X_4 | \{X_2, X_3, X_5\}, \quad X_1 \perp\!\!\!\perp X_5 | \{X_2, X_3, X_4\}, \quad X_2 \perp\!\!\!\perp X_5 | \{X_1, X_3, X_4\}$$

and

$$X_3 \perp\!\!\!\perp X_5 | \{X_1, X_2, X_4\}.$$

By the global Markov property it follows that, for example,

$$X_1 \perp\!\!\!\perp \{X_4, X_5\} | \{X_2, X_3\}$$

and by the local Markov property it follows that, for example,

$$X_5 \perp\!\!\!\perp \{X_1, X_2, X_3\} | X_4.$$

2.5 Graphical Log-linear Models

The first log-linear model was developed by Birch in the 1960s. (For further details, see [3–5] and the books by Bishop et al. [7, 8] and Agresti [1]). Wermuth [18] gave analogies between log-linear models that exhibit conditional independence and the covariance selection models for continuous random variables of Dempster [12]. The work of Darroch et al. [10] then gave a way of displaying the independence structure of certain log-linear models using independence graphs.

2.5.1 Notation

Consider a p -dimensional contingency table cross-classifying the p random variables $\vec{X}_V = \vec{X} = (X_1, \dots, X_p)^T$, where $V = \{1, \dots, p\}$ and X_i has r_i categories labelled $1, \dots, r_i$. Hence, the sample space for X_i is $\mathcal{X}_i = \{1, \dots, r_i\}$ and the number of cells in the table is $r = \prod_{i=1}^p r_i$.

Let $n_V(\vec{x}_V) = n(\vec{x})$ be the observed cell counts and $P(\vec{X}_V = \vec{x}_V) = \pi_V(\vec{x}_V) = \pi(\vec{x})$ be the underlying cell probabilities. For example, if $p = 2$, $r_1 = 2$ and $r_2 = 3$, then the possible \vec{x} are $(1, 1), (1, 2), (1, 3), (2, 1), (2, 2), (2, 3)$ and the observed cell counts and underlying probabilities are displayed in Table 2.1a and b, respectively.

Let $n_A(\vec{x}_A)$ be the observed marginal cell count corresponding to a subset of the variables $A \subset V$ and let

$$\pi_A(\vec{x}_A) = \sum_{\vec{x}_B} \pi_V(\vec{x}_A, \vec{x}_B)$$

be the marginal cell probabilities; see Table 2.1a and b, respectively, for examples. Note that n_\emptyset is the sample size, and denote the p -dimensional table of cell counts by $\vec{n} = \{n_V(\vec{x}), \vec{x} \in \mathcal{X}\}$.

2.5.2 Log-linear Models

A saturated log-linear model for \vec{X}_V is

$$\log \pi_V(\vec{x}_V) = \sum_{D \subseteq V} \lambda_D(\vec{x}_D),$$

where the sum is over all possible subsets of V including the empty set \emptyset . For example, the saturated log-linear model for $\vec{X}_{12} = \vec{X}_{\{1,2\}}$ is

$$\log \pi_{12}(\vec{x}_{12}) = \lambda_\emptyset + \lambda_1(x_1) + \lambda_2(x_2) + \lambda_{12}(x_1, x_2). \tag{2.3}$$

For identifiability, set $\lambda_D(\vec{x}_D) = 0$ if any component of \vec{x}_D is one, for all $D \subseteq V$. These are called the corner-point constraints and are, for example, if $p = 2$, $r_1 = 2$

Table 2.1 (a) Observed cell counts and (b) underlying probabilities for a 2×3 contingency table

(a)				(b)			
$n(1, 1)$	$n(1, 2)$	$n(1, 3)$	$n_1(1)$	$\pi(1, 1)$	$\pi(1, 2)$	$\pi(1, 3)$	$\pi_1(1)$
$n(2, 1)$	$n(2, 2)$	$n(2, 3)$	$n_2(2)$	$\pi(2, 1)$	$\pi(2, 2)$	$\pi(2, 3)$	$\pi_2(2)$
$n_2(1)$	$n_2(1, 2)$	$n(1, 3)$	n_\emptyset	$\pi_2(1)$	$\pi_2(1, 2)$	$\pi(1, 3)$	1

and $r_2 = 3$,

$$\lambda_1(1) = \lambda_2(1) = \lambda_{12}(1, 1) = \lambda_{12}(1, 2) = \lambda_{12}(1, 3) = \lambda_{12}(2, 1) = 0.$$

Simpler models for \vec{X}_V can be specified by setting one or more of the $\lambda_D(\vec{x}_D)$ to zero for all \vec{x}_D , and the subsets of V for which $\lambda_D(\vec{x}_D) \neq 0$ for some \vec{x}_D define the model:

$$\log \pi_V(\vec{x}_V) = \sum_{D \in \mathcal{D}} \lambda_D(\vec{x}_D),$$

where $\mathcal{D} = \{D : \lambda_D(\vec{x}_D) \neq 0 \text{ for some } \vec{x}_D\}$. For example, if $\lambda_D(\vec{x}_{12}) = 0$ for all \vec{x}_{12} in model (2.3), then the simpler model is defined by the subsets $\{1\}$ and $\{2\}$. Furthermore, in the simpler model $X_1 \perp\!\!\!\perp X_2$, which follows from the factorisation criteria for independence, see expression (2.1), since $\pi_{12}(\vec{x}_{12}) = \lambda_{\emptyset} \lambda_1(x_1) \lambda_2(x_2) = g(x_1)h(x_2)$, say.

2.5.3 Hierarchical Log-linear Models

A log-linear model is called *hierarchical* if $\lambda_C(\vec{x}_C) = 0$ for all $\vec{x}_C \implies \lambda_D(\vec{x}_D) = 0$ for all $C \subseteq D$ and \vec{x}_D . For example, if $\lambda_2(x_2) = 0$ for all x_2 in (2.3), then the model is hierarchical only if $\lambda_{12}(x_1, x_2) = 0$ for all x_1 and x_2 ; otherwise it is non-hierarchical. Therefore, rather than specifying \mathcal{D} , all the subsets of V for which $\lambda_D(\vec{x}_D) \neq 0$ for some \vec{x}_D , to define a hierarchical log-linear model, only the maximal subsets, maximal with respect to inclusion, are needed. These subsets are called the *generators* of the hierarchical log-linear model and the set of generators is called the *generating class*: $\mathcal{C} = \{C \in \mathcal{D} : C \text{ is not a strict subset of any } D \in \mathcal{D}\}$.

Independence is not easily modelled using log-linear models for \vec{X}_V , $p > 2$, since it is a property of marginal distributions. However, conditional independence is easily modelled. For example, if $p = 3$, then

$$X_1 \perp\!\!\!\perp X_2 | X_3 \iff \lambda_{123}(x_1, x_2, x_3) = \lambda_{12}(x_1, x_2) = 0$$

and

$$X_1 \perp\!\!\!\perp X_3 | X_2 \iff \lambda_{123}(x_1, x_2, x_3) = \lambda_{13}(x_1, x_3) = 0.$$

In general, if $\vec{X}_V = \{\vec{X}_A, \vec{X}_B, \vec{X}_C\}$, then

$$\vec{X}_B \perp\!\!\!\perp \vec{X}_C | \vec{X}_A \iff \lambda_D(\vec{x}_D) = 0$$

for all D with one or more elements from B and one or more elements from C . This follows from the factorisation criteria for conditional independence, see

expression (2.2), since

$$\begin{aligned} \vec{X}_B \perp\!\!\!\perp \vec{X}_C | \vec{X}_A &\iff \log \pi_V(\vec{x}_V) = \sum_{D \subseteq P} \lambda_D(\vec{x}_D) = g(\vec{x}_A, \vec{x}_B) + h(\vec{x}_A, \vec{x}_C) \\ &\iff \lambda_D(\vec{x}_D) = 0 \end{aligned}$$

if D has one or more elements from B and one or more elements from C . Note that

$$X_i \perp\!\!\!\perp X_j | \vec{X}_R \iff \lambda_{ijA}(\vec{x}_{ijA}) = 0, \quad \text{for all } A \subseteq R.$$

Hence, a conditional independence graph can be drawn for any log-linear model:

$$(i, j) \notin E \iff \lambda_{ijA}(\vec{x}_{ijA}) = 0, \quad \text{for all } A \subseteq R.$$

Table 2.2 presents examples of possible graphs for a 3-dimensional contingency table along with the conditional independencies and log-linear parameters constrained to be zero in the log-linear model

$$\begin{aligned} \log \pi_{123}(\vec{x}_{123}) &= \lambda_\emptyset + \lambda_1(x_1) + \lambda_2(x_2) + \lambda_3(x_3) + \lambda_{12}(x_1, x_2) + \lambda_{13}(x_1, x_3) \\ &\quad + \lambda_{23}(x_2, x_3) + \lambda_{123}(x_1, x_2, x_3). \end{aligned} \quad (2.4)$$

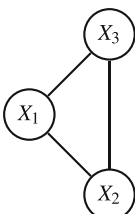
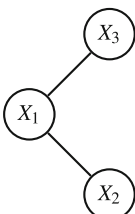
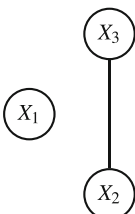
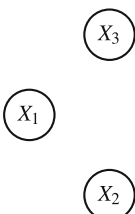
Also given in Table 2.2 are the generators of the models.

If $\lambda_{123}(x_1, x_2, x_3) = 0$, for all x_1, x_2 and x_3 , but each of the two-way λ -parameters are non-zero for some values of x_1, x_2 and x_3 , then there are no conditional independencies and hence the model with no three-way interaction has the same independence graph as model (a) in Table 2.2. Therefore, every graph does not correspond to a unique hierarchical log-linear model. The following hierarchical log-linear models all have the graph in Fig. 2.2a:

- (i) $\{1, 2, 3\}, \{2, 3, 4\}, \{4, 5\}$;
- (ii) $\{1, 2\}, \{1, 3\}, \{2, 3, 4\}, \{4, 5\}$;
- (iii) $\{1, 2, 3\}, \{2, 4\}, \{3, 4\}, \{4, 5\}$;
- (iv) $\{1, 2\}, \{1, 3\}, \{2, 3\}, \{2, 4\}, \{3, 4\}, \{4, 5\}$.

While the pairwise conditional independencies that define model (c) are listed in Table 2.2, note that it follows from the equivalence of Markov properties that $X_1 \perp\!\!\!\perp \{X_2, X_3\}$, since the empty set separates $\{X_1\}$ and $\{X_2, X_3\}$ (Global Markov Property) or, equivalently, since X_1 has no neighbours (Local Markov property). Similarly, under model (d) it follows that $X_1 \perp\!\!\!\perp X_2$, $X_1 \perp\!\!\!\perp X_3$ and $X_2 \perp\!\!\!\perp X_3$. This demonstrates the importance of the equivalence of Markov properties result if graphs are to be used to display the association structure of a random vector, since it would be odd not to be able to conclude from models (c) or (d) that $X_1 \perp\!\!\!\perp X_2$, etc.

Table 2.2 Examples of graphs, with corresponding conditional independencies, constrained log-linear parameters and generators for $p = 3$

Model	Graph	Independencies	Constraints	Generators
(a)		none	none	$\{1, 2, 3\}$
(b)		$X_2 \perp\!\!\!\perp X_3 X_1$	$\lambda_{23}(x_2, x_3) = 0$ $\lambda_{123}(x_1, x_2, x_3) = 0$	$\{1, 2\}, \{1, 3\}$
(c)		$X_1 \perp\!\!\!\perp X_2 X_3$ $X_1 \perp\!\!\!\perp X_3 X_2$	$\lambda_{12}(x_1, x_2) = 0$ $\lambda_{13}(x_1, x_3) = 0$ $\lambda_{123}(x_1, x_2, x_3) = 0$	$\{1\}, \{2, 3\}$
(d)		$X_1 \perp\!\!\!\perp X_2 X_3$ $X_1 \perp\!\!\!\perp X_3 X_2$ $X_2 \perp\!\!\!\perp X_3 X_1$	$\lambda_{12}(x_1, x_2) = 0$ $\lambda_{13}(x_1, x_3) = 0$ $\lambda_{23}(x_2, x_3) = 0$ $\lambda_{123}(x_1, x_2, x_3) = 0$	$\{1\}, \{2\}, \{3\}$

2.5.4 Graphical Log-linear Models

A log-linear model is called *graphical* if, and only if, its generators correspond to the cliques of the graph. Hence, the models defined in Table 2.2 are graphical, but the model with only $\lambda_{123}(x_1, x_2, x_3) = 0$, for all x_1, x_2 and x_3 , is not. It follows from the definition, that every graph has a unique graphical log-linear model. For example, model (i) above is the graphical model corresponding to the graph in Fig. 2.2a. Note that graphical log-linear models are a subset of hierarchical models.

2.5.5 Fitting Log-linear Models

Log-linear models can be fitted by maximising the log-likelihood:

$$\begin{aligned} l(\vec{\lambda}, \vec{n}) &= \log \prod_{\vec{x}} \pi(\vec{x})^{n(\vec{x})} \\ &= \sum_{\vec{x}} n(\vec{x}) \log \pi(\vec{x}) \\ &= \sum_{\vec{x}} n(\vec{x}) \sum_{D \in \mathcal{D}} \lambda_D(\vec{x}_D), \end{aligned}$$

subject to $\sum_{\vec{x}} \pi(\vec{x}) = 1$; see, for example, [1, Sections 9.6.1 to 9.6.5].

An interesting point here is that the parameters of a log-linear model have direct estimates if, and only if, the model is a graphical log-linear model and the corresponding independence graph is decomposable or, equivalently, triangulated [10]. Such models are called *decomposable* models. For example, the parameters of the models defined in Table 2.2 all have direct estimates, but those for the model with only $\lambda_{123}(x_1, x_2, x_3) = 0$ for all x_1, x_2 and x_3 do not, nor do those for the graphical log-linear model corresponding to the graph in Fig. 2.2d. Furthermore, only model (i) of the models with the graph in Fig. 2.2a listed above has direct estimates.

To summarise, the types of log-linear model introduced above can be ordered as:

$$\text{decomposable} \subset \text{graphical} \subset \text{hierarchical} \subset \text{log-linear}.$$

Overall goodness of fit of a model can be assessed by using the deviance:

$$\text{dev}(M) = 2n_{\emptyset} \sum_{\vec{x}} n(\vec{x}) \log \frac{n(\vec{x})}{n_{\emptyset} \hat{\pi}(\vec{x})},$$

where $\hat{\pi}(\vec{x})$ is the maximum likelihood estimate under model M ; see, for example, [1, Section 9.6.6]. Under the null hypothesis that model M generated the data, $\text{dev}(M)$ has, asymptotically as $n_{\emptyset} \rightarrow \infty$, a chi-squared distribution with degrees of freedom equal to $r - q$, where q is the number of unconstrained λ -parameters in M .

Two nested models $M_0 \subset M_1$ can be compared using the differences in deviances:

$$\text{dev}(M_0|M_1) = \text{dev}(M_0) - \text{dev}(M_1);$$

see, for example, [1, Section 4.5.4]. Under the null hypothesis that model M_0 generated the data, $\text{dev}(M_0|M_1)$ has, asymptotically as $n_{\emptyset} \rightarrow \infty$, a chi-squared distribution with degrees of freedom equal to $q_1 - q_0$, where q_m is the number of unconstrained λ -parameters in model M_m .

The *edge exclusion deviances* are often useful when selecting a graphical model that best represents the conditional independence structure of a random vector. These are the deviances between two models whose graphs differ by one edge:

$$edev(i, j|M_1) = dev(M_0|M_1),$$

where M_0 is the model with edge (i, j) removed from M_1 . For example, the edge exclusion deviances for comparing the models (a) and (b), and (c) and (d) in Table 2.2 are $edev(2, 3|\{1, 2, 3\})$ and $edev(2, 3|\{1\}, \{2, 3\})$, respectively. Note that, if M_1 is the saturated model, M_S say, and model M_0 is the model with only edge (i, j) missing, model $M_{\setminus(i,j)}$ say, then

$$edev(i, j|M_S) = dev(M_{\setminus(i,j)}).$$

While there are packages specially developed to fit graphical models, including some for R (for more details see [14]), any package which can fit log-linear models can be used to fit graphical log-linear models.

2.5.6 Example: Infant Survival Data

The infant survival data set presented in Bishop [6] and Bishop et al. [7, 8] has been used by Whittaker [19] and Roverato [17] to illustrate the use of graphical log-linear models and will be used here. Table 2.3 presents a three-way contingency table cross-classifying infants by the clinic they attended (1 or 2), the amount of care they received (less or more) and their survival status (yes or no).

Table 2.4a presents the edge exclusion deviances from the saturated model, $edev(i, j|M_S)$, for the infant survival data. There is strong evidence of a conditional independence between care and survival given clinic when these deviances are compared to a χ^2_2 distribution, but no evidence of either of the other two conditional independencies. In general, one approach to model selection would be now to remove the non-significant edge and test if further edges could be removed. For example, Table 2.4b presents these edge exclusion deviances, $edev(i, j|M_{\setminus(2,3)})$,

Table 2.3 Infant survival data

Clinic (X_1)	Care (X_2)	Survival (X_3)	
		No	Yes
1	Less	3	176
	More	4	293
2	Less	17	197
	More	2	23

Reproduced with permission from part of Table 2 of [6]

Table 2.4 Edge exclusion deviances for the infant survival data

(a)	$e\text{dev}(i, j M_S)$				(b)	$e\text{dev}(i, j M_{\setminus(2,3)})$			
	Clinic (X_1)					Clinic (X_1)			
	Care (X_2)	188.1				Care (X_2)	193.7		
	Survival (X_3)	12.22	0.082			Survival (X_3)	17.75	–	
		X_1	X_2	X_3			X_1	X_2	X_3

which should be compared with a χ_1^2 distribution. As expected from the edge exclusion deviances in Table 2.4a, the edges between care and clinic, and survival and clinic remain significant. Hence, model (b) in Table 2.2 best represents the conditional independence structure for these data and the conclusion is that after controlling for clinic, care is independent of survival. However, clinics have significantly different survival rates and significantly different care profiles.

For further examples of analyses using graphical log-linear models, see [19, Sections 7.5 to 7.7 and 8.5], [13, Chapter 2] and [14, Chapter 2].

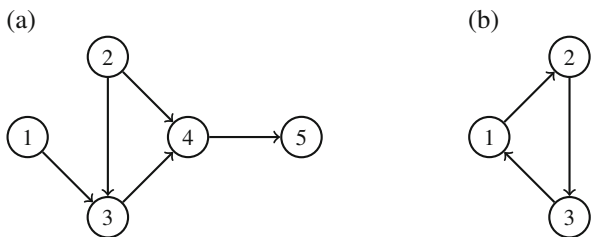
2.6 Directed Graphical Models

The graphical models discussed in Sect. 2.5 treat all the variables on an equal footing, whereas often the variables can or should be partially or fully ordered such that some variables are considered as purely explanatory variables, others as intermediate variables and some as response variables. When the variables can be completely ordered then direct graphical models, the subject of this section, are appropriate, whereas when they can only be partially ordered graphical chain models, the subject of Sect. 2.7, may be used.

2.6.1 Directed Acyclic Graphs

If the vertices can be completely ordered and only edges from $i < j$ are permitted, then the resulting graph will be a directed acyclic graph (DAG): directed since only directed edges are permitted, often drawn as arrows, and acyclic since it is impossible for a *directed path* i_1, i_2, \dots, i_m , where $i_l < i_{l+1}$, to be a cycle. An example of a DAG is given in Fig. 2.3a, whereas the graph in Fig. 2.3b, while directed, is not acyclic. Two vertices $i < j$ are *adjacent* in a DAG if $(i, j) \in E$. Note that a directed path in a DAG follows the direction of the arrows, whereas an *undirected path* is a sequence of adjacent vertices. For example, in Fig. 2.3a vertices 1, 3, 4 form a directed path and vertices 1, 3, 2 form an undirected path. While [15] reserves *path* for a directed path and uses *chain* to refer to an undirected path, below *path* refers to an undirected path as it does in [13] and [17].

Fig. 2.3 Examples of (a) a DAG and (b) a cyclic graph



The set of parents of vertex i is denoted by $pa(i)$ and the set of children of vertex i by $ch(i)$, where $j < i$ is a *parent* of i if $(j, i) \in E$ and $j > i$ is a *child* of i if $(i, j) \in E$. For example, in Fig. 2.3a $pa(4) = \{2, 3\}$ and $ch(4) = \{5\}$.

The *ancestors* of a vertex i , $an(i)$, is the set of vertices that have a directed path to i and the *descendants* of i , $de(i)$, is the set vertices that have a directed path from i . For example, in Fig. 2.3a, $an(4) = \{1, 2, 3\}$ and $de(3) = \{4, 5\}$. The ancestors of a set A , $an(A)$, are all the ancestors of $i \in A$ that are not in A , that is, $an(A) = \{\cup_{i \in A} an(i)\} \setminus A$. Finally, following [13], $an^+(A) = A \cup an(A)$. For example, Fig. 2.3a, $an(\{3, 4\}) = \{1, 2\}$ and $an^+(\{3, 4\}) = \{1, 2, 3, 4\}$.

The directed global Markov property given in Sect. 2.6.2 uses d-separation. To define d-separation we need to distinguish between collider and non-collider vertices in a path. A vertex j is called a *collider* in a path if edges (i, j) and (j, k) are in the path. It is call a *non-collider* if it is neither of the terminal vertices nor a collider in the path. For example, vertex 4 is a collider and vertex 3 is a non-collider in the path 1, 3, 4, 2 in Fig. 2.3a. A subset of vertices A *blocks* a path between $i \notin A$ and $j \notin A$ if the path has either (i) a non-collider in A or (ii) the path has a collider k , say, such that k nor its descendants are in A , that is, $\{k, de(k)\} \cap A = \emptyset$. For example, in Fig. 2.3a $A = \{3\}$ does not block the path 1, 3, 2, since 3 is a collider in this path, whereas $A = \{3\}$ blocks the path 1, 3, 4, since 3 is a non-collider in this path. Similarly, $A = \{2\}$ blocks the path 3, 2, 4. Note that $A = \emptyset$ blocks the path 1, 3, 2, since 3 is a collider and, by definition, neither it nor its descendant are in \emptyset . However, $A = \{4, 5\}$ does not block the path 1, 3, 2, since 4 and 5 are descendants of the collider 3 in the path.

A subset of vertices A *d-separates* subsets of vertices B and C (A , B and C disjoint) if for every pair of vertices $i \in B$ and $j \in C$ it blocks all paths between i and j . For example, in Fig. 2.3a $A = \{4\}$ d-separates $B = \{2, 3\}$ from $C = \{5\}$, but $A = \{3\}$ does not d-separate $B = \{1\}$ from $C = \{5\}$, since while it blocks the path 1, 3, 4, 5, it does not block the path 1, 3, 2, 4, 5. See [17, pages 113 and 114] for further examples.

Rather than checking if all paths are blocked, another method to check for d-separation uses the moral graph of an ancestral subgraph. The *moral graph* $\mathcal{G}^m = (V, E^m)$ corresponding to the DAG $\mathcal{G} = (V, E)$ has undirected edges added between the parents of each of the vertices and all the other directed edges replaced by undirected edges, that is,

$$E^m = \{(i, j) : (i, j) \in pa(k), k \in V\} \cup \{(i, j) : (i, j) \text{ or } (j, i) \in E\}.$$

For example, the moral graph of the DAG in Fig. 2.3a is the graph in Fig. 2.2a. An undirected edge has been added between vertices 1 and 2, the parents of vertex 3, and between vertices 2 and 3, the parents of vertex 4, and all the other directed edges replaced by undirected edges. Note that the moral graph is an undirected graph.

The *ancestral subgraph* of a set of vertices A is the subgraph $\mathcal{G}_{an^+(A)}$. For example, the ancestral subgraph of $\{3, 4\}$ is the DAG in Fig. 2.3a without vertex 5 and the arrow to it.

Lauritzen [15, Proposition 3.25] states that a subset of vertices A d-separates subsets of vertices B and C (A , B and C disjoint) in the DAG $\mathcal{G} = (V, E)$ if, and only if, A separates subsets of vertices B and C in the moral graph of the ancestral graph of $A \cup B \cup C$, that is, $(\mathcal{G}_{an^+(A \cup B \cup C)})^m$.

2.6.2 Directed Markov Properties

The Markov properties can be modified for DAGs. Denote by $pr(i) = \{j : j < i \in V\}$ the vertices prior to i . Then a random vector \vec{X} with sample space \mathcal{X} exhibits, relative to the DAG $\mathcal{G} = (V, E)$:

- the *direct pairwise Markov property* if, for vertices $i < j$,

$$(i, j) \notin E \iff X_i \perp\!\!\!\perp X_j | \vec{X}_{pr(j) \setminus \{i\}};$$

- the *directed global Markov property* if A , B and C are disjoint subsets of V and A d-separates B and C implies $\vec{X}_B \perp\!\!\!\perp \vec{X}_C | \vec{X}_A$;
- the *directed local Markov property* if, for any $i \in V$, $X_i \perp\!\!\!\perp \vec{X}_{pr(i) \setminus pa(i)} | \vec{X}_{pa(i)}$.

For example, if \vec{X} exhibits the directed pairwise Markov property relative to the DAG in Fig. 2.3a, then

$$\begin{aligned} X_1 \perp\!\!\!\perp X_2, \quad X_1 \perp\!\!\!\perp X_4 | \{X_2, X_3\}, \quad X_1 \perp\!\!\!\perp X_5 | \{X_2, X_3, X_4\}, \\ X_2 \perp\!\!\!\perp X_5 | \{X_1, X_3, X_4\} \end{aligned}$$

and

$$X_3 \perp\!\!\!\perp X_5 | \{X_1, X_2, X_4\}.$$

If it exhibits the directed global Markov property, then, for example,

1. $X_1 \perp\!\!\!\perp X_2$, since, as noted above, $A = \emptyset$ d-separates vertices 1 and 2. Alternatively, this follows since there is no edge in (the moral graph of) the ancestral graph of

{1, 2}. However, we cannot conclude that $X_1 \perp\!\!\!\perp X_2 | X_3$, since the moral graph of the ancestral graph of {1, 2, 3} has no edges missing.

2. $X_1 \perp\!\!\!\perp X_5 | \{X_2, X_3\}$ and $X_1 \perp\!\!\!\perp X_5 | X_4$, since both {2, 3} and {4} separate vertices 1 and 5 in the moral graph of Fig. 2.3a, that is the graph in Fig. 2.2a. However, we cannot conclude that $X_1 \perp\!\!\!\perp X_5 | X_3$, since 3 is a collider in the path 1, 3, 2, 4, 5.

If \vec{X} exhibits the directed local Markov property, then

$$X_4 \perp\!\!\!\perp X_1 | \{X_2, X_3\} \quad \text{and} \quad X_5 \perp\!\!\!\perp \{X_1, X_2, X_3\} | X_4.$$

Under the same conditions as for the undirected case, the directed pairwise, local and global Markov properties are equivalent. For further details, see [15, Section 3.2.2].

2.6.3 Models

In general, given an ordering of the p random variables, the joint probability mass function (p.m.f.) of \vec{X}_V , $P(\vec{X}_V = \vec{x}_V)$, can be factorised into the product of a univariate marginal p.m.f. and $p - 1$ univariate conditional p.m.f.s:

$$P(\vec{X}_V = \vec{x}_V) = P(X_1 = x_1) \prod_{i=2}^p P\left(X_i = x_i | \vec{X}_{pr(i)} = \vec{x}_{pr(i)}\right). \quad (2.5)$$

Now if \vec{X}_V exhibits the directed local Markov property relative to a DAG, then the conditioning sets in Eq. (2.5) can be replaced by $pa(i)$, since $X_i \perp\!\!\!\perp \vec{X}_{pr(i) \setminus pa(i)} | \vec{X}_{pa(i)}$ implies

$$P\left(X_i = x_i | \vec{X}_{pr(i)} = \vec{x}_{pr(i)}\right) = P\left(X_i = x_i | \vec{X}_{pa(i)} = \vec{x}_{pa(i)}\right)$$

and hence

$$P(\vec{X}_V = \vec{x}_V) = P(X_1 = x_1) \prod_{i=2}^p P\left(X_i = x_i | \vec{X}_{pa(i)} = \vec{x}_{pa(i)}\right). \quad (2.6)$$

While there are packages available to fit directed graphical models to categorical data, by taking advantage of factorisation (2.5), a directed graphical model can be selected by considering a series of $p - 1$ models for $P\left(X_i = x_i | \vec{X}_{pr(i)} = \vec{x}_{pr(i)}\right)$, $i = 2, \dots, p$. If $r_i = 2$, then a binary logistic regression model can be considered, whereas if $r_i > 2$ either a multinomial or possibly an ordinal logistic regression model is appropriate. The explanatory variables considered in the model are $\vec{X}_{pr(i)}$

and those that are significant form $\vec{X}_{pa(i)}$. For further details regarding logistic regression models, see [1, Chapters 5 and 8]. Alternatively, for unordered variables, as described in [2], a hierarchical log-linear model can be fitted to $\{X_i, \vec{X}_{pr(i)}\}$, where the highest order interaction between the explanatory variables, $\vec{X}_{pr(i)}$, is included in the model. The subset of variables in $\vec{X}_{pr(i)}$ that have significant interaction with X_i forms $\vec{X}_{pa(i)}$.

2.6.4 Example: Infant Survival Data (Continued)

Rather than treat all the variables on a equal footing, as in Sect. 2.5.6, it might be more appropriate to order variables in the infant survival example as clinic (X_1), care (X_2) and survival (X_3), where clinic is a potential cause of the type of care and it, along with care, is a potential cause of survival. While this is not a complex example, it does permit comparison with the analysis in Sect. 2.5.6 where the three variables were treated on an equal footing.

The test for an edge (arrow) between X_1 and X_2 in the DAG corresponds to the test of $X_1 \perp\!\!\!\perp X_2$ or, equivalently, the test of $\lambda_{12}(2, 2) = 0$ in log-linear model (2.3). In this case the edge exclusion deviance is $edev(1, 2|\{1, 2\}) = 193.7$, which, when compared to a χ^2_1 distribution, is highly significant. The tests for edges between X_1 and X_3 , and between X_2 and X_3 in the DAG correspond, respectively, to the tests of $X_1 \perp\!\!\!\perp X_3|X_2$ and $X_2 \perp\!\!\!\perp X_3|X_1$ or, equivalently, the tests of $\lambda_{13}(2, 2) = \lambda_{123}(2, 2, 2) = 0$ and $\lambda_{23}(2, 2) = \lambda_{123}(2, 2, 2) = 0$ in log-linear model (2.4). In this case the edge exclusion deviances are given in the last row of Table 2.4a and the conclusion is that there should be an edge between X_1 and X_3 , but not between X_2 and X_3 . The selected DAG is presented in Fig. 2.4b.

While the conclusions are the same as for the analysis using undirected graphs, note that the presence of an edge between X_1 and X_2 in the DAG, presented in Fig. 2.4b, corresponds to the lack of independence between these two variables, whereas the presence of the corresponding edge in the undirected graph, presented in Fig. 2.4a, corresponds to lack of conditional independence.

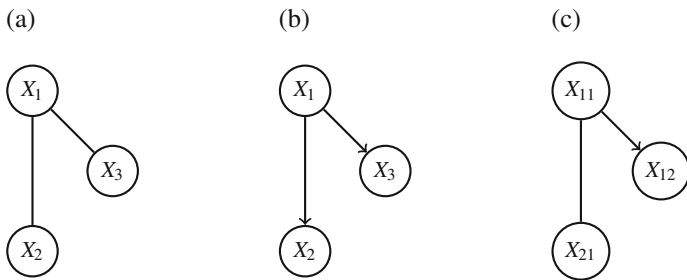


Fig. 2.4 The selected (a) undirected graph, (b) DAG and (c) chain graph for the infant survival data

2.7 Graphical Chain Models

If the variable can be partitioned into B ordered blocks, where the variables in a block are to be treated on an equal footing and the variables in succeeding blocks can be considered as responses to the variables in the preceding blocks, then a graphical chain model might be appropriate and the association structure between the variables can be presented in a chain graph.

2.7.1 Chain Graphs

Let $ib \in V, i = 1, \dots, p_b$ and $b = 1, \dots, B$ denote the i th vertex in the b th block. If only undirected edges are permitted between vertices within a block, $(ib, jb) \in E \iff (jb, ib) \in E$, and only directed edges are permitted from vertices in block b to vertices block $c > b$, then the resulting graph will be a chain graph. An example of a chain graph with three blocks, with three vertices in block 1, two in block 2 and one in block 3, is given in Fig. 2.5a.

In a chain graph a vertex can have neighbours, parents and children which are the adjacent vertices in the same, preceding and succeeding blocks, respectively. The sets of neighbours, parents and children of vertices ib are denoted, respectively, by $ne(ib) = \{jb : (ib, jb) \in E\}$, $pa(ib) = \{jc : (jc, ib) \in E, b > c\}$ and $ch(ib) = \{jc : (ib, jc) \in E, b < c\}$. The *boundary* of vertex ib is the set of its neighbours and parents: $bd(ib) = ne(ib) \cup pa(ib)$. For example, in Fig. 2.5a $ne(22) = \{12\}$, $pa(22) = \{21\}$, $ch(22) = \{13\}$ and hence $bd(22) = \{21, 12\}$, and $ne(21) = \{11, 31\}$, $pa(21) = \emptyset$, $ch(21) = \{12, 22\}$ and hence $bd(21) = ne(21)$.

The neighbours, parents, children and boundary of a subset of vertices $A \subset V$ is the union of those of the vertices in A minus the vertices in A . For example, $bd(A) = \cup_{ib \in A} bd(ib) \setminus A$ and in Fig. 2.5a, if $A = \{21, 12\}$, $bd(A) = \{11, 31, 22\}$.

The concept of a moral graph \mathcal{G}^m extends to chain graphs, but now, as well as adding edges between parents, edges are added between parents whose children are in the same block and are connected by a path entirely in that block. For example, the moral graph of the chain graph in Fig. 2.5b is given in Fig. 2.5c. Here $ch(11) = \{12\}$

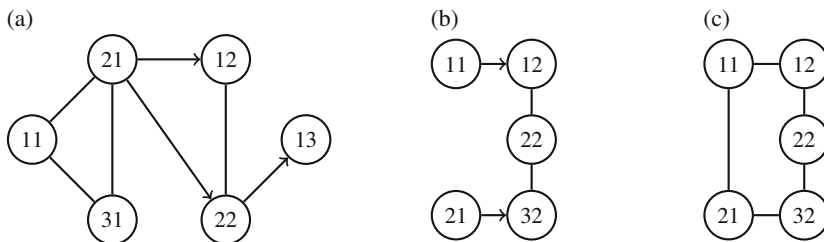


Fig. 2.5 Examples of (a) and (b) chain graphs and (c) the moral graph of (b)

and $ch(21) = \{32\}$ are connected by the path 12, 22, 32 entirely in block 2, and therefore an edge is added between 11 and 21 in the moral graph. Note that the moral graph of the chain graph in Fig. 2.5a has no edges added, just the directed edges replaced by undirected edges.

The definition of the ancestral set $an^+(A)$ extends to chain graphs and is also called the anterior set of A . As well as including all the vertices in A and the ancestors of all the vertices in A , the *anterior set* of A , again here denoted by $an^+(A)$, also includes any vertex that is in the same block as a vertex in A and connected to it by a path entirely in that block, and their ancestors. For example, in Fig. 2.5b, $an^+(\{12, 22\}) = \{11, 21, 12, 22, 32\}$, since vertex 32 is connected to vertex 22 by a path entirely in block 2, and vertex 21 is an ancestor of vertex 32. Note that this is also the anterior set of any non-empty subset of the vertices in block 2.

The concept of d-separation can be extended to chain graphs and has been called c-separation. A subset of vertices A *c-separates* subsets of vertices B and C (A , B and C disjoint) in the chain graph $\mathcal{G} = (V, E)$ if, and only if, A separates subsets of vertices B and C in the moral graph of the anterior graph of $A \cup B \cup C$, that is, $(\mathcal{G}_{an^+(A \cup B \cup C)})^m$ (see, for example, [14], pages 13 and 14). For example, in Fig. 2.5b $A = \emptyset$ c-separates vertices 11 and 21, since $an^+(\{11, 21\}) = \{11, 21\}$ and there is no edge added in $(\mathcal{G}_{\{11, 21\}})^m$. However, vertex 22 does not c-separate vertices 11 and 21, since there is an edge between these vertices in the moral graph of the anterior graph of $\{11, 21, 22\}$, Fig. 2.5c.

2.7.2 Chain Graph Markov Properties

Again the Markov properties can be modified for chain graphs. Denote now by $pr(ib) = \{jc \in V : c < b\} \cup \{jb \in V : j \neq i\}$ the vertices in blocks preceding block b and the vertices other than i in block b . Then a random vector \vec{X} with sample space \mathcal{X} exhibits, relative to the chain graph $\mathcal{G} = (V, E)$:

- the *pairwise chain Markov property* if, for $b \leq c$,

$$(ib, jc) \notin E \iff X_{ib} \perp\!\!\!\perp X_{jc} \mid \vec{X}_{pr(jc) \setminus \{ib\}};$$

- the *global chain Markov property* if A , B and C are disjoint subsets of V and A c-separates B and C implies $\vec{X}_B \perp\!\!\!\perp \vec{X}_C \mid \vec{X}_A$;
- the *local chain Markov property* if, for any $ib \in V$, $X_{ib} \perp\!\!\!\perp \vec{X}_{pr(ib) \setminus bd(ib)} \mid \vec{X}_{bd(ib)}$.

For example, if \vec{X} exhibits the pairwise chain Markov property relative to the chain graph in Fig. 2.5a, then

$$\begin{aligned} X_{11} \perp\!\!\!\perp X_{12} \mid \{X_{21}, X_{31}, X_{22}\}, \quad X_{11} \perp\!\!\!\perp X_{22} \mid \{X_{21}, X_{31}, X_{12}\}, \\ X_{11} \perp\!\!\!\perp X_{13} \mid \{X_{21}, X_{31}, X_{12}, X_{22}\}, \end{aligned}$$

$$X_{21} \perp\!\!\!\perp X_{13} | \{X_{11}, X_{31}, X_{12}, X_{22}\}, \quad X_{31} \perp\!\!\!\perp X_{12} | \{X_{11}, X_{21}, X_{22}\}, \\ X_{31} \perp\!\!\!\perp X_{22} | \{X_{11}, X_{21}, X_{12}\}$$

$$X_{31} \perp\!\!\!\perp X_{13} | \{X_{11}, X_{21}, X_{12}, X_{22}\} \quad \text{and} \quad X_{12} \perp\!\!\!\perp X_{13} | \{X_{11}, X_{21}, X_{31}, X_{22}\}.$$

If it exhibits the local chain Markov property, then

$$X_{12} \perp\!\!\!\perp \{X_{11}, X_{31}\} | \{X_{21}, X_{22}\}, \quad X_{22} \perp\!\!\!\perp \{X_{11}, X_{31}\} | \{X_{21}, X_{12}\}$$

and

$$X_{13} \perp\!\!\!\perp \{X_{11}, X_{21}, X_{31}, X_{12}\} | X_{22}.$$

If \vec{X} exhibits the global chain Markov property relative to the chain graph in Fig. 2.5b, then, for example,

1. $X_{11} \perp\!\!\!\perp X_{21}$, since as noted above, $A = \emptyset$ c-separates vertices 11 and 21. However, we cannot conclude that $X_{11} \perp\!\!\!\perp X_{21} | X_{22}$, since as also noted above, vertex 22 does not separate vertices 12 and 22 in the moral graph of Fig. 2.5b, that is the graph in Fig. 2.5c.
2. $X_{11} \perp\!\!\!\perp X_{32} | \{X_{21}, X_{12}\}$ and $X_{11} \perp\!\!\!\perp X_{32} | \{X_{21}, X_{22}\}$, since both $\{21, 12\}$ and $\{21, 22\}$ separate vertices 11 and 32 in the moral graph of Fig. 2.5b.

Under the same conditions as for the undirected case, the pairwise and local and global chain Markov properties are equivalent. For further details, see [15, Section 3.2.3].

2.7.3 Models

In general, if the variables are partitioned into B blocks, then the joint p.m.f. of \vec{X}_V , $P(\vec{X}_V = \vec{x}_V)$, can be factorised into the product of a marginal p.m.f. and $B - 1$ conditional p.m.f.s:

$$P(\vec{X}_V = \vec{x}_V) = P(\vec{X}_1 = \vec{x}_1) \prod_{b=2}^B P(\vec{X}_b = \vec{x}_b | \vec{X}_1 = \vec{x}_1, \dots, \vec{X}_{b-1} = \vec{x}_{b-1}), \quad (2.7)$$

where $\vec{X}_b = \{X_{1b}, \dots, X_{pb}\}$ is the set of variables in block b . Note here that the marginal and conditional densities are in general multi-dimensions with dimensions p_1 and p_b , $b = 2, \dots, B$, respectively.

Let $A_b = \{ib, i = 1, \dots, p_b\}$ be the set of vertices in block b . Now if \vec{X}_V exhibits the Markov properties relative to a chain graph, then the conditioning sets in Eq. (2.5) can be replaced by $bd(A_b)$ and hence

$$P(\vec{X}_V = \vec{x}_V) = P(\vec{X}_1 = \vec{x}_1) \prod_{b=2}^B P\left(\vec{X}_b = \vec{x}_b \mid \vec{X}_{bd(A_b)} = \vec{x}_{bd(A_b)}\right). \quad (2.8)$$

Similar to a DAG, a chain graph can be selected by fitting a series of models corresponding to the factorisation (2.7). For the first block, an undirected log-linear model is selected for \vec{X}_1 . For blocks that contain only one variable ($p_b = 1$), the same approach as for a DAG can be used and binary, ordinal or multinomial logistic regression models or log-linear models can be selected. For blocks with more than one variable ($p_b > 1$), then, as described in Asmussen and Edwards [2], a hierarchical log-linear model can be fitted to $\{\vec{X}_1, \dots, \vec{X}_b\}$, where the highest order interaction between the explanatory variables, $\{\vec{X}_1, \dots, \vec{X}_{b-1}\}$, is included in the model. An undirected edge (ib, jb) within block b is required if there is a significant interaction between X_{ib} and X_{jb} , and a directed edge (ib, jc) , $b < c$, between vertex i in block b and vertex j in block c is required if there is a significant interaction between X_{ib} and X_{jc} .

Mohamed et al. [16] used this general approach based on the factorisation (2.7) to select graphical chain models to identify the determinants of neonatal and post-neonatal mortality in Malaysia, although they were not always able to include all of the interaction between the explanatory variables because of the sparseness of the data.

2.7.4 Example: Infant Survival Data (Continued)

For the infant survival data, rather than treat all the variables on an equal footing, as in Sect. 2.5.6, or to completely order them, as in Sect. 2.6.4, another option is to treat clinic ($X_{11} = X_1$) and care ($X_{21} = X_2$) on an equal footing and survival ($X_{21} = X_3$) as a response variable. For this very simple example, the test for an undirected edge between X_{11} and X_{21} in the chain graph is the same as the test for a directed edge between X_1 and X_2 in the DAG. Similarly, the tests for directed edges (arrows) between X_{11} and X_{12} , and between X_{21} and X_{12} in the chain graph are the same as for the DAG. Hence, the selected chain graph, presented in Fig. 2.4c, has edges between the same variables as the selected DAG, presented in Fig. 2.4b. However, note the difference in the type of edge between clinic and care.

2.8 Further Reading

This chapter has focussed on graphical models for categorical variables. However, as mentioned in the Introduction, there are graphical Gaussian models for when all the random variables are continuous and mixed interaction models for when there is a mixture of continuous and discrete random variables. The three comprehensive texts on graphical models [13, 15, 19] all contain chapters on graphical Gaussian and mixed interaction models, as well as on graphical log-linear models. For further details on graphical models for categorical data, see also Roverato [17]. All these texts discuss undirected, directed and chain graphical models. Another book, which places greater emphasis on directed and chain graph models, is Cox and Wermuth [9].

References

1. Agresti, A.: *Categorical Data Analysis*, 3rd edn. Wiley, New Jersey (2013)
2. Asmussen, S., Edwards, D.: Collapsibility and response variables in contingency tables. *Biometrika* **70**, 567–78 (1983)
3. Birch, M.W.: Maximum likelihood in three-way contingency tables. *J. R. Statist. Soc. B* **25**, 220–223 (1963)
4. Birch, M.W.: The detection of partial association, I: the 2×2 case. *J. R. Statist. Soc. B* **26**, 313–324 (1964)
5. Birch, M.W.: The detection of partial association, II: the general case. *J. R. Statist. Soc. B* **27**, 111–124 (1965)
6. Bishop, Y.M.: Full contingency tables, logits and split contingency tables. *Biometrics* **25**, 383–399 (1969)
7. Bishop, Y.M., Fienberg, S., Holland, P.: *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge (1975)
8. Bishop, Y.M., Fienberg, S.E., Holland, P.W.: *Discrete Multivariate Analysis: Theory and Practice*. Springer, New York (2007)
9. Cox, D.R., Wermuth, N.: *Multivariate Dependencies: Models, Analysis, and Interpretation*. Chapman and Hall, Boca Raton (1996)
10. Darroch, J.N., Lauritzen, S. L., Speed, T.P.: Markov fields and log-linear interaction models for contingency tables. *Ann. Statist.* **8**, 522–539 (1980)
11. Dawid, A.P.: Conditional independence in statistical theory (with discussion). *J. R. Statist. Soc. B* **41**, 1–31 (1979)
12. Dempster, A.P.: Covariance selection. *Biometrics* **28**, 157–175 (1972)
13. Edwards, D.: *Introduction to Graphical Modelling*, 2nd edn. Springer, New York (2000)
14. Højsgaard, S., Edwards, D., Lauritzen, S.: *Graphical Models with R*. Springer, New York (2012)
15. Lauritzen, S.L.: *Graphical Models*. Clarendon Press, Oxford (1996)
16. Mohamed, W.N., Diamond, I.D., Smith, P.W.F.: The determinants of infant mortality in Malaysia: a graphical chain modelling approach. *J. R. Statist. Soc. A* **161**, 349–366 (1998)
17. Roverato, A. (2017) *Graphical Models for Categorical Data*. Cambridge: Cambridge University Press.

18. Wermuth, N.: Analogies between multiplicative models in contingency tables and covariance selection. *Biometrics* **32**, 95–108 (1976)
19. Whittaker, J.C.: *Graphical Models in Applied Multivariate Statistics*. Wiley, Chichester (1990)
20. Wright, S.: The theory of path coefficients: a reply to Niles' criticism. *Genetics* **8**, 239–255 (1923)
21. Wright, S.: The method of path coefficients. *Ann. Math. Statist.* **5**, 161–215 (1934)

Chapter 3

Marginal Models: An Overview



Tamás Rudas and Wicher Bergsma

3.1 Introduction

We start with motivating examples in Sect. 3.2, including repeated measurements, missing data, and graphical models. All these involve the application of models which apply restrictions only on subsets of the variables, that is, on marginals of the contingency table containing their joint distribution.

The restrictions imposed by marginal models apply to the association structures within subsets of variables. The association is captured by log-linear parameters calculated in marginals of the table and Sect. 3.3 deals with general aspects of parameters and parameterizations, including variation independence.

Marginal log-linear parameterizations are developed in Sect. 3.4 and some of their fundamental properties, like variation independence, smoothness, and collapsibility are also discussed. Depending on the choice of the marginals in which the log-linear parameters are determined, marginal log-linear parameterizations are appropriate to capture several characteristics of the marginal and conditional association structure.

Marginal log-linear models are defined by restricting some marginal log-linear parameters to zero, as described in Sect. 3.5.

Marginal log-linear parameters are the standard log-linear parameters calculated from marginal distributions and measure the strength of conditional and/or marginal

T. Rudas
Department of Statistics, Faculty of Social Sciences, Eötvös Loránd University, Budapest,
Hungary
e-mail: trudas@elte.hu

W. Bergsma (✉)
Department of Statistics, London School of Economics and Political Science, London, UK
e-mail: w.p.bergsma@lse.ac.uk

association. Marginal log-linear parameters are based on ordinary odds and odds ratios and their higher-dimensional generalizations. But, as described in Sect. 3.6, other types of odds ratios, which are particularly useful for ordinal data, may also be used to define marginal log-linear models.

Section 3.7 contains results concerning the general type of conditional independence models, including the case when some conditional independences apply to subsets of the variables, which may be formulated as marginal log-linear models.

Section 3.8 deals with estimation and testing. Lagrangian and Fisher scoring methods for maximum likelihood estimation are described and compared. The generalized estimating equations (GEE) approach for estimating marginal models is described as well.

Section 3.9 discusses areas of applications where marginal log-linear models either provide a general way of implementing the standard analysis or a new approach to answer the research question. These include directed graphical models, path models, and latent variable models, but many other applications are mentioned, too.

Very few proofs are included, as most of the results are quoted from research publications.

3.2 Motivation

There are several types of statistical problems where marginal distributions of higher dimensional joint distributions play a central role. In this section, we discuss three such broad types of problems.

3.2.1 Repeated Measurements and Panel Studies

In many experimental and observational settings, subjects are measured or observed repeatedly. The reasons for measuring repeatedly include to study the within-subject variability of the measurements, or to reduce measurement error by taking the average measurement value. In such cases, the measurements are made close to each other in time. Another reason for repeated measurements is to investigate the effect of a treatment applied to the subjects between the measurements, in which case one measurement is taken before, and another one after, the treatment. Sometimes the variability or stability of the measurement results over time is of interest, without any treatment being applied.

For example, variables A_1 and B_1 are observed in a first measurement, a treatment is applied, and then the same variables are measured again, denoted as A_2 and B_2 . There are a number of relevant hypotheses to test. The first one, say H_1 , is that A and B are independent both before and after treatment. One may argue that H_1 is true if and only if both H_{11} : “ A_1 is independent of B_1 ” and

H_{12} : “ A_2 is independent of B_2 ” are true. This is correct, but a test of H_1 with a given level cannot be constructed, in general, from separate tests of the hypotheses H_{11} and H_{12} . This would be possible if the samples for the pairs of variables A_1, B_1 and A_2, B_2 were independent, which is not the case in the current repeated measurements setup. Instead, one has observations for each unit in the sample for the variables A_1, B_1, A_2, B_2 , and H_1 states that in this 4-dimensional distribution there are two marginal independences, one for A_1 and B_1 , and one for A_2 and B_2 . This is a marginal model.

Another relevant model, say H_2 , in this setup is that the distributions of the two measurements of A are identical, that is, the treatment does not change the distribution on the population level, and similarly for B , but the results of the second measurements are independent. Thus H_2 contains restrictions on the $A_1 \times A_2$ (marginal homogeneity), $B_1 \times B_2$ (marginal homogeneity), and $A_2 \times B_2$ (independence) marginals.

The hypotheses H_1 and H_2 assume marginal models about the joint distribution.

A closely related longitudinal design is called a panel study, see, e.g., Frees and Kim [36], where the individuals in a sample are interviewed repeatedly at regular intervals. The advantage¹ of such a design is that changes in opinions, preferences, or attitudes may be studied in a more valid way than by simultaneously asking about current and also previous positions in a cross-sectional study. The main limitation of such an approach is that earlier opinions or attitudes are often not remembered and sometimes are not reported truthfully.

In the analysis of panel data, the transition probabilities from one position into another one are of central interest. In particular, the dependence of the transition probabilities on earlier positions is an important question because this determines the fragmentation of the data. More precisely, if the panel has, say, 5 waves and A_1, A_2, A_3, A_4, A_5 denotes the positions of a respondent regarding a particular question during the waves, then one is interested in deciding whether, for instance,

$$P(A_5|A_4, A_3) = P(A_5|A_4)$$

holds. If it does, then the position at wave 5 cannot be better predicted if, in addition to the position at wave 4, the position at wave 3 is also taken into account. For example, in this case the chance of supporting a particular political party at the time of wave 5 may depend on the preferred party at the time of wave 4, but if the latter is known, the party preference at the time of wave 3 provides no additional information.

Slightly more generally, if

$$P(A_t|A_{t-1}, A_{t-2}, \dots, A_1) = P(A_t|A_{t-1})$$

¹ The design also has disadvantages, of course. These include panel attrition and the fact that, even if originally selected appropriately, with passing time the sample will become different in composition from the current population.

holds for all waves (time points) t , then the joint distribution is called a one-step Markov chain. It is easy to see that this property is equivalent to the following conditional independence

$$A_t \perp\!\!\!\perp A_{t-2}, \dots, A_1 | A_{t-1}.$$

In detail, for the 5 waves this means that

$$A_3 \perp\!\!\!\perp A_1 | A_2,$$

$$A_4 \perp\!\!\!\perp A_2, A_1 | A_3,$$

$$A_5 \perp\!\!\!\perp A_3, A_2, A_1 | A_4.$$

For the joint distribution of the variables A_1, A_2, A_3, A_4, A_5 the model prescribes conditional independences on the $A_1 \times A_2, A_1 \times A_2 \times A_3, A_1 \times A_2 \times A_3 \times A_4$ and $A_1 \times A_2, \times A_3 \times A_4 \times A_5$ marginals.

3.2.2 *Missing Data and Data Fusion*

The statistical problems discussed next lead to the task of generating a joint distribution with given marginal distributions. Thus, the restrictions implied by design or the type of data collected in these cases fully determine some marginal distributions (or make it possible to estimate them) and do not only specify a model for them as in the previous examples.

One group of such problems is related to incomplete observations or missing data, see Little and Rubin [55]. When the data are collected through a survey of a human population, usually not all individuals selected by the sampling procedure answer the questions. Some are not found, some are found but are not willing to participate in the survey, and some do participate but choose not to answer some questions. While dealing with those who do not provide any information is a serious issue, the best utilization of the often only partial answers collected is an important statistical problem [see Little and Rubin 55]. A similar situation occurs when data are collected in an experimental setting, because of the dropout of the participants. One approach is to consider the responses collected for a particular subset of the questions and use them to estimate the joint distribution of the answers. These are estimates of some marginal distributions of the joint distribution of all answers. This procedure is justified, because the smaller a subset of questions is, the more individuals gave responses to all of them, and their joint distribution may be better estimated than the joint distribution of all variables.

For example, let the questionnaire contain 4 yes-no questions and let the variables A_1, A_2, A_3, A_4 contain the answers. Then, the distribution of A_1 may be estimated based on all the answers provided to the first question, and similarly for all other variables. Thus, the one-way marginal distributions are estimated based on different

subsets of the sample. Next, the $A_1 \times A_2$ marginal distribution is estimated based on the one-way marginal estimates already obtained and on the observations which contained responses to both A_1 and A_2 . From the latter, one may estimate the odds ratio [see, e.g., Rudas 75] between A_1 and A_2 and combine this with the one-way marginals to estimate the $A_1 \times A_2$ distribution. In theory, the procedure can be continued until the $A_1 \times A_2 \times A_3 \times A_4$ distribution is estimated, although it raises many compatibility and optimality issues, and as will be seen later, the feasibility of such a procedure depends heavily on the patterns of missing data.

Sometimes, the missing data pattern is not observed but is implied by design. When the questionnaire is too long, or answering all questions could be seen as a breach of the respondents' privacy, some of them may be asked A_1, A_2, A_3 , others A_1, A_4, A_5 , where now these may not be individual questions rather blocks of questions, and, of course, other patterns are also possible. The $A_1 \times A_2 \times A_3$ and the $A_1 \times A_4 \times A_5$ marginal distributions may be estimated, and from these the joint distribution. The design is called a split questionnaire as described in Rhemtulla and Little [68] but similar problems arise in so-called register-based censuses, see, e.g., Eppmann et al. [28].

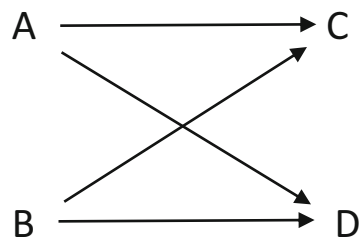
In a register-based census, now applied by several countries, instead of collecting information from all inhabitants of the country, data from existing registers (driving licences, health care access, etc.) are combined to find out the relevant information. The individual registers provide certain conditional and/or marginal distributions, and the task is to estimate the joint distributions. This problem is called data fusion, see, e.g., D'Orazio et al. [26], and it also occurs in other areas, see e.g., Cocchi [17].

3.2.3 Graphical Modelling

Graphical Markov models associated with directed acyclic graphs (also called Bayesian nets) are widely used in expert systems, artificial intelligence, and also in some approaches to modelling causal effects.

A simple example of a directed acyclic graph (DAG) is shown in Fig. 3.1. It has four nodes, A, B, C, D , which are identified with variables and the intuitive interpretation of the arrows is that they represent direct effects. The graph is acyclic, because there is no sequence of nodes in the order of arrows with the same starting and ending node.

Fig. 3.1 A directed acyclic graph



A precise interpretation of the Markov model associated with a DAG is that it assumes (conditional) independences among the variables. It is the missing arrows which imply the conditional independences defined by the directed Markov property, see, e.g., Lauritzen [50].²

Graphical Markov models associated with DAGs generalize the conditional independence property of Markov chains. In a Markov chain, the conditional independence is implied by the lack of temporal adjacency, and in the more general structures considered here, the temporal adjacency is replaced by the adjacency read off from a graph. When this graph is undirected, one Markov property, called the local Markov property, says that a variable is conditionally independent of its non-neighbours, given its neighbours.³ Notice that such a conditional independence involves all variables.

The assumption that the joint distribution of the variables obeys the conditional independences implied by the local Markov property applied to a particular graph is a graphical log-linear model, see Lauritzen [50].

To define the Markov property associated with a DAG, call those nodes from which an arrow goes to A the parents of node A and denote them as $\text{pa}(A)$. Further, call those nodes into which no directed path leads which starts in A the non-descendants of A and denote these nodes as $\text{nd}(A)$. Note that $\text{pa}(A) \subseteq \text{nd}(A)$, because otherwise the graph would contain a directed cycle. Then the local Markov property is that

$$A \perp\!\!\!\perp \text{nd}(A) \setminus \text{pa}(A) \mid \text{pa}(A).$$

In the case of the DAG in Fig. 3.1, the directed Markov property implies that

$$A \perp\!\!\!\perp B \tag{3.1}$$

$$C \perp\!\!\!\perp D \mid A, B. \tag{3.2}$$

Out of the two (conditional) independences, one is on the $A \times B$ marginal, and the other one is on the full table.

In general, the conditional independences defining a Markov model associated with a DAG are on marginals containing a variable and its non-descendants. Therefore, these models are marginal models.

An important property of the distributions which are Markov according to a DAG is that they factorize into the product of conditional distributions of the variables given their parents, see, e.g., Lauritzen [50]. For the DAG in Fig. 3.1, the factorization is

$$P(ABCD) = P(A)P(B)P(C|A, B)P(D|A, B).$$

² See Bergsma et al. [12], Chapter 5, for applications in causal analysis and the relationship with structural equation models.

³ The neighbours of a node A are those nodes with which A is connected by an edge.

3.3 Parameterizations of Discrete Probability Distributions

A marginal model is defined most conveniently using a particular parameterization of the joint distribution of the variables of interest, the so-called marginal log-linear parameterization, to be discussed in Sect. 3.4. This section deals with various general characteristics of parameters and parameterizations.

3.3.1 Parameters and Parameterizations

A parameter is an arbitrary function of the distribution, and is often multidimensional, i.e., vector-valued. For example, in the case of a 2×2 distribution, with the usual notation, (p_{11}, p_{12}) is a parameter, so is $(p_{11}, p_{12}, p_{21}, p_{22})$ or (p_{11}, p_{1+}, p_{+1}) . The parameters which are of interest in statistical analysis usually express some relevant property of the distribution. For example, a widely used measure of association between the two variables forming the table, see e.g., Rudas [75], is the odds ratio

$$\frac{p_{11}p_{22}}{p_{12}p_{21}},$$

which is also a parameter of the distribution. It measures a characteristic (strength and direction of association) which is not directly seen from the probabilities.

Therefore, a parameter represents information from the distribution. In many cases, one is interested in looking at parameters that carry all information in the distribution. A more formal way of imposing this is to consider, instead of the value of a certain parameter, the function which yields that parameter and to require that this function is invertible. If this holds, the parameter is called a parameterization.

In the case of a 2×2 distribution, the (p_{11}, p_{12}) parameter is not a parameterization, because if its value is known, the distribution cannot be reconstructed. But the $(p_{11}, p_{12}, p_{21}, p_{22})$ and (p_{11}, p_{1+}, p_{+1}) parameters are parameterizations. The cell probabilities are given in the first case, and are easily determined in the second. Also, the odds ratio and the marginal probabilities p_{1+} and p_{+1} form a parameterization, see Rudas [75]. In this case, inverting the parameterization, i.e. calculating the cell probabilities, needs to be done using numerical algorithms such as the Iterative Proportional Fitting or Scaling algorithm, see, e.g., Rudas [75].

While a parameter vector may have arbitrary dimension, a parameterization has a minimal dimension, which is the dimension of the distribution. In the case of a 2×2 distribution, although one has 4 probabilities, in the 4-dimensional space the distributions are in a 3-dimensional subspace, as their sum is 1. The same fact may also be formulated by saying that out of the 4 probabilities, only 3 are linearly independent. Therefore, the minimal dimension of a parameterization is 3.

3.3.2 Variation Independence

One of the most desirable properties of parameters and parameterizations is the variation independence of their components. Before giving a general definition, a simple example adopted from Rudas and Bergsma [76] is given to illustrate the concept.

Suppose in an experiment a 2×2 treatment by outcome table was observed, and as a measure of effect of the treatment, the difference in proportion of positive outcomes among those treated and among the control is used. Assume the data given in Tables 3.1 and 3.2 were observed for male and female participants, respectively.

The selected measure of the size of the effect takes on the value of

$$\frac{20}{100} - \frac{10}{100} = 0.1$$

for men and is

$$\frac{60}{100} - \frac{40}{100} = 0.2$$

for women. Is, then, the treatment twice as effective for women than for men?

The answer to this question is not clear. While the difference in the probabilities of positive outcomes under treatment and control seems like a meaningful measure of effect and its value is twice as big for women as for men, the following argument is also possible: for both men and women, 100 individuals received the treatment and 100 the control, but for men, there were 30 and for women 100 positive outcomes. Given this, the maximum possible value of the measure of effect is

$$\frac{30}{100} - \frac{0}{100} = 0.3$$

for men and is

$$\frac{100}{100} - \frac{0}{100} = 1$$

Table 3.1 Hypothetical experimental results for men

Outcome	Positive	Negative	Total
Treatment	20	80	100
Control	10	90	100

Table 3.2 Hypothetical experimental results for women

Outcome	Positive	Negative	Total
Treatment	60	40	100
Control	40	60	100

for women. Thus, the actual value of the measure of treatment efficacy is $1/3$ of its maximum possible value for men, while for women the actual value is only $1/5$ of its theoretical maximum.

This example illustrates that the possible range of the measure is affected by the values of the other parameters and, in such a case, the assessment of the actual value may be very different when the other parameters are or are not taken into account. Put differently, a parameter which is not variation independent of the other parameters lacks calibration.

Variation independence means that the above dependence does not occur. Two parameters are variation independent if their joint range is the Cartesian product of their individual ranges, i.e., any otherwise possible value of one can be combined with any otherwise possible value of the other.

In the case of the example discussed above, the measure of treatment efficacy was

$$\frac{p_{11}}{p_{+1}} - \frac{p_{12}}{p_{+2}} \quad (3.3)$$

and this was not variation independent of the other parameters p_{1+} and p_{+1} . Indeed, its minimum value is zero and its maximum value is

$$\min\left(1, \frac{p_{1+}}{p_{+1}}\right),$$

so its range is $[0, \min(1, p_{1+}/p_{+1})]$. To put it differently, the range of the measure (3.3) depends on the marginal distributions, and it is often not clear whether the inference should or should not condition on the marginals.

The odds ratio is variation independent of (p_{1+}, p_{+1}) and is, therefore, a parameter of the association with a calibration which does not depend on the marginals. Consequently, its values may be compared, even if calculated for tables with different marginal distributions. The value of the odds ratio is $1800/800 = 2.25$ for men and $3600/1600 = 2.25$ for women, suggesting that the strength of association between treatment and positive outcome is the same for men as for women, in contrast with the naive comparison of the measures calculated originally. This is, however, not to say that the odds ratio would be without problematic characteristics when used as a measure of treatment efficacy, see Rudas [73].

For higher dimensional tables, there are parameterizations which rely on the odds ratio and its generalizations with variation independence properties, leading to a natural definition of log-linear models, see Rudas [75], but for the definition of marginal models another type of parameterization, based on marginal and conditional distributions, is of more immediate use.

In the case of a 2×2 distribution, a parameterization with the p_{+1} marginal probability and the $p_{1|1}$ and $p_{1|2}$ conditional probabilities is also possible. Indeed,

$$p_{11} = p_{1|1}p_{+1}$$

$$p_{21} = (1 - p_{1|1})p_{+1}$$

$$p_{12} = p_{1|2}(1 - p_{+1})$$

$$p_{22} = (1 - p_{1|1})(1 - p_{+1}).$$

An important feature of this parameterization is that all three parameters in it are variation independent.

A similar parameterization of a distribution on a 4-way $A \times B \times C \times D$ table may parameterize the distribution on the $A \times B$ marginal, and then parameterize the conditional distribution on $C \times D$, given the marginal distribution on $A \times B$. Here, the two groups of parameters are variation independent. Further, within this parameterization, one may impose the marginal independence of A and B , and then the conditional independence of C and D , given A and B . Note that this is exactly the marginal model defined in (3.1) and (3.2), as implied by the Markov property applied to the graph in Fig. 3.1.

Alternatively, the following parameters constitute a parameterization of the 4-way table in the binary case, with OR denoting the odds ratio and COR the conditional odds ratio:

$$\theta_1 = (P(A = 1), P(B = 1))$$

$$\theta_2 = OR(A, B)$$

$$\theta_3 = (P(C = 1|A = 1, B = 1), P(C = 1|A = 1, B = 2),$$

$$P(C = 1|A = 2, B = 1), P(C = 1|A = 2, B = 2), P(D = 1|A = 1, B = 1),$$

$$P(D = 1|A = 1, B = 2), P(D = 1|A = 2, B = 1), P(D = 1|A = 2, B = 2))$$

$$\theta_4 = (COR(C, D|A = 1, B = 1), COR(C, D|A = 1, B = 2),$$

$$COR(C, D|A = 2, B = 1), COR(C, D|A = 2, B = 2)).$$

In this example, θ_1 is equivalent to the marginal distributions of variables A and of B and θ_3 gives the conditional distributions of C and of D , given any possible category combinations of A and B . The parameter θ_2 is the odds ratio in the marginal distribution $A \times B$, and θ_4 is the collection of the conditional odds ratios of C and D , given all possible category combinations of A and B . Thus, θ_1 and θ_2 determine

the $A \times B$ marginal distribution, and θ_3 and θ_4 determine the $C \times D$ conditional distribution, given $A \times B$.

Here, θ_1 and θ_2 are variation independent. Further, θ_3 is variation independent of θ_1 and θ_2 , and θ_4 is variation independent of all the other three parameters. But also, θ_2 and θ_4 are variation independent of θ_1 and θ_3 .

A statistical model may be obtained by fixing the values of some parameters of a parameterization. If variation independence between the fixed and the other parameters holds, this has no implication for the possible values of the other parameters, that is, there will be exactly one distribution in the model for every possible value of the other parameters. In other words, the unrestricted (components of the) parameter parameterize the model obtained by restricting the other (components of the) parameter.

This is illustrated most easily with the parameterization of a 2×2 , $A \times B$ table with $\eta_1 = (p_{1+}, p_{+1})$ and η_2 the odds ratio $\text{OR}(A, B)$. As η_1 and η_2 are variation independent, if one defines a model by imposing $\eta_2 = 1$, i.e., the independence of A and B , then there is exactly one independent distribution for every choice of the marginal probabilities in η_2 . Related models are obtained by fixing the odds ratio at a different value, see Rudas [71] and Rudas and Leimer [79].

In the case of the 2×2 example above, when η_2 is fixed at 1 and one obtains the model of independence for the 2×2 table, the number of degrees of freedom is 1. It may sound counter-intuitive that when more parameters are fixed by the model, then the number of degrees of freedom is higher. To accept this, one has to remember that the number of degrees of freedom is related to the amount of deviation between observed and expected frequencies tolerated before one would decide the data provide evidence against the model. The more restrictive is the model, the larger is the deviation between observed and expected frequencies that one is ready to tolerate without rejecting the model. Thus, to have a testing procedure with fixed type I error probability, one wishes to use critical values which increase monotonically with the number of parameters fixed by the model, and chi-squared distributions with larger degrees of freedom have larger critical values. Therefore, the number of degrees of freedom associated with a model is not related to the parameters left free, rather it is related to (more precisely, is equal to) the number of parameters fixed by the model.

In the case of the 4-dimensional example, setting $\theta_2 = 1$ implies (3.1) and setting $\theta_4 = (1, 1, 1, 1)$ implies (3.2), yielding the graphical model associated with the graph in Fig. 3.1. These two parameters are variation independent of the other two parameters, thus, θ_1 and θ_3 parameterize all distributions which are Markov according to the graph in Fig. 3.1. Further, as there are 5 parameter values fixed by the model (θ_2 and θ_4), the standard Pearson and likelihood ratio statistics have an asymptotic chi-squared distribution on 5 degrees of freedom, when the model holds true, and data and maximum likelihood estimates are compared to test model fit.

Following Bergsma and Rudas [9], marginal models will be defined in this chapter as a generalization of the procedure above.

3.4 Marginal Log-linear Parameterizations

Marginal log-linear parameters and parameterizations generalize the example discussed for the 4-way table in the last section. Marginal models will be defined by setting some marginal log-linear parameters to zero. The definition of marginal log-linear parameters, and of marginal log-linear models, provides flexible applications which can capture various useful properties of the joint distribution of several categorical variables.

3.4.1 Definition

Let \mathcal{V} be a set of categorical variables, and let $\mathcal{M} \subseteq \mathcal{V}$ denote a marginal. In the sequel, the word marginal will be used, depending on the context, as a subset of the variables, a marginal table of the associated contingency table, or the marginal distribution derived from a joint distribution.

Following Bergsma and Rudas [9], marginal log-linear parameters are defined as log-linear parameters (see e.g., Agresti [1], Bishop et al. [15], and Rudas [75]), calculated in marginals of the table. For simplicity, only distributions with positive cell probabilities are considered in this chapter.

Every subset $\mathcal{E} \subseteq \mathcal{V}$ of the variables may have an effect associated with them, which affects the joint distribution. To emphasize this, subsets of the variables will also be referred to as effects. The strength of the effect (associated with a subset of variables) may be quantified in different ways. A particular quantification is, of course, a parameter. In this section a specific choice of the parameters is used, and alternatives will be discussed later. Many of the properties of the models do not depend on the particular choice of the parameters; however, this becomes relevant when estimated parameter values are used to describe distributions in the model.

A classical log-linear parameter (see e.g., Agresti [1], Bishop et al. [15], and Rudas [75]) for an effect, \mathcal{E} associates a value with every category combination e of the variables \mathcal{E} , denoted as $\lambda_e^{\mathcal{E}}$. These parameters are defined via the following recursion:

$$\lambda^{\emptyset} = \frac{1}{c_{\mathcal{V}}} \sum_v \log P(v),$$

$$\lambda_e^{\mathcal{E}} = \frac{1}{c_{\mathcal{V} \setminus \mathcal{E}}} \sum_{v:(v)_{\mathcal{E}}=e} \log P(v) - \sum_{\mathcal{F} \subsetneq \mathcal{E}} \lambda_{(e)_{\mathcal{F}}}^{\mathcal{F}} \quad (3.4)$$

where e is a joint category of the variables \mathcal{E} , $c_{\mathcal{V} \setminus \mathcal{E}}$ denotes the number of joint categories of the variables in $\mathcal{V} \setminus \mathcal{E}$, and $(v)_{\mathcal{E}}$ denotes the categories out of v which belong to the variables in \mathcal{E} .

When all the variables are binary, the log-linear parameters can be shown to be equal to various averages of the logarithms of the roots of $(l - 1)$ th order conditional

odds ratios of the l variables in \mathcal{E} , given all possible category combinations of the variables $\mathcal{V} \setminus \mathcal{E}$, (see, e.g., Rudas [75] and Section 6 of this chapter). For example, in a binary $A \times B \times C$ table

$$\begin{aligned}
 \lambda_{21}^{AB} &= \frac{1}{2} (\log P(2, 1, 1) + \log P(2, 1, 2)) \\
 &- \frac{1}{4} (\log P(2, 1, 1) + \log P(2, 1, 2) + \log P(2, 2, 1) + \log P(2, 2, 2)) \\
 &- \frac{1}{4} (\log P(1, 1, 1) + \log P(1, 1, 2) + \log P(2, 1, 1) + \log P(2, 1, 2)) \\
 &+ \frac{1}{8} (\log P(1, 1, 1) + \log P(1, 1, 2) + \log P(1, 2, 1) + \log P(1, 2, 2)) \\
 &+ \log (P(2, 1, 1) + \log P(2, 1, 2) + \log P(2, 2, 1) + \log P(2, 2, 2)) \\
 &= \log \sqrt[8]{\frac{P(1, 2, 1)P(1, 2, 2)P(2, 1, 1)P(2, 1, 2)}{P(1, 1, 1)P(1, 1, 2)P(2, 2, 1)P(2, 2, 2)}} \\
 &= \frac{1}{2} \left(\log \sqrt[4]{\frac{P(1, 2, 1)P(2, 1, 1)}{P(1, 1, 1)P(2, 2, 1)}} + \log \sqrt[4]{\frac{P(1, 2, 2)P(2, 1, 2)}{P(1, 1, 2)P(2, 2, 2)}} \right) \\
 &= \frac{1}{2} \left(\log \sqrt[4]{COR(A, B|C=1)} + \sqrt[4]{COR(A, B|C=2)} \right). \tag{3.5}
 \end{aligned}$$

In general, the log-linear parameter for every effect will be considered as vector-valued, with one component for every category combination of the variables in \mathcal{E} , except for the combinations where any of the variables is in, say, its first category, in order to avoid linear dependence of the components of the parameter. So if $\mathcal{E} = \{V_1, V_2, \dots, V_l\}$, and these variables have c_1, c_2, \dots, c_l categories, then the log-linear parameter has

$$(c_1 - 1)(c_2 - 1) \cdots (c_l - 1) \tag{3.6}$$

components and these components are, in general, linearly independent.

The $(l - 1)$ th order conditional odds ratio of the variables in \mathcal{E} (when conditioned on any category combination of the variables $\mathcal{V} \setminus \mathcal{E}$) is variation independent of the marginal distributions of the variables in any proper subset of \mathcal{E} , see, e.g., Rudas [75]. This was illustrated above for the simple cases of 2- and 4-way tables. Therefore, the log-linear parameters, which are functions of the conditional odds ratios, are widely used as measures of the amount of association within an effect, that cannot be attributed to a proper subset of the variables in the effect.

Calculating the log-linear parameter for \mathcal{E} in a marginal \mathcal{M} , with $\mathcal{E} \subseteq \mathcal{M}$, means that the marginal probabilities of \mathcal{M} are used, instead of the joint probabilities of \mathcal{V} . For example, in a 4-way binary $A \times B \times C \times D$ table, in the $A \times B \times C$ marginal, the value of the marginal log-linear parameter for the AB effect is

$$\frac{1}{2} \sum_{k=1}^2 \log \left(\frac{P(1, 1, k, +)P(2, 2, k, +)}{P(1, 2, k, +)P(2, 1, k, +)} \right)^{1/4}. \tag{3.7}$$

The value in (3.7) is denoted as λ_{AB}^{ABC} and in general as $\lambda_{\mathcal{E}}^{\mathcal{M}}$. The parameter λ_{AB}^{ABC} is a measure of average (over the categories of C) conditional association between variables A and B , calculated in the $A \times B \times C$ marginal of the four-way distribution.

The parameter λ_{AB}^{ABC} has a single value, as both A and B are binary. If, for instance, B has three categories, so the table is of the size $2 \times 3 \times 2 \times 2$, then λ_{AB}^{ABC} has 2 components, one for the (2, 2) and one for the (2, 3) indices of A and B . Out of these, the one associated with (2, 2) is as given in (3.7), and the one associated with (2, 3) depends on the type of odds ratio selected, see Sect. 3.6. Such choices are governed by the characteristics of the research question.

Let

$$\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_k$$

be a sequence of marginals, such that

$$\mathcal{M}_j \not\subseteq \mathcal{M}_i, \text{ if } i < j$$

and

$$\mathcal{M}_k = \mathcal{V}.$$

Such a sequence will be called *non-decreasing*. Marginal log-linear parameters calculated in such sequences of marginals play a central role in this chapter.

To define the marginal log-linear parameters, for every effect \mathcal{E} , let $\mathcal{M}(\mathcal{E})$ be the first marginal in the non-decreasing order, that contains it:

$$\mathcal{M}(\mathcal{E}) = \mathcal{M}_i \text{ if } \mathcal{E} \subseteq \mathcal{M}_i \text{ and } \mathcal{E} \not\subseteq \mathcal{M}_j \text{ if } j < i. \tag{3.8}$$

Let now $\lambda_{\mathcal{E}}^{\mathcal{M}(\mathcal{E})}$ denote the log-linear parameter of the effect \mathcal{E} calculated within the $\mathcal{M}(\mathcal{E})$ marginal. This is a log-linear parameter calculated not in the joint distribution of all variables, but rather in a marginal distribution. As illustrated above, the parameter is usually vector-valued, but this fact will be suppressed in the sequel. Marginal log-linear parameters measure the strength of marginal and conditional associations at the same time. By the choice of the marginal, in which the parameter for an effect is defined, some variables are disregarded, and then one conditions upon the variables which are in the marginal but do not belong to the

effect. For further discussion of conditional and marginal association, see Bergsma and Rudas [11]. The marginal log-linear parameters $\{\lambda_{\mathcal{E}}^{\mathcal{M}(\mathcal{E})}, \mathcal{E} \subseteq \mathcal{V}\}$ are called *hierarchical and complete*.

The marginal log-linear parameters, as defined here, obviously contain as special cases the ordinary log-linear parameters, but also the multivariate logistic parameters introduced by Glonek and McCullagh [59] and McCullagh and Nelder [40] as well as the mixed parameters considered in Glonek [39].

Note that the parameters defined at the end of the previous section are one-to-one functions of marginal log-linear parameters. In this example, $\mathcal{V} = \{A, B, C, D\}$, $\mathcal{M}_1 = \{A, B\}$, and $\mathcal{M}_2 = \{A, B, C, D\}$. Thus $\mathcal{M}(\emptyset) = \mathcal{M}(A) = \mathcal{M}(B) = \mathcal{M}(A, B) = \mathcal{M}_1$ and $\mathcal{M}(C) = \mathcal{M}(A, C) = \mathcal{M}(B, C) = \mathcal{M}(A, B, C) = \mathcal{M}(D) = \mathcal{M}(A, D) = \mathcal{M}(B, D) = \mathcal{M}(A, B, D) = \mathcal{M}(C, D) = \mathcal{M}(A, C, D) = \mathcal{M}(B, C, D) = \mathcal{M}(A, B, C, D) = \mathcal{M}_2$. The parameters specified are one-to-one functions of the marginal log-linear parameters of the effects. In particular, setting θ_2 and θ_4 equal to 1 is the same as setting

$$\lambda_{AB}^{AB} = 0 \text{ and } \lambda_{CD}^{ABCD} = 0.$$

In order to obtain analytical properties of marginal log-linear parameters, it is important to remove from among them those which are redundant in the sense that they can be calculated from the others; this is assumed to be the case throughout the whole chapter. The formula in (3.6) only took the non-redundant values of the parameter into account.

3.4.2 Basic Properties

Marginal log-linear parameters have a number of desirable properties. With $\mathcal{M}(\mathcal{E})$ defined by (3.8):

Theorem 3.1 *The parameters $\{\lambda_{\mathcal{E}}^{\mathcal{M}(\mathcal{E})} : \mathcal{E} \subseteq \mathcal{V}\}$ constitute a parameterization of the joint distribution of the variables \mathcal{V} .*

Proof This is part of Theorem 2 in Bergsma and Rudas [9]. Technically, the proof is based on repeated applications of the Iterative Proportional Scaling procedure to determine joint distributions based on mixed parameterizations of exponential families, see, e.g., Rudas [75]. The general relevant result on mixed parameterizations is given by Barndorff-Nielsen [4].

The argument in the proof above also implies that for any $1 \leq i \leq k$, the marginal log-linear parameters calculated in $\mathcal{M}_1, \dots, \mathcal{M}_i$ can be used to determine the joint distribution of the variables in \mathcal{M}_i . This implies the following result.

Theorem 3.2 *If $\mathcal{M}_i \setminus \cup_{j<i} \mathcal{M}_j \neq \emptyset$, then the marginal log-linear parameters $\{\lambda_{\mathcal{E}}^{\mathcal{M}(\mathcal{E})} : \mathcal{M}(\mathcal{E}) = \mathcal{M}_i\}$ determine the conditional joint distribution of the variables in $\mathcal{M}_i \setminus \cup_{j<i} \mathcal{M}_j$, given the joint distributions of variables in $\mathcal{M}_i \cap (\cup_{j<i} \mathcal{M}_j)$.*

To illustrate Theorem 3.2, for the variables A, B, C let $\mathcal{M}_1 = \{AB\}$ and $\mathcal{M}_2 = \{ABC\}$. Then $\mathcal{M}_2 \setminus \mathcal{M}_1 = \{C\}$, and the effects which have their marginal log-linear parameters calculated in \mathcal{M}_2 are C, AC, BC, ABC and they parameterize the conditional distribution of C , given the joint distribution AB . As the joint distribution of AB is parameterized in the marginal \mathcal{M}_1 , the marginal log-linear parameters in the two marginals parameterize the ABC joint distribution. The marginal log-linear parameters determined in the two marginals are variation independent.

For a less straightforward example, let $\mathcal{M}_1 = \{A\}$, $\mathcal{M}_2 = \{B\}$, and $\mathcal{M}_3 = \{ABC\}$. In this case, the theorem is about the conditional distribution of C , given AB , but now the marginal log-linear parameters determined in $\mathcal{M}_1 = \{A\}$ and $\mathcal{M}_2 = \{B\}$ do not determine the AB joint distribution, only its 1-way marginal distributions. In this case, out of the marginal log-linear parameters determined in \mathcal{M}_3 , those belonging to the effects C, AC, BC, ABC determine the conditional distribution. This is most easily seen by including the AB marginal as the third one, which would not change these parameters but would make the setup essentially the same as in the previous example, because in this case the AB joint distribution would be parameterized before the parameters in the $\{ABC\}$ marginal are calculated.

If, however, $\mathcal{M}_1 = \{AB\}$, $\mathcal{M}_2 = \{AC\}$, and $\mathcal{M}_3 = \{ABC\}$, then $\mathcal{M}_3 \setminus (\mathcal{M}_1 \cup \mathcal{M}_2) = \emptyset$, and Theorem 3.2 does not apply. Indeed, if conditioned on $AB \cup AC$, no conditional distribution remains. What do the marginal log-linear parameters given in \mathcal{M}_3 , which are for the effects BC and ABC , determine? In this case, they determine the parameters which are needed in addition to the AB and AC marginal distributions to parameterize the ABC distribution: the second-order odds ratio of ABC and the conditional odds ratio of B and C , given A , see, e.g., Rudas [75].

It is not true, in general, that all components of a marginal log-linear parameterization would be variation independent. The following result gives a necessary and sufficient condition for the components of a hierarchical and complete marginal log-linear parameterization to be variation independent.

Theorem 3.3 *The components of a hierarchical and complete marginal log-linear parameterization based on a non-decreasing sequence of marginals*

$$\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_k = \mathcal{V}$$

are variation independent, if and only if the following condition holds. Either $k = 2$ or for every $j = 3, \dots, k$, the maximal elements out of

$$\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_j,$$

say

$$\mathcal{H}_1, \mathcal{H}_2, \dots, \mathcal{H}_l$$

are such that either $l = 2$ or for every $3 \leq h \leq l$, there is $1 \leq g = g(h) \leq h - 1$, such that

$$(\mathcal{H}_1 \cup \dots \cup \mathcal{H}_{h-1}) \cap \mathcal{H}_h = \mathcal{H}_g \cap \mathcal{H}_h.$$

Proof This is Theorem 4 in Bergsma and Rudas [9].

The property formulated in the previous theorem is called *ordered decomposability*. If the marginals $\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_k$ are all incomparable with respect to inclusion, thus all are maximal, then ordered decomposability means the standard decomposability concept, see, e.g., Rudas [75].

For example, in the case discussed last, with $\mathcal{M}_1 = \{AB\}$, $\mathcal{M}_2 = \{AC\}$, and $\mathcal{M}_3 = \{ABC\}$, ordered decomposability holds. But if $\mathcal{M}_1 = \{AB\}$, $\mathcal{M}_2 = \{AC\}$, $\mathcal{M}_3 = \{BC\}$, and $\mathcal{M}_4 = \{ABC\}$, ordered decomposability does not hold, and it is easy to find values of the marginal log-linear parameters defined in \mathcal{M}_1 and \mathcal{M}_2 , which restrict the range of the parameters in \mathcal{M}_3 ; see Bergsma and Rudas [9]. The three 2-way marginal (frequency) distributions presented in Table 3.3 are weakly compatible but not strongly compatible, that is, although the generated 1-way marginals are all uniform, there is no 3-way distribution with these marginals.

Indeed, if one had such a distribution, one would have for the frequencies that $f(1, 1, 2) \leq 1$ (from the BC marginal) and $f(1, 2, 2) \leq 1$ (from the AB marginal), but the sum of these two frequencies would have to be 3 (from the AC marginal). This means that the three 2-way marginals, and consequently the corresponding marginal log-linear parameters, are not variation independent.

This is an important difference between the standard and the marginal log-linear parameters. If the log-linear parameterization is calculated in the $\{ABC\}$ table, that is one has standard log-linear parameters, the parameter belonging to the BC effect is essentially the conditional odds ratio $COR(B, C|A = a)$ and this is variation independent of the AB and AC marginal distributions. But if a marginal log-linear parameterization is considered based on the marginals $\{AB\}$, $\{AC\}$, and $\{BC\}$, then the parameter belonging to the BC effect is the marginal odds ratio $OR(B, C)$ and this is not variation independent of the AB and AC marginal distributions.

Table 3.3 Marginal distributions which are weakly compatible but not strongly compatible

	B = 1	B = 2		C = 1	C = 2		C = 1	C = 2
A = 1	3	1	A = 1	1	3	B = 1	3	1
A = 2	1	3	A = 2	3	1	B = 2	1	3

Table 3.4 Structure of a distribution with AB and AC marginals as given in Table 3.3

A = 1	C = 1	C = 2	A = 2	C = 1	C = 2
B = 1	t	3 - t	B = 1	u	1 - u
B = 2	1 - t	t	B = 2	3 - u	u

To have the AB and AC marginal distributions as prescribed in Table 3.3, the 3-way table has to have the structure shown in Table 3.4, implying that $t \leq 1$ and $u \leq 1$. The conditional odds ratios are

$$\text{COR}(B, C|A = 1) = \frac{t^2}{(1 - t)(3 - t)}$$

and

$$\text{COR}(B, C|A = 2) = \frac{u^2}{(1 - u)(3 - u)}$$

and their values are not restricted, i.e., depending on t and u , may be anywhere on the interval $(0, \infty)$. But the marginal odds ratio is

$$\text{OR}(B, C) = \frac{(t + u)^2}{(4 - t - u)^2}$$

and this is restricted to be not more than 1.

However, even in this case, the marginal log-linear parameters calculated in the marginals $\{AB\}$, $\{AC\}$, and $\{BC\}$ on the one hand, and the parameters calculated in $\{ABC\}$, on the other hand, are variation independent.

3.4.3 Smoothness of Marginal Log-linear Parameters

Marginal log-linear models will be defined by assuming that some marginal log-linear parameters are zero. Many of the statistical properties of these models, including the behaviour of maximum likelihood estimates and asymptotic distributions of test statistics depend on analytical properties of the parameterizations used.

A parameter is called smooth if, as a function of the (probability or frequency) distribution, it is continuous, invertible, twice continuously differentiable, and its Jacobian has full rank everywhere.

Theorem 3.4 *The hierarchical and complete marginal log-linear parameters are a smooth parameterization of the frequency distribution.*

Proof This is Theorem 2 in Bergsma and Rudas [9]. Note that smoothness holds only if the redundant parameter values are omitted.

To obtain a smooth parameterization of the probability distribution, the parameter referring to the empty set, $\lambda_{\emptyset}^{\mathcal{M}(\emptyset)}$, must be omitted because its value is determined by the other parameters through the requirement that the probabilities must sum to 1.

Bergsma and Rudas [9] showed (their Theorem 3) that for two marginals \mathcal{M} and \mathcal{N} and effect $\mathcal{E} \subseteq \mathcal{M} \cap \mathcal{N}$, the partial derivatives of the parameters $\lambda_{\mathcal{E}}^{\mathcal{N}}$ and $\lambda_{\mathcal{E}}^{\mathcal{M}}$ according to the components of the probability distribution, evaluated at the uniform distribution, are equal and therefore these parameters cannot be parts of a smooth parameterization of all distributions, because the partial derivative matrix would not always be of full rank. A more detailed analysis of this issue is given by Colombi and Forcina [22], using a different marginal log-linear parameterization which does not involve averaging over the categories of the conditioning variables as in (3.7).⁴

One has the following result connecting marginal log-linear parameters of the same effect calculated in different marginals.

Theorem 3.5 *Let all the variables be binary, and then each marginal log-linear parameter has one non-redundant value. Let further $\mathcal{E} \subseteq \mathcal{M} \subset \mathcal{N}$. Then*

$$\lambda_{\mathcal{E}}^{\mathcal{N}} = \lambda_{\mathcal{E}}^{\mathcal{M}} + f(\mathbf{A}_{\mathcal{N}|\mathcal{M}}),$$

for some smooth function f , with

$$\mathbf{A}_{\mathcal{N}|\mathcal{M}} = \{\lambda_{\mathcal{F}}^{\mathcal{N}} : \mathcal{F} \subseteq \mathcal{N}, \mathcal{F} \not\subseteq \mathcal{M}\}.$$

Further,

$$f(\mathbf{A}_{\mathcal{N}|\mathcal{M}}) = 0 \text{ if } (\mathcal{N} \setminus \mathcal{M}) \perp\!\!\!\perp A \mid (\mathcal{M} \setminus A) \quad (3.9)$$

for some $A \in \mathcal{E}$.

Proof This is part of Theorem 3.1 in Evans [29].

For example, the second claim of the theorem implies that if $A \perp\!\!\!\perp B \mid C$, then $\lambda_B^{ABC} = \lambda_B^{BC}$. This is directly seen by noting that these log-linear parameters are simple functions of the conditional odds of the categories of B . For the first one, conditioning is on A and C and for the second one conditioning is on C only. But if the conditional independence in (3.9) holds, the conditioning on A does not provide further information after conditioning on C in the sense that

$$P(B = j \mid C = k) = P(B = j \mid A = i, C = k),$$

so the conditional probabilities entering the formulas for the log-linear parameters are the same. In general, this implies that if condition (3.9) holds for a distribution,

⁴ For alternative parameterizations see Sect. 3.5.

then $\lambda_{\mathcal{E}}^{\mathcal{M}}$ and $\lambda_{\mathcal{E}}^{\mathcal{N}}$ cannot be both contained in a smooth parameterization, because then the Jacobian could not be of full rank.

3.4.4 Collapsibility

The final property of marginal log-linear parameterizations that we consider before giving the general definition of marginal log-linear models, is collapsibility.

Collapsibility of a parameter is a desirable property but it cannot always be achieved. The concept of collapsibility has many variants, and it refers to the property that some aspect of the inference from a full table is identical to the corresponding inference based on a marginal table. For example, in a 3-way binary table, $\lambda_{AB}^{ABC} = 0$ does not generally imply that $\lambda_{AB}^{AB} = 0$, so the inference with respect to the strength of association between variables A and B is not the same, whether it is considered in the full table or in the AB marginal.

In general, a marginal log-linear parameter $\lambda_{\mathcal{E}}^{\mathcal{N}}$ would be called collapsible, see e.g., Ghosh and Vellaisamy [38] if, for $\mathcal{M} \subseteq \mathcal{N}$, $\lambda_{\mathcal{E}}^{\mathcal{N}} = \lambda_{\mathcal{E}}^{\mathcal{M}}$ held. Of course, this cannot be true in general, as in this case marginal log-linear parameters would not be different from the standard log-linear ones. Even for a much weaker requirement, called directional collapsibility, where only the direction of the association is retained, Rudas [74] showed that there is essentially only one parameterization of multivariate binary distributions which is directionally collapsible for every distribution, and it is not a log-linear, but rather a linear function of the cell probabilities.

Thus, collapsibility is often interpreted as a property not associated with a parameter, but rather with a parameter and a particular distribution. For example, Ghosh and Vellaisamy [38] gave the following result.

Theorem 3.6 *Let $\emptyset \neq \mathcal{E} \subseteq \mathcal{M} \subsetneq \mathcal{N} \subseteq \mathcal{V}$ be fixed. Then, in the binary case, collapsibility in the sense that*

$$\lambda_{\mathcal{F}}^{\mathcal{M}} - \lambda_{\mathcal{F}}^{\mathcal{N}} = 0, \text{ for all } \mathcal{F} \subseteq \mathcal{E}$$

holds if and only if for the distribution P ,

$$\sum_{\mathcal{F} \subseteq \mathcal{E}} \frac{(-1)^{|\mathcal{E} \setminus \mathcal{F}|}}{2^{|\mathcal{M} \setminus \mathcal{F}|}} \sum_{m: (m)_{\mathcal{F}} = (m^*)_{\mathcal{F}}} d(\mathcal{M}, m) = 0$$

for all category combinations m^ of the variables in \mathcal{M} , where*

$$d(\mathcal{M}, m) = \log P_{\mathcal{M}}(m) - \frac{1}{2^{|\mathcal{N} \setminus \mathcal{M}|}} \sum_{n: (n)_{\mathcal{M}} = m} \log P_{\mathcal{N}}(n).$$

Proof This is part of Theorem 3.1 in Ghosh and Vellaisamy [38].⁵

3.5 Marginal Log-linear Models

Marginal log-linear models are obtained from marginal log-linear parameterizations by applying a linear restriction to the components. If in the example of Sect. 3.2.1, one wishes to assume that the strength of association between the first and second measurements are the same, that is, treatment does not affect association, then this model may be formulated by requiring that

$$\lambda_{A_1 B_1}^{A_1 B_1} = \lambda_{A_2 B_2}^{A_2 B_2}.$$

For example, the graphical model associated with Fig. 3.1, which has been discussed repeatedly, is equivalent to the restrictions in (3.1) and (3.2) in Sect. 3.2.3. Then, in Sect. 3.3.2, a parameterization of the joint distribution of the variables based on the marginals

$$\{AB\}, \{ABC\}, \{ABD\}, \{ABCD\}$$

was considered and it was shown that the restrictions defining the model may be imposed by restricting some resulting parameters. In Sect. 3.4.1 it was mentioned that the restrictions are the same as

$$\lambda_{AB}^{AB} = 0 \text{ and } \lambda_{CD}^{ABCD} = 0. \quad (3.10)$$

This is the marginal log-linear definition of the graphical model associated with the DAG in Fig. 3.1. Sect. 3.9.1 will discuss the marginal log-linear approach to graphical modelling in general.

To define a marginal log-linear model, a non-decreasing sequence of marginals is selected and the implied marginal log-linear parameterization is considered. Remember that only non-redundant parameters are included in the parameterization, which is thus smooth, see Theorem 3.4. In the generality considered in Bergsma and Rudas [9], a marginal log-linear model is obtained by assuming that the parameters belong to a linear subspace of the parameter space and marginal log-linear models are the special case when the subspace is defined by the equality-to-zero assumptions.

These models provide a rich family of generalizations of the log-linear model. The actual meaning of the model depends on the marginals selected and on the restrictions applied. Several examples will be discussed later on in the chapter.

In this section, we concentrate on the general properties of marginal log-linear models. The first property is that these models always exist.

⁵ Note that formula (iii) in Theorem 3.1 in Ghosh and Vellaisamy [38] appears to have a typo.

Theorem 3.7 *A marginal log-linear model based on a non-decreasing ordering of the marginals is never empty.*

Proof This is implied directly by Theorem 7 of Bergsma and Rudas [9].

An example is the uniform distribution over a contingency table, which satisfies any marginal log-linear model referred to in the theorem. Note that variation independence is not required here.

The smoothness of the parameterization (see Theorem 3.4) from which marginal log-linear models are derived implies that the usual desirable asymptotic behaviour holds under multinomial (see, e.g., Rudas [75]) sampling.

Theorem 3.8 *Assume a marginal log-linear model based on a non-decreasing sequence of marginals contains the true distribution. Then, under multinomial sampling, the probability that a unique maximum likelihood estimate of the true distribution (or of its parameters) exists tends to 1 as the sample size goes to infinity. Further, the asymptotic distribution of the maximum likelihood estimator is normal, with expected value equal to the true distribution.*

Proof This follows from Theorem 8 in Bergsma and Rudas [9].

This result also implies the standard asymptotic behaviour of goodness-of-fit statistics.

3.6 Alternative Parameterizations of Marginal Log-linear Models

There are several ways in which odds ratios may be defined and used to parameterize distributions. These lead to alternative definitions of marginal log-linear parameterizations and models, adding further flexibility of interpretation to the approach described in this chapter.

It was illustrated in Sect. 3.4.1 that marginal log-linear parameters are closely related to local odds ratios and their higher dimensional generalizations (see, e.g., Rudas [72, 75]). In fact, the marginal log-linear parameters may be derived from the local $(l - 1)$ th order odds ratios in the marginal tables, where l is the number of variables in the marginal. To define these in marginal tables, let the marginal probabilities in the marginal \mathcal{M} be denoted as $P_{\mathcal{M}}$ and let variable V_j have indices $1, \dots, c_j$. Then, the local odds ratio of order $l - 1$ in the marginal table for every $(i_1, \dots, i_l) : i_j \geq 2, j = 1, \dots, l$, has the form

$$\prod_{m_j \in \{0,1\}, j=1,\dots,l} P_{\mathcal{M}}^{(-1)^{m_1+\dots+m_l}}(i_1 - m_1, \dots, i_l - m_l). \quad (3.11)$$

The expression in (3.11) is a product of probabilities or their reciprocals. The probabilities involved are in the marginal table \mathcal{M} and are associated with adjacent

cells which are obtained by reducing some indices in (i_1, \dots, i_l) by 1. Whether or not (3.11) contains a probability or its reciprocal depends on the parity of the number of indices which were reduced.

Instead of local odds ratios (of any order), spanning cell odds ratios could also be used to define marginal log-linear parameters. For a marginal table \mathcal{M} , the spanning cell odds ratios are the odds ratios in the 2^l subtables spanned by the reference cell with all indices equal to 1 and a cell (i_1, \dots, i_l) , with all indices greater than 1. The spanning cell odds ratios of order $l - 1$ are of the form

$$\prod_{m_j \in \{0, i_j - 1\}, j=1, \dots, l} P_{\mathcal{M}}^{(-1)^m}(i_1 - m_1, \dots, i_l - m_l), \quad (3.12)$$

where m is the number of indices j where $m_j \neq 0$. In this case, the relevant cells are obtained by replacing some indices by 1.

The intuitive meaning of the higher order odds ratios – whether local or spanning cell – is best understood through a recursive definition involving ratios of lower order conditional odds ratios. While local odds ratios measure the strength of association in adjacent cells, and are also relevant when the categories of the variables have orderings, spanning cell odds ratios measure the strength of association when categories are compared to the reference category coded as 1.

Bartolucci et al. [6] considered various marginal interaction parameters which, if calculated in a non-decreasing set of marginals, may also be used to define marginal models. These generalized marginal interactions are contrasts of logarithms of sums of (marginal) probabilities. Note that the marginal log-linear parameters considered so far in this chapter are also contrasts of logarithms of (marginal) probabilities.

The central concept in the definition of the interaction parameters by Bartolucci et al. [6] is the lumped table. While local and spanning cell odds ratios derive binary sub-tables from a marginal table by selecting various subsets of the cells, and then calculate the odds ratios for these subsets, the approach of Bartolucci et al. [6] derives binary sub-tables by collapsing categories of variables. The global and continuation odds ratios resulting from collapsing categories are particularly useful when the variables are ordinal. A table formed by the variables with collapsed categories is called a lumped table.

For example, if one considers a bivariate marginal \mathcal{M} with $I \times J$ categories of the variables, and probabilities $P_{\mathcal{M}}(i, j)$, then for each $i^* = 2, \dots, I$ and $j^* = 2, \dots, J$, one may consider the following quantities:

$$Q_{\mathcal{M}, i^*, j^*}(l, l) = \sum_{i=1, \dots, i^*-1, j=1, \dots, j^*-1} P_{\mathcal{M}}(i, j)$$

$$Q_{\mathcal{M}, i^*, j^*}(l, nl) = \sum_{i=1, \dots, i^*-1, j=j^*, \dots, J} P_{\mathcal{M}}(i, j)$$

$$Q_{\mathcal{M},i^*,j^*}(nl, l) = \sum_{i=i^*, \dots, I, j=1, \dots, j^*-1} P_{\mathcal{M}}(i, j)$$

$$Q_{\mathcal{M},i^*,j^*}(nl, nl) = \sum_{i=i^*, \dots, I, j=j^*, \dots, J} P_{\mathcal{M}}(i, j).$$

Here, the summation of the marginal cell probabilities goes for the indices less (l) or not less (nl) than the specified i^* and j^* .

Then, the lumped table is of the size 2×2 , and the lumped distribution is $Q_{\mathcal{M},i^*,j^*}$. This kind of lumping divides the cells of the marginal table into 4 rectangles and combines the probabilities within each. The odds ratio of the lumped distribution is

$$\frac{Q_{\mathcal{M},i^*,j^*}(l, l)Q_{\mathcal{M},i^*,j^*}(nl, nl)}{Q_{\mathcal{M},i^*,j^*}(l, nl)Q_{\mathcal{M},i^*,j^*}(nl, l)},$$

which is called the global odds ratio belonging to cell (i^*, j^*) . Similar lumping is also possible for l -dimensional tables, and the $(l - 1)$ th order odds ratio in the resulting 2^l table is also called a global odds ratio. There are $(c_1 - 1)(c_2 - 1) \cdots (c_l - 1)$ global odds ratios for an effect \mathcal{E} .

Another type of odds ratio is obtained by the following partial lumping for 2-way $I \times J$ tables, for each $i^* = 1, \dots, I - 1$, and $j^* = 1, \dots, J - 1$:

$$R_{\mathcal{M},i^*,j^*}(e, e) = P_{\mathcal{M},i^*,j^*}(i^*, j^*)$$

$$R_{\mathcal{M},i^*,j^*}(n, e) = P_{\mathcal{M}}(i^* + 1, j^*)$$

$$R_{\mathcal{M},i^*,j^*}(e, m) = \sum_{j=j^*+1, \dots, J} P_{\mathcal{M}}(i^*, j)$$

$$R_{\mathcal{M},i^*,j^*}(n, m) = \sum_{j=j^*+1, \dots, J} P_{\mathcal{M}}(i^* + 1, j),$$

where e stands for equal, n stands for next, and m stands for more than.

The odds ratio obtained for the lumped 2×2 distribution,

$$\frac{R_{\mathcal{M},i^*,j^*}(e, e)R_{\mathcal{M},i^*,j^*}(n, m)}{R_{\mathcal{M},i^*,j^*}(e, m)R_{\mathcal{M},i^*,j^*}(n, e)},$$

is called the continuation odds ratio. Its meaning is best seen by writing it as

$$\frac{R_{\mathcal{M},i^*,j^*}(n, m)/R_{\mathcal{M},i^*,j^*}(n, e)}{R_{\mathcal{M},i^*,j^*}(e, m)/R_{\mathcal{M},i^*,j^*}(e, e)},$$

which is the ratio of the conditional odds of the ‘continuation’ of the second variable, as opposed to not changing it, when conditioned on the next or on the current category of the first variable.

In multivariate generalizations of the continuation odds ratios, lumping does occur for the response variables but not for the explanatory variables, if such a distinction among the variables exists.

Bartolucci et al. [6] define extended interaction parameters as contrasts of logarithms of generalized odds ratios including global and continuation odds ratios (and also local and spanning cell odds ratios) and show that for the models obtained by linear restrictions on these, many of the results presented so far in this chapter apply, too. They called this more general model class *hierarchical marginal models*.

3.7 Marginal Log-linear Parameterization of Conditional Independence Models

Many of the relevant marginal models assume conditional independences in various marginals of the table. The most important group of such models are graphical models, of which models associated with DAGs have already been considered. A more detailed account will be given later in this chapter. In this section, we present general results about formulating conditional independence models as marginal models, that is, by restricting some parameters in a hierarchical and complete marginal log-linear parameterization.

For $i = 1, \dots, h$, let $\mathcal{A}_i \neq \emptyset$, $\mathcal{B}_i \neq \emptyset$, and \mathcal{C}_i be pairwise disjoint sets of variables. The goal is to formulate the following conditional independences jointly, as a marginal log-linear model:

$$\mathcal{A}_i \perp\!\!\!\perp \mathcal{B}_i \mid \mathcal{C}_i, \text{ for all } i = 1, \dots, h. \quad (3.13)$$

For example, the graphical model associated with the DAG in Fig. 3.1 is equivalent to imposing the conditional independences (3.1) and (3.2). In this case, $h = 2$, $\mathcal{A}_1 = \{A\}$, $\mathcal{B}_1 = \{B\}$, $\mathcal{C}_1 = \emptyset$ and $\mathcal{A}_2 = \{C\}$, $\mathcal{B}_2 = \{D\}$, $\mathcal{C}_2 = \{A, B\}$.

To explore when (3.13) may be formulated as a marginal log-linear model, define

$$\mathbb{D}_i = \mathbb{P}(\mathcal{A}_i \cup \mathcal{B}_i \cup \mathcal{C}_i) \setminus [\mathbb{P}(\mathcal{A}_i \cup \mathcal{C}_i) \cup \mathbb{P}(\mathcal{B}_i \cup \mathcal{C}_i)],$$

where $\mathbb{P}(\cdot)$ denotes the power set. That is, for every i , \mathbb{D}_i is the collection of those subsets of $\mathcal{A}_i \cup \mathcal{B}_i \cup \mathcal{C}_i$ that contain variables from both \mathcal{A}_i and \mathcal{B}_i . In the case of the DAG example, $\mathbb{D}_2 = \{CD, ACD, BCD, ABCD\}$. A sufficient condition is given by the following result.

Theorem 3.9 *Let*

$$\mathcal{M}_1, \dots, \mathcal{M}_k = \mathcal{V}$$

be a non-decreasing sequence of marginals with the following property:

$$C_i \subseteq \mathcal{M}(\mathcal{E}) \subseteq A_i \cup B_i \cup C_i, \text{ for all } \mathcal{E} \in \cup_{i=1}^h \mathbb{D}_i. \quad (3.14)$$

Then, the conditional independences in (3.13) define a marginal log-linear model based on these marginals. More specifically, (3.13) holds for a distribution P if and only if

$$\lambda_{\mathcal{E}}^{\mathcal{M}(\mathcal{E})} = 0 \text{ for all } \mathcal{E} \in \cup_{i=1}^h \mathbb{D}_i$$

for this distribution. Further, the distributions in the model are smoothly parameterized by the remaining marginal log-linear parameters:

$$\{\lambda_{\mathcal{E}}^{\mathcal{M}(\mathcal{E})} : \mathcal{E} \notin \cup_{i=1}^h \mathbb{D}_i\}.$$

Proof This is part of Theorem 1 in Rudas et al. [78].

Condition (3.14) means that for any effect \mathcal{E} which contains variables from any two subsets A_i and B_i of variables which are assumed to be conditionally independent, the first marginal in the sequence which contains \mathcal{E} has to be big enough to contain the conditioning set C_i , but has to be small enough to be contained in $A_i \cup B_i \cup C_i$. The condition requires the sequence of marginals to be sufficiently rich.

For example, in the case of the model defined by the DAG in Fig. 3.1, there are various choices of the sequence of marginals with property (3.14). Clearly, the $AB, ABCD$ sequence is one such choice. But $A, AB, ABCD$ or A, B, AB, ABC, ABD , and $ABCD$ are also appropriate sequences of marginals in order to be able to define the model by setting some marginal log-linear parameters to zero. However, Theorem 3.9 does not imply that the DAG model would be a marginal log-linear model based on the sequence of marginals $A, ACD, ABC, ABCD$, because the variables C and D , which have to be conditionally independent, are present together, without their conditioning set AB . In the case of the sequence of marginals A, AB , and $ABCD$, the effects which are to be set to zero to specify the DAG model are

$$AB, CD, ACD, BCD, ABCD;$$

the first one in the AB marginal, and the others in the $ABCD$ marginal. Note that this is the same specification as the one given in (3.10), taking into account that the marginal log-linear parameters are log-linear parameters calculated in a marginal, thus the second equality in (3.10) implies that

$$\lambda_{ACD}^{ABCD} = \lambda_{BCD}^{ABCD} = \lambda_{ABCD}^{ABCD} = 0;$$

see Rudas [75].

The marginal log-linear parameters which parameterize the distributions in the DAG model belong to the following effects:

$$\emptyset, A, B, C, D, AC, BC, ABC, AD, BD, ABD.$$

Out of these, the first two are calculated in the first marginal, the third one in the second marginal, and the rest in the last marginal.

Further applications of Theorem 3.9 will be given in Sect. 3.9.1.

The result in Theorem 3.9 raises the question of how to determine, for a given list of conditional independences, whether a smooth marginal log-linear definition and parameterization of the model is possible. This would require considering the non-decreasing sequences of marginals in which the required conditional independences in (3.13) may be formulated, and to see whether (3.14) holds for any such sequence. An apparent difficulty is that a particular effect \mathcal{E} may be a subset of \mathbb{D}_i for more than one i and, thus, (3.14) may impose several restrictions on $\mathcal{M}(\mathcal{E})$. An obvious necessary condition for the existence of a smooth marginal log-linear definition is the following. If for a subset of the variables \mathcal{E} , $I_{\mathcal{E}}$ denotes the indices from among $1, \dots, h$, for which $\mathcal{E} \subseteq \mathbb{D}_i$, then $\mathcal{M}(\mathcal{E})$ should be such that

$$\cup_{i \in I_{\mathcal{E}}} \mathcal{C}_i \subseteq \mathcal{M}(\mathcal{E}) \subseteq \cap_{i \in I_{\mathcal{E}}} \mathcal{A}_i \cup \mathcal{B}_i \cup \mathcal{C}_i$$

and if

$$\cup_{i \in I_{\mathcal{E}}} \mathcal{C}_i \not\subseteq \cap_{i \in I_{\mathcal{E}}} \mathcal{A}_i \cup \mathcal{B}_i \cup \mathcal{C}_i,$$

then, an appropriate sequence of marginals does not exist and the sufficient condition of Theorem 3.9 does not hold.

Forcina et al. [35] arrived at similar results using a different approach. They proposed an algorithm to decide whether a model defined by a given set of conditional independences admits a marginal log-linear definition in the sense discussed here, and is, thus, smooth.⁶

Forcina [33] discussed further questions related to the smoothness of models defined by various, more general, collections of conditional independence statements, when, for any ordering of the relevant marginals, the marginal log-linear parameters which have to be set to zero in order to obtain the prescribed conditional independences, cannot be specified in the first marginal where the effect occurs. An example discussed in Forcina [33] is for four binary variables and requires that

$$X_1 \perp\!\!\!\perp X_2 \mid X_3,$$

$$X_2 \perp\!\!\!\perp X_3 \mid X_4,$$

$$X_2 \perp\!\!\!\perp X_4 \mid X_1.$$

⁶ A model is called smooth if it admits a smooth parameterization.

He showed the model is smooth, even though condition (3.14) does not hold, and Theorem 3.9 does not apply. Indeed, if, for example, the three marginals appeared in the following order,

$$X_1 X_2 X_3, X_2 X_3 X_4, X_1 X_2 X_4,$$

then $X_4 \notin \mathcal{M}(X_2 X_3) = \{X_1 X_2 X_3\}$. Similarly, all other orderings of the marginals would lead to a violation of (3.14).

Forcina [33] offered an iterative algorithm to construct the distributions in the model based on the mixed parameterization of exponential families, see, e.g. Rudas [75], and proved that the convergence of the algorithm implies smoothness of the model.

3.8 Estimation and Testing

In this section, we describe the maximum likelihood (ML) and the GEE approaches to estimating marginal log-linear models. Both estimation methods provide asymptotically unbiased estimators, but ML estimators have the advantage over GEE ones of being asymptotically efficient. On the other hand, the GEE approach has the advantage of being computationally more efficient, which is important because of the large possible sizes of contingency tables. For example, 8 variables with 5 categories gives a contingency table of size $5^8 = 390,625$. The ML method requires all expected cell frequencies to be estimated, whereas the GEE method only estimates first and second moments of observed marginal frequencies. Nevertheless, the ML method can handle large tables; for example, we found that tables with one million cells can be estimated without too much difficulty.

ML estimators of marginal log-linear models are, in general, not available in closed form and iterative methods need to be used. There are two main approaches. Firstly, there are algorithms based on the approach developed by Aitchison and Silvey [3], who used the Lagrange multiplier technique. Lang and Agresti [49] first used this method for marginal models, and a modification, which seems to have improved practical performance, was given by Bergsma [8]. A difficulty with these methods is that a search is done for a saddle point, hence convergence may be difficult to monitor. Bergsma and Rapcsak [14] resolved this problem by developing an alternative Lagrangian method, which turns the constrained maximization problem into an unconstrained one.

A second approach to ML estimation is to maximize the likelihood parameterized in terms of a hierarchical and complete marginal log-linear parameter vector, for example, using a Fisher scoring algorithm. The drawback of this approach is that it involves ‘iteration within iteration’, that is, at each Fisher scoring step, the cell probabilities need to be computed from the current estimated marginal log-linear parameter (this can be done with the iterative proportional fitting algorithm, which has guaranteed convergence). Therefore, this approach is computationally

burdensome and Lagrange multiplier methods are more attractive. We describe the approach for completeness.

As far as we are aware, the GEE method has not been described in the literature for general marginal models. Section 3.8.5 gives an outline, including a suitable choice of the working covariance matrix.

3.8.1 Matrix Formulation of Marginal Models

Let \mathbf{m} be a vector containing the expected cell frequencies in a contingency table. A marginal log-linear parameter $\boldsymbol{\lambda}$ can be represented as

$$\boldsymbol{\lambda} = \mathbf{B}' \log \mathbf{M}' \mathbf{m} \quad (3.15)$$

where \mathbf{B} and \mathbf{M} are appropriately defined matrices and a prime represents the transpose. This formulation includes the marginal log-linear parameterizations of Bergsma and Rudas [9, 10], see Sect. 3.4.

A marginal log-linear model is then defined by

$$\boldsymbol{\lambda} = \mathbf{X}\boldsymbol{\beta} \quad (3.16)$$

for a matrix \mathbf{X} and parameter vector $\boldsymbol{\beta}$ of smaller length than $\boldsymbol{\lambda}$. Equivalently, a marginal log-linear model can be specified as

$$\mathbf{C}'\boldsymbol{\lambda} = \mathbf{0} \quad (3.17)$$

for an appropriate matrix \mathbf{C} . Taking \mathbf{C} to be the orthogonal complement of \mathbf{X} , in the sense that $\mathbf{C}'\mathbf{X} = \mathbf{0}$ and (\mathbf{X}, \mathbf{C}) is an invertible matrix, the two formulations are seen to be equivalent. These formulations have been called *freedom* and *constraint* specifications, see Lang [46].

For example, consider a 2×2 table with expected cell frequencies $(m_{11}, m_{12}, m_{21}, m_{22})$. The marginal homogeneity model in the constraint specification is $\log(m_{i+}) - \log(m_{+i}) = 0$ ($i = 1, 2$), where a plus in the subscript denotes summation over that subscript. In matrix notation, this is

$$\begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix} \log \left[\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} m_{11} \\ m_{12} \\ m_{21} \\ m_{22} \end{pmatrix} \right] = 0. \quad (3.18)$$

In the freedom specification, the model is $(m_{i+}, m_{+i}) = (\beta_i, \beta_i)$ ($i = 1, 2$), which in matrix notation is

$$\log \left[\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} m_{11} \\ m_{12} \\ m_{21} \\ m_{22} \end{pmatrix} \right] = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}.$$

3.8.2 Characterization of ML Estimators

In this section we give a score equation, and a Lagrangian score equation, whose solutions, under some conditions, are the ML estimators of a marginal model. Algorithms for solving these equations are postponed to Sect. 3.8.4.

Let \mathbf{n} be a vector of observed cell counts of a contingency table. We assume \mathbf{n} has a multinomial or independent Poisson distribution with expected frequency vector $\mathbf{m} = E(\mathbf{n})$. The log-likelihood for \mathbf{m} then is

$$L(\mathbf{m}|\mathbf{n}) = \mathbf{n}' \log(\mathbf{m}) - \mathbf{1}'\mathbf{m} + c \tag{3.19}$$

where $\mathbf{1}$ is a vector of ones of appropriate length and c is a constant. In the multinomial case, the constraint $\mathbf{1}'\mathbf{m} = \mathbf{1}'\mathbf{n}$ holds, but this does not affect maximum likelihood estimation or inference in the present case, see Lang [47]. Hence, for notational simplicity, we will ignore the multinomial constraint below. The maximum likelihood estimator $\hat{\mathbf{m}}$ of \mathbf{m} under a marginal log-linear model maximizes the log-likelihood $L(\mathbf{m}|\mathbf{n})$ subject to a constraint of the form (3.16) or (3.17). The maximum likelihood estimator $\hat{\mathbf{m}}$ of \mathbf{m} has been characterized in two equivalent ways, namely as the solution to (i) equations involving Lagrange multipliers, or (ii) the score equation for β . The former is due to Aitchison and Silvey [3] and Lang [46] and the latter was considered by Glonek and McCullagh [40] and Colombi and Forcina [18].

The Lagrange multiplier method seeks a stationary point of the Lagrangian log-likelihood

$$L(\mathbf{m}, \boldsymbol{\tau}|\mathbf{n}) = \mathbf{n}' \log(\mathbf{m}) - \mathbf{1}'\mathbf{m} - \boldsymbol{\tau}'\mathbf{C}'\boldsymbol{\lambda}$$

where $\boldsymbol{\tau}$ is a vector of Lagrange multipliers and $\boldsymbol{\lambda}$ is a marginal log-linear parameter of the form (3.15). Denote the Jacobian of $\boldsymbol{\lambda}$ as \mathbf{A} , given by

$$\mathbf{A} = \frac{d\boldsymbol{\lambda}'}{d\mathbf{m}} = \mathbf{M}\mathbf{D}_{\mathbf{M}\mathbf{m}}^{-1}\mathbf{B} \tag{3.20}$$

where \mathbf{D} is the diagonal matrix with its subscript on the main diagonal. Differentiating the log-likelihood L with respect to \mathbf{m} and equating to zero gives

$$\frac{\mathbf{n}}{\mathbf{m}} - \mathbf{1} + \mathbf{A}\mathbf{C}\boldsymbol{\tau} = \mathbf{0} \quad (3.21)$$

where the division in \mathbf{n}/\mathbf{m} is element-wise.

Under some conditions, the ML estimator $\hat{\mathbf{m}}$ is a solution to the simultaneous equations (3.21) and (3.17). Sufficient conditions include (i) all observed frequencies are strictly positive, and (ii) the Jacobian $\mathbf{A}\mathbf{C}$ has full column rank. For most, if not all, marginal models of practical interest, the second condition is satisfied; see Sect. 3.4.3. However, the positivity of all observed frequencies is often not satisfied in practice; for example, for many real-world problems the number of cells in the table is larger than the sample size, implying there must be some cells with zero observations. A heuristic solution to this problem is to replace all zero observed frequencies by a small constant, so that the total contribution to the likelihood will be negligible, as is described in Bergsma et al. [12].

To illustrate the problem with zero observed cells, note that for the marginal homogeneity model defined by (3.18), (3.21) becomes

$$\frac{n_{ij}}{m_{ij}} - 1 - \frac{\lambda_i}{m_{i+}} + \frac{\lambda_j}{m_{+j}} = 0 \quad i, j = 1, 2.$$

Consider now the equation for $(i, j) = (1, 1)$. Since $m_{1+} = m_{+1}$, we obtain

$$\frac{n_{11}}{m_{11}} - 1 = 0.$$

The solution is $\hat{m}_{11} = n_{11}$ except if $n_{11} = 0$, in which case there is no solution. The true ML estimator in this case is $\hat{m}_{11} = 0$, and replacing n_{11} by a small number makes negligible difference for inferential purposes.

An alternative to the Lagrange multiplier method for characterizing the ML estimator is by means of the score equation for $\boldsymbol{\beta}$ in (3.16). The likelihood is parameterized in terms of $\boldsymbol{\beta}$ and the ML estimator is obtained by computing the score equation and solving for $\boldsymbol{\beta}$. This approach is facilitated if $\boldsymbol{\lambda}$ is a marginal log-linear parameterization, in which case its Jacobian \mathbf{A} is invertible. Differentiating the log-likelihood then gives the score vector

$$\mathbf{s}(\boldsymbol{\beta}) = \frac{dL}{d\boldsymbol{\beta}} = \frac{d\boldsymbol{\lambda}'}{d\boldsymbol{\beta}} \frac{d\mathbf{m}}{d\boldsymbol{\lambda}'} \frac{dL}{d\mathbf{m}} = \frac{d\boldsymbol{\lambda}'}{d\boldsymbol{\beta}} \left(\frac{d\boldsymbol{\lambda}'}{d\mathbf{m}} \right)^{-1} \frac{dL}{d\mathbf{m}} = \mathbf{X}'\mathbf{A}^{-1} \left(\frac{\mathbf{n}}{\mathbf{m}} - \mathbf{1} \right). \quad (3.22)$$

Provided all observed cell frequencies are positive, the ML estimator $\hat{\boldsymbol{\beta}}$ satisfies $\mathbf{s}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$. As in the Lagrange multiplier case, we suggest replacing zero observed cell frequencies by a small constant. If $\boldsymbol{\lambda}$ is a smooth parameterization, then \mathbf{A} is invertible, and $\mathbf{s}(\boldsymbol{\beta}) = \mathbf{0}$ is equivalent to the Lagrangian equation (3.21), since $\mathbf{C}'\mathbf{X} = \mathbf{0}$ and (\mathbf{X}, \mathbf{C}) is an invertible matrix. The score function is potentially

computationally expensive to evaluate, because the matrix \mathbf{A} needs to be computed and inverted.

3.8.3 Likelihood Ratio Tests and Asymptotic Distribution of ML Estimators

Suppose model (3.16) holds. The setup of Aitchison and Silvey [3] and Lang [46] applies, implying that the maximum likelihood estimator $\hat{\mathbf{m}}$ under this model has an approximate large sample multivariate normal distribution, with mean \mathbf{m} and covariance matrix

$$\text{cov}(\hat{\mathbf{m}}) \approx \mathbf{D}_{\mathbf{m}} - \mathbf{A}(\mathbf{A}'\mathbf{D}_{\mathbf{m}}\mathbf{A})^{-1}\mathbf{A}.$$

The estimated parameter vector $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\log \hat{\mathbf{m}}$ also has a large sample multivariate normal distribution, with mean $\boldsymbol{\beta}$ and covariance matrix

$$\text{cov}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_{\mathbf{m}}^{-1}\text{cov}(\hat{\mathbf{m}} - \mathbf{m})\mathbf{D}_{\mathbf{m}}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.$$

The usual likelihood ratio test can be used for selecting nested models. Let H_0 and H_1 be nested models, i.e., if H_0 is true then H_1 is true, and let $\hat{\mathbf{m}}_k$ be the ML estimate of \mathbf{m} under H_k ($k = 0, 1$). The log likelihood ratio test statistic is

$$G^2 = 2\mathbf{n}' \log \frac{\hat{\mathbf{m}}_0}{\hat{\mathbf{m}}_1}$$

Under some regularity conditions, if H_0 is true then G^2 has an asymptotic chi-square distribution with degrees of freedom (df) equal to the dimension of H_1 minus the dimension of H_0 .

Non-nested models can be compared using various information criteria, such as the Bayesian information criterion (BIC),

$$\text{BIC} = G^2 + 2 \text{df} \log(N)$$

where N is the sample size.

3.8.4 Algorithms for Finding ML Estimators

3.8.4.1 Lagrangian Methods

Several Lagrangian algorithms have been proposed to find the ML estimators of a marginal model. In their seminal paper, Aitchison and Silvey [3] described

Lagrangian methods for constrained maximum likelihood in some generality. Lang and Agresti [49] and Lang [46] introduced Lagrangian methods for categorical marginal models. The algorithm we describe here is a slightly modified algorithm developed by Bergsma [8], which practical experience indicated has improved convergence properties compared to the original Aitchison and Silvey algorithm.

The first step of the algorithm is to choose an appropriate starting point $\mathbf{m}^{(0)}$, after which subsequent estimates $\mathbf{m}^{(k+1)}$ ($k = 0, 1, 2, \dots$) are calculated iteratively using the formula

$$\log \mathbf{m}^{(k+1)} = \log \mathbf{m}^{(k)} + \text{step}^{(k)} \mathbf{u}(\mathbf{m}^{(k)}) \quad (3.23)$$

where $\text{step}^{(k)}$ is an appropriately chosen step size and

$$\mathbf{u}(\mathbf{m}) = \frac{\mathbf{n}}{\mathbf{m}} - \mathbf{1} - \mathbf{A}\mathbf{C}(\mathbf{C}'\mathbf{A}'\mathbf{D}_m\mathbf{A}\mathbf{C})^{-1}[\mathbf{C}'\mathbf{A}'\mathbf{M}'(\mathbf{n} - \mathbf{m}) + \mathbf{C}'\boldsymbol{\lambda}].$$

Here, \mathbf{A} is defined by (3.20) and depends on \mathbf{m} . A suggested starting point is $\mathbf{m}^{(0)} = \mathbf{n} + \epsilon$, where ϵ is some small constant, such as 10^{-6} . For further details, see Bergsma et al. [12, Section 2.3.5].

A closely related algorithm was given by Colombi and Forcina [18], which, being based on updating $\boldsymbol{\beta}$ in (3.16), was named the ‘regression algorithm’. The two algorithms were shown to be equivalent by Evans and Forcina [30]. They showed the two algorithms have rather different numerical properties depending on whether the design matrix \mathbf{X} has a block diagonal structure, arising with continuous covariates: if this is the case, the regression algorithm (3.23) tends to be much more efficient but if not, Bergsma’s algorithm tends to be much more efficient in practice.

Although we have very good practical experience with convergence of the algorithm (3.23) to the ML estimator, theoretical results are lacking. Generally speaking, convergence properties of constrained optimization problems are more difficult to establish than those of unconstrained ones. The Lagrange multiplier method turns a constrained optimization problem into the problem of finding a saddle point of the Lagrangian function, but finding such a saddle point may be more difficult than finding a global (unconstrained) maximum or a minimum. Two ways of reformulating the ML estimation problem for marginal models as an unconstrained optimization problem have been described.

Firstly, Bergsma and Rapcsák [14] provided a general method for turning a constrained optimization problem into an unconstrained one and applied this to ML estimation of marginal models. The advantage of this algorithm is good theoretical properties, and it is similar in computational efficiency to the algorithm defined by (3.23).

The second way is via the Fisher scoring algorithm described next.

3.8.4.2 Fisher Scoring

In this section we build on the Fisher scoring algorithm for marginal models described by Colombi and Forcina [18]. We wish to find the value of $\boldsymbol{\beta}$ in (3.16) maximizing the log-likelihood (3.19). Here, $\boldsymbol{\lambda}$ is a marginal log-linear parameterization as described in Sect. 3.3. Then, by Theorem 3.4, $\boldsymbol{\Lambda}$ defined by (3.20) is invertible. Differentiating the log-likelihood gives the score vector $\mathbf{s}(\boldsymbol{\beta})$ given by (3.22). The Fisher information on $\boldsymbol{\beta}$ is

$$\mathbf{I}(\boldsymbol{\beta}) = -E\left[\frac{d\mathbf{s}(\boldsymbol{\beta})}{d\boldsymbol{\beta}'}\right] = \mathbf{X}'\boldsymbol{\Lambda}^{-1}\mathbf{D}_m^{-1}\boldsymbol{\Lambda}'^{-1}\mathbf{X}.$$

The Fisher scoring algorithm is given by

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + \text{step}^{(k)}\mathbf{I}(\boldsymbol{\beta}^{(k)})^{-1}\mathbf{s}(\boldsymbol{\beta}^{(k)}). \quad (3.24)$$

At each iteration, the vector of expected cell frequencies \mathbf{m} needs to be computed from $\boldsymbol{\lambda}$, which can be done using the iterative proportional fitting algorithm [see Bergsma and Rudas 9]. However, the Newton-Raphson scheme proposed by Glonek and McCullagh [see Bergsma and Rudas 9 40] may be numerically more efficient.

A major potential numerical bottleneck for (3.24) is that $\boldsymbol{\Lambda}$ needs to be stored and inverted at each iteration. In particular, if there are K cells in the table $\boldsymbol{\Lambda}$ is a $K \times K$ matrix. A normally much more efficient algorithm can be obtained by updating $\boldsymbol{\lambda} = \mathbf{X}\boldsymbol{\beta}$ directly. We obtain the updating step

$$\begin{aligned} \boldsymbol{\lambda} &\rightarrow \boldsymbol{\lambda} + \text{step} \mathbf{X} \cdot \mathbf{I}(\boldsymbol{\beta})^{-1}\mathbf{s}(\boldsymbol{\beta}) \\ &= \boldsymbol{\lambda} - \mathbf{X}(\mathbf{X}'\boldsymbol{\Lambda}^{-1}\mathbf{D}_m^{-1}\boldsymbol{\Lambda}'^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Lambda}^{-1}\mathbf{X}'\boldsymbol{\Lambda}^{-1}\mathbf{D}_m^{-1}(\mathbf{n} - \mathbf{m}) \\ &= \boldsymbol{\lambda} - [\boldsymbol{\Lambda}'\mathbf{D}_m\boldsymbol{\Lambda} - \boldsymbol{\Lambda}'\mathbf{D}_m\boldsymbol{\Lambda}\mathbf{C}(\mathbf{C}'\boldsymbol{\Lambda}'\mathbf{D}_m\boldsymbol{\Lambda}\mathbf{C})^{-1}\mathbf{C}'\boldsymbol{\Lambda}'\mathbf{D}_m\boldsymbol{\Lambda}]\mathbf{D}_m^{-1}(\mathbf{n} - \mathbf{m}), \end{aligned}$$

where \mathbf{C} is an orthogonal complement of \mathbf{X} (see (3.17)). In practice, the matrix \mathbf{C} typically has low column rank, making the latter updating step relatively efficient if implemented well; see Colombi and Forcina [18] for details.

Overall, the Fisher scoring algorithm appears more cumbersome to implement than Lagrangian algorithms, in particular if numerical efficiency is desired. Furthermore, due to the required ‘iteration within iteration’, Fisher scoring algorithms can be expected to be slower than Lagrangian algorithms. If parameterizations based on a set of marginals which is not ordered decomposable are used, out-of-range estimates (negative probabilities) can be obtained as described by Colombi and Forcina [18].

In more general settings, a drawback of Fisher scoring is that it requires a parameterization of the distribution in terms of parameters of interest. Such parameterizations are available for marginal log-linear models, but not for the more general models based on non-log-linear parameters considered by Bergsma [8], Lang [48], and Bergsma et al. [12].

3.8.4.3 Software

The following three R packages are available for marginal modelling: `cmm` by Wicher Bergsma and Andries van der Ark, `mph.fit` by Joseph Lang, and `hmmm` by Roberto Colombi, Sabrina Giordano, and Manuela Cazzaro. A detailed description of the `cmm` package can be found at stats.lse.ac.uk/bergsma/cmm/index.html. The website contains R code with explanations for all the data examples in Bergsma et al. [12]. Documentation for `mph.fit` can be found at homepage.stat.uiowa.edu/~jblang/mph.fitting/index.htm and for `hmmm` at rdrr.io/cran/hmmm/; see also Colombi et al. [22]. All three packages can estimate a wide variety of models. A special feature of `cmm` is that it can handle marginal models with latent variables, while `hmmm` can handle hidden Markov models and inequality constraints. For features of `hmmm`, see also Sect. 3.6.

3.8.5 The GEE Method

A drawback of ML estimation of marginal models is that all cells in the contingency table need to be estimated, making it computationally infeasible if the number of cells is large. The GEE method is a quasi-likelihood method which models the covariance matrix between marginal observations, while ignoring higher order associations, allowing greater computational efficiency at the cost of some statistical efficiency. A detailed general overview of the GEE methodology is provided by Molenberghs and Verbeke [60, Chapter 8]. In most literature on GEE, the association is modelled using correlations. Lipsitz et al. [54] developed the GEE methodology based on odds ratios for univariate binary responses. Touloumis et al. [82] gave a more general development for multinomial responses. Below, we adapt the GEE procedure for general marginal models as described in this chapter, i.e., the association is modelled using log-linear parameters and the marginals of interest may be multivariate.

The GEE method derives from the score vector for a generalized linear model for a multivariate marginal mean; if $\mathbf{y} \sim \text{MVN}(\boldsymbol{\mu}, \mathbf{V}_y)$ and $\boldsymbol{\mu} = g(\mathbf{X}\boldsymbol{\beta})$ for some link function g , the score equation yielding the maximum likelihood estimator of $\boldsymbol{\beta}$ is

$$\frac{d\boldsymbol{\mu}'}{d\boldsymbol{\beta}} \mathbf{V}_y^{-1} (\mathbf{y} - \boldsymbol{\mu}) = \mathbf{0}. \quad (3.25)$$

This equation can also yield a consistent estimator of $\boldsymbol{\beta}$ if \mathbf{y} is non-normal [see Wedderburn 84]. However, there is the difficulty that \mathbf{V}_y is typically unknown and potentially difficult to estimate. Liang and Zeger [53] proposed replacing \mathbf{V}_y with a potentially incorrect ‘working’ covariance matrix $\tilde{\mathbf{V}}_y$, giving the GEE

$$\frac{d\boldsymbol{\mu}'}{d\boldsymbol{\beta}} \tilde{\mathbf{V}}_y^{-1} (\mathbf{y} - \boldsymbol{\mu}) = \mathbf{0}. \quad (3.26)$$

Here, $\tilde{\mathbf{V}}_y$ can depend on parameters, in particular $\boldsymbol{\mu}$ and parameters describing the correlation structure of \mathbf{y} . Liang and Zeger [53] showed that under some conditions, the GEE yields a consistent estimator $\tilde{\boldsymbol{\mu}}$ of $\boldsymbol{\mu}$. Huber's [43] large sample sandwich estimator of the covariance matrix of $\tilde{\boldsymbol{\mu}}$ is then

$$\text{cov}(\tilde{\boldsymbol{\beta}}) = \tilde{\mathbf{I}}^{-1} \tilde{\mathbf{J}} \tilde{\mathbf{I}}^{-1}$$

where

$$\tilde{\mathbf{I}} = \left. \frac{d\boldsymbol{\mu}'}{d\boldsymbol{\beta}} \tilde{\mathbf{V}}_y^{-1} \frac{d\boldsymbol{\mu}}{d\boldsymbol{\beta}'} \right|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}} \quad \tilde{\mathbf{J}} = \left. \frac{d\boldsymbol{\mu}'}{d\boldsymbol{\beta}} \tilde{\mathbf{V}}_y^{-1} \mathbf{V}_y^* \tilde{\mathbf{V}}_y^{-1} \frac{d\boldsymbol{\mu}}{d\boldsymbol{\beta}'} \right|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}}.$$

Here, \mathbf{V}_y^* is a consistent estimator of \mathbf{V}_y .

Let us now give the GEE method for estimating $\boldsymbol{\beta}$ in the marginal model (3.16), denoting the marginal observed frequency vector by $\mathbf{y} = \mathbf{M}'\mathbf{n}$ and the corresponding expected frequency vector by $\boldsymbol{\mu} = E(\mathbf{y}) = \mathbf{M}'\mathbf{m}$. Then

$$\mathbf{V}_y = \mathbf{M}'\mathbf{D}_m\mathbf{M} - N^{-1}\boldsymbol{\mu}\boldsymbol{\mu}' \quad (3.27)$$

where $N = \mathbf{1}'\mathbf{n}$ is the sample size. We can write the marginal model (3.16), with $\boldsymbol{\lambda}$ given by (3.15), as

$$\boldsymbol{\mu} = \exp(\mathbf{U}\mathbf{X}\boldsymbol{\beta})$$

where \mathbf{U} is an orthogonal complement of \mathbf{B} , that is, $\mathbf{B}'\mathbf{U} = \mathbf{0}$ and (\mathbf{B}, \mathbf{U}) is an invertible matrix. Hence,

$$\frac{d\boldsymbol{\mu}'}{d\boldsymbol{\beta}} = \mathbf{X}'\mathbf{U}'\mathbf{D}_\mu$$

so that (3.25) becomes

$$\mathbf{X}'\mathbf{U}'\mathbf{D}_\mu \tilde{\mathbf{V}}_y^{-1} (\mathbf{y} - \boldsymbol{\mu}) = \mathbf{0}. \quad (3.28)$$

A difficulty is that \mathbf{V}_y is typically not invertible, in which case we can replace (3.28) by

$$\mathbf{y} - \boldsymbol{\mu} + \mathbf{V}_y \mathbf{D}_\mu^{-1} \mathbf{B}\mathbf{C}\boldsymbol{\tau} = \mathbf{0} \quad (3.29)$$

where $\boldsymbol{\tau}$ is a parameter to be estimated. Straightforward calculations show that if \mathbf{V}_y is invertible, (3.28) and (3.29) are equivalent. Note that (3.29) follows from the Lagrangian score equation (3.21) by pre-multiplying the left- and right-hand sides by $\mathbf{M}'\mathbf{D}_m$. Replacing \mathbf{V}_y in (3.29) by a working covariance matrix $\tilde{\mathbf{V}}_y$ gives a GEE for marginal models. A consistent estimator \mathbf{V}_y^* of \mathbf{V}_y is needed to compute $\tilde{\mathbf{J}}$, and for this we can take $\mathbf{V}_y^* = \mathbf{M}'\mathbf{D}_n\mathbf{M}$.

It remains to find a working covariance $\tilde{\mathbf{V}}_{\mathbf{y}}$. A simple way to do this is as follows. Suppose the marginal model is based on non-nested marginals $\mathcal{M}_1, \dots, \mathcal{M}_k$. Then $\mathbf{V}_{\mathbf{y}}$ given by (3.27) is a function of the expected marginal frequencies for the following marginals

$$\{\mathcal{M}_i \cup \mathcal{M}_j | i, j = 1, \dots, k\}. \quad (3.30)$$

A simple choice of working covariance matrix is obtained by assuming a (potentially incorrect) conditional independence model for the marginal $\mathcal{M}_i \cup \mathcal{M}_j$:

$$(\mathcal{M}_i \setminus \mathcal{M}_j) \perp\!\!\!\perp (\mathcal{M}_j \setminus \mathcal{M}_i) | \mathcal{M}_i \cap \mathcal{M}_j.$$

This gives a closed-form expression for the expected marginal frequencies in the $\mathcal{M}_i \cup \mathcal{M}_j$ in terms of the expected marginal frequencies in the \mathcal{M}_i , so that (3.29) subject to (3.16) can be solved for $\boldsymbol{\beta}$, using, for example, the Newton-Raphson method.

3.8.5.1 Remarks on the GEE Method

If the working covariance matrix is incorrect, the GEE method loses asymptotic efficiency compared to the asymptotically optimal ML method. Above, we proposed a simple working covariance, which for univariate marginals corresponds to an independence working assumption. Touloumis et al. [82] showed that this leads to a potentially big loss of efficiency if there is a strong dependence among the marginal observations. Efficiency can be improved by specifying a working covariance matrix that is closer to the truth, which can be done by specifying and estimating an appropriate parametric model for the marginals in (3.30); Touloumis et al. [82] obtained major improvements for univariate marginal models by modelling the bivariate marginals using homogeneous association models (see Agresti [1], Chapter 9, or Forcina and Kateri [34], for overviews of association models).

The GEE method is a *quasi-likelihood* method. Another popular quasi-likelihood method is *composite likelihood*, which is based on a quasi-likelihood defined by multiplying certain marginal likelihoods; see, e.g., Molenberghs and Verbeke [60, Chapter 9] for an overview. Composite likelihood has the advantage that it can be used both for marginal and conditional models. On the other hand, the GEE method has the advantage that, by improving the specification of the working covariance matrix, its asymptotic efficiency can be arbitrarily close to that of the ML method.

Model comparison using GEE estimation is more difficult than using ML estimation. Model comparison and goodness-of-fit tools were developed by Rotnitzky and Jewell [69], and the quasi-likelihood information criterion (QIC) developed by Pan [66] is particularly popular.

3.9 Areas of Application

3.9.1 Directed Graphical Models

Graphical models for categorical data associated with DAGs, or the more general chain graphs described by Lauritzen [50], are marginal log-linear models in the sense of Bergsma and Rudas [9, 10]. Parameterizations of these models have received considerable attention recently, see Evans and Forcina [61], Marchetti and Lupporelli [58], Németh and Rudas [30], Rudas [73], and Nicolussi and Colombi [64]. For DAGs, the Markov property is

$$V_i \perp\!\!\!\perp \text{nd}(V_i) \mid \text{pa}(V_i). \quad (3.31)$$

Here, for every variable V_i , $\text{nd}(V_i)$ denotes the non-descendants and $\text{pa}(V_i)$ denotes the parents of V_i . The marginal log-linear parameterization of such models given in Rudas et al. [73] is based on a well-numbering of the variables described by Lauritzen et al. [51], such that (3.31) is equivalent to

$$V_i \perp\!\!\!\perp \text{pre}(V_i) \setminus \text{pa}(V_i) \mid \text{pa}(V_i), \quad (3.32)$$

where $\text{pre}(V_i)$ is the set of variables preceding V_i in the well-numbering. The parameterization proposed by Rudas et al. [73] is based on the marginals $\{V_i\} \cup \text{pre}(V_i)$ which allows a parameterization as in Theorem 3.1.

Earlier work on statistical models associated with chain graphs includes Lauritzen and Wermuth [52], Frydenberg [37], Cox and Wermuth [24], Andersson et al. [2], Richardson [31], Wermuth and Cox [24], and Drton [27]. For a component $\mathcal{K} \subseteq \mathcal{V}$ of a chain graph, $\text{ND}(\mathcal{K})$ is the set of nondescendants of \mathcal{K} , i.e., the union of those components, except \mathcal{K} , for which no semi-directed path leads from any node in \mathcal{K} to any node in these components. $\text{PA}(\mathcal{K})$ is the set of parents of \mathcal{K} , i.e., the union of those components from which an arrow points to a node in \mathcal{K} . The set of neighbours of $\mathcal{X} \subseteq \mathcal{K}$, $\text{nb}(\mathcal{X})$, is the set of nodes in \mathcal{K} that are connected to a node in \mathcal{X} and $\text{pa}(\mathcal{X})$ is the set of nodes in \mathcal{K} from which an arrow points to any node in \mathcal{X} .

Chain graph models are defined by combinations of some of the following properties.

- P1 For all components \mathcal{K} , $\mathcal{K} \perp\!\!\!\perp \{\text{ND}(\mathcal{K}) \setminus \text{PA}(\mathcal{K})\} \mid \text{PA}(\mathcal{K})$,
- P2a For all \mathcal{K} and $\mathcal{X} \subseteq \mathcal{K}$, $\mathcal{X} \perp\!\!\!\perp \{\mathcal{K} \setminus \mathcal{X} \setminus \text{nb}(\mathcal{X})\} \mid \{\text{PA}(\mathcal{K}) \cup \text{nb}(\mathcal{X})\}$,
- P2b For all \mathcal{K} and $\mathcal{X} \subseteq \mathcal{K}$, $\mathcal{X} \perp\!\!\!\perp \{\mathcal{K} \setminus \mathcal{X} \setminus \text{nb}(\mathcal{X})\} \mid \text{PA}(\mathcal{K})$,
- P3a For all \mathcal{K} and $\mathcal{X} \subseteq \mathcal{K}$, $\mathcal{X} \perp\!\!\!\perp \{\text{PA}(\mathcal{K}) \setminus \text{pa}(\mathcal{X})\} \mid \{\text{pa}(\mathcal{X}) \cup \text{nb}(\mathcal{X})\}$,
- P3b For all \mathcal{K} and $\mathcal{X} \subseteq \mathcal{K}$, $\mathcal{X} \perp\!\!\!\perp \{\text{PA}(\mathcal{K}) \setminus \text{pa}(\mathcal{X})\} \mid \text{pa}(\mathcal{X})$.

The Type I Markov property (P1, P2a, P3a) is also called the Lauritzen–Wermuth–Frydenberg block-recursive Markov property, see Lauritzen and Wermuth [52] and Frydenberg [37], and the Type II Markov property (P1, P2a, P3b) is

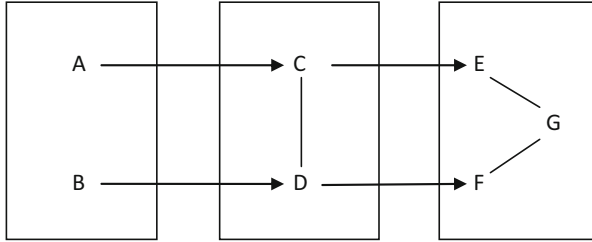


Fig. 3.2 Chain graph whose Andersson–Madigan–Perlman interpretation is a smooth model

also called the Andersson–Madigan–Perlman block-recursive Markov property, see Andersson et al. [2].

Smoothness of Type I models is implied by the results of Frydenberg [37] and is also easily obtained applying Theorem 3.4.

The following example illustrates that marginal log-linear parameterizations may be used to establish smoothness of chain graph models belonging to model classes which also contain nonsmooth models. The graph in Fig. 3.2 with Type II interpretation is a smooth model and may be parameterized using the marginals $AB, ABC, ABD, CDE, CDF, CDG, CDEG, CDFG, CDEFG, ABCDEFG$. Type II models are not smooth in general, see Drton [27], but in this case Theorem 3.4 implies smoothness immediately.

Drton [27] showed that Type IV models (P1, P2b, P3b) are smooth and gave a parameterization. Lupparelli et al. [56] illustrated through examples that these models are marginal log-linear. We now apply the general method in Theorem 3.4 to obtain smoothness based on an interpretable parameterization, also implying the number of degrees of freedom associated with a Type IV model.

Theorem 3.10 *Assuming strictly positive discrete distributions, a Type IV model for a chain graph is a hierarchical marginal log-linear model, and is, therefore, smooth. If the chain graph has components $\mathcal{K}_1, \dots, \mathcal{K}_T$, that are well-numbered, the parameterization is based on the marginals*

$$\{\text{PA}(\mathcal{K}_t) \cup \mathcal{X} : \mathcal{X} \subseteq \mathcal{K}_t\}^*, \mathcal{K}_1 \cup \dots \cup \mathcal{K}_t, t = 1, \dots, T, \tag{3.33}$$

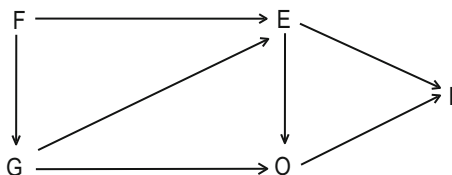
where $\{\ }^*$ denotes a non-decreasing ordering of the elements of the set. The parameters set to zero to define the model are those associated with the effects in

$$\{\mathbf{D}(\mathcal{X}, \mathcal{K}_t \setminus \mathcal{X} \setminus \text{nb}(\mathcal{X}), \text{PA}(\mathcal{K}_t)) : \mathcal{X} \subseteq \mathcal{K}_t\} \cup \tag{3.34}$$

$$\{\mathbf{D}(\mathcal{X}, \text{PA}(\mathcal{K}_t) \setminus \text{pa}(\mathcal{X}), \text{pa}(\mathcal{X})) : \mathcal{X} \subseteq \mathcal{K}_t\} \cup \mathbf{D}(\mathcal{K}_t, \text{PRE}(\mathcal{K}_t) \setminus \text{PA}(\mathcal{K}_t), \text{PA}(\mathcal{K}_t)),$$

for all components \mathcal{K}_t , where $\text{PRE}(\mathcal{K}_t)$ is the set of components that precede \mathcal{K}_t .

Fig. 3.3 The graph of a path model



The proof is given in Rudas et al. [73]. The parameters not set to zero, i.e., the ones not corresponding to (3.34), parameterize the model. These parameters are associated with the same effects as those found by Marchetti and Lupporelli [58] to have non-zero values in the examples they investigated, although the marginals used for the parameterization are different.

Further relevant work includes Marchetti and Lupporelli [58], who described marginal log-linear parameterizations of chain graph models of the multivariate regression type. Evans and Richardson [30] introduced a class of marginal models corresponding to Acyclic Directed Mixed Graphs (ADMGs), which contain both directed and bidirected edges. These models were shown to possess a smooth parameterization, and conditions were given for the parameterization to have a variation independence property. Nicolussi and Colombi [64] considered Type II chain graph models. This class of models, as mentioned above, is known to be not smooth, in general, but, by using a marginal log-linear parameterization, a smooth subclass could be identified.

3.9.2 Path Models

Path models have a long history in statistics and the basic idea is illustrated using Fig. 3.3. Intuitively, one may wish to use path models to describe a situation when variable F influences E and G , G influences E and O , E influences O and I , and O influences I .⁷ In this case, however, one may say that in addition to the direct influence of F on E , F also has an indirect influence on E through G . Similarly, E influences I directly and indirectly. Also, one may wish to assume that only the influences depicted in the graph exist among the variables. There is a further assumption, which is often made but usually remains implicit, namely that variables not taken into account only have negligible influences on those analysed. Path analysis aims at formulating these assumptions precisely and also at quantifying the magnitudes of the influences.

To achieve these goals, motivated by Goodman [41], Rudas et al. [77] proposed the following 2-step approach.

First, interpret the graph in Fig. 3.3 as a graphical Markov model and parameterize the distributions in this model as a marginal log-linear model.

⁷ The notations for the variables are going to be clarified later.

The conditional independences associated with the graph are

$$O \perp\!\!\!\perp F|GE$$

and

$$I \perp\!\!\!\perp FG|EO.$$

These conditional independences may be conveniently imposed in a marginal log-linear model based on the marginals

$$FGEO \text{ and } FGEOI,$$

and are obtained, as implied by Theorem 3.9, by setting to zero the following marginal log-linear parameters⁸

$$\lambda_{FO}^{FGEO}, \lambda_{FEO}^{FGEO}, \lambda_{FGO}^{FGEO}, \lambda_{FGE}^{FGEO},$$

in the $FGEO$ marginal, and also

$$\lambda_{FI}^{FGEOI}, \lambda_{GI}^{FGEOI}, \lambda_{FGI}^{FGEOI}, \lambda_{FEI}^{FGEOI}, \lambda_{FOI}^{FGEOI}, \lambda_{GEI}^{FGEOI},$$

$$\lambda_{GOI}^{FGEOI}, \lambda_{FGEI}^{FGEOI}, \lambda_{FGOI}^{FGEOI}, \lambda_{FEOI}^{FGEOI}, \lambda_{GEOI}^{FGEOI}, \lambda_{FGE}^{FGEOI}$$

in the $FGEOI$ marginal, that is, in the whole table.

There is a total of $2^5 = 32$ parameters, and 16 out of them are set to zero to imply the conditional independences. The remaining parameters parameterize the distributions in the Markov model. The parameters in the present case are interpreted as measuring the strength of influence instead of association because of the inherent assumption behind using a directed graph to formulate the research hypothesis.

The remaining parameters belong to four disjoint groups, depending on the number of variables included in their effects: the marginal log-linear parameter of the empty effect; the marginal log-linear parameters with a single variable in their effects; the marginal log-linear parameters with two variables in their effects; and the marginal log-linear parameters with more than two variables in their effects.

The parameters with more than 2 variables in their effects quantify the joint influence of several variables on one variable, as these parameters are log-linear parameters (determined in a particular marginal) and possess the standard properties of log-linear parameters. For example, λ_{FGE}^{FGEO} , which is not set to zero, is a measure of the joint influence of F and G on E , in addition to their individual influences.

⁸ Setting a parameter to zero means setting it zero for all category combinations of the variables in the effect.

Although the intention of the path model was to assume that such higher-order influences do not exist, their existence is not yet excluded. Indeed, it is very easy to find distributions for, say, three categorical variables, where there are no individual influences (all 2-way marginal distributions are uniform), but two variables together completely determine the third one, illustrating that joint influences on top of the individual influences do exist [see, e.g., Rudas 75].

Therefore, in the next step of the path model definition, such higher-order interactions are excluded.

Second, assume that among the marginal log-linear parameters not set to zero in the first step, all those with more than two variables in their effect are equal to zero.

In the example, this implies setting to zero the following marginal log-linear parameters

$$\lambda_{FGE}^{FGEO}, \lambda_{GEO}^{FGEO},$$

and

$$\lambda_{EOI}^{FGEOI}.$$

To define a path model from the graphical model, a further three parameters are set to zero. This means that the existence of the joint influence of F and G on E and the joint influence of G and E on O in the $FGEO$ marginal, and of the joint influence of E and O on I in the $FGEOI$ marginal, are excluded.

The remaining $32 - (16 + 3) = 13$ marginal log-linear parameters parameterize all the distributions in the path model associated with the graph in Fig. 3.3. These parameters are the univariate distributions of the variables and the strengths of the influences associated with the arrows in Fig. 3.3.

The steps of the definition and parameterization of the model are summarized in Table 3.5.

Table 3.5 The definition and parameterization of the path model associated with Fig. 3.3

Marginals	$FGEO$	$FGEOI$
Effects	$\emptyset, F, G, E, O, FG, FE, FO, GE, GO, EO, FGO, FGE, FEO, GEO, FGEO$	$I, FI, GI, EI, OI, FGI, FEI, FOI, GEI, GOI, EOI, FGEI, FGOI, FEOI, GEOI, FGEOI$
Effects set to zero to define the graphical model	$FO, FEO, FGO, FGEO$	$FI, GI, FGI, FEI, FOI, GEI, GOI, FGEI, FGOI, FEOI, GEOI, FGEOI$
Effects set to zero to define the path model	FGE, GEO	EOI
Remaining effects which parameterize the path model	$\emptyset, F, G, E, O, FG, FE, GE, GO, EO$	I, EI, OI

It has to be pointed out that these parameters provide a parameterization of all distributions in the path model. In a practical data analytic situation, this means that if a particular path model is used, then all relevant information from the data is summarized by the estimates of these parameters obtained from the data. Németh and Rudas [61] provide such an example in the context of social status attainment with variables F —father’s education, G —father’s occupation, E —son’s education, O —son’s occupation, and I —son’s income. They found the path model associated with the graph in Fig. 3.3 well fitting to data for several countries, and gave estimates and interpretations of the parameters of the model. For further details of applications of marginal models to social mobility research, see Németh and Rudas [62].

3.9.3 Latent Variable Models

When some relevant variables in an analysis cannot be observed, i.e., are latent, then the analysis of the observed variables applies to a marginal of the entire table. Therefore, latent variable models and marginal models are closely related. Under certain modelling assumptions, the joint distribution of the latent and observed variables may be estimated, but even in this case, testing of the model has to be restricted to a comparison of the estimated and observed marginal distributions.

For example, if the true position of someone on a left-right political scale, say X , is difficult to observe, then one may ask two related questions, say A and B , which are indicators of X , but may not measure it precisely, rather with some measurement error. Thus, if A and B are equal to X perturbed by measurement errors independent of X and of each other, then one has

$$A \perp\!\!\!\perp B | X. \quad (3.35)$$

Measurement errors are usually assumed to be additive when the observations are numerical. For categorical data, measurement errors may also take different forms. For example, in the case of a binary variable the measurement error may change the category with a given probability. Then, the error is independent of the true category, if the probability of change does not depend on it. Or, for variables with multiple categories, the independent error may alter the category, so that reporting any category other than the true one has the same probability, which does not depend on the true category.

As X is latent, and A and B are observed, (3.35) is a simple latent variable model. As seen above, it has many straightforward marginal log-linear model definitions. One can use the X and XAB marginals, or the X , A , B , XAB marginals, but the definition may also be based on the XAB whole table. In either case, the model is defined as

$$\lambda_{AB}^{XAB} = \lambda_{XAB}^{XAB} = 0.$$

To test the latent variable model (3.35), one has to rely on the observed data for the AB marginal. The usual procedure is to specify the number of categories of the latent variable X and to obtain estimates for the distribution AB , subject to (3.35) so that the likelihood of the observed data is maximized. To determine such estimates, usually the EM algorithm is applied, see, e.g., Rudas [75]. Then, the estimates and the actual observations are compared using some statistical test.

Several more involved applications of the marginal modelling approach to latent variable models are described by Bergsma et al. [12], Bergsma et al. [13], and Hagenaaers et al. [42]. In one problem, there are two latent variables, Y and Z , which are related. Their example refers to election forecasting and Y is political party preference out of three parties and Z is candidate preference, out of their respective candidates. These are seen as latent variables, which may only be observed in an imprecise way. The observed variables are the responses in two waves of a panel study to the party preference (A and B) and to the candidate preference (C and D) questions. This setup may be seen as an instance of the repeated measurement designs considered in Sect. 3.2.1. The model they fit is a graphical model of the path analysis type in the sense that the highest order interactions allowed are YZ, YA, YB, ZC, ZD .

To provide a marginal log-linear definition of this model, one may use the marginals $YZ, XA, XB, ZC, ZD, YZABCD$ and set all marginal log-linear parameters which are defined in the $YZABCD$ marginal equal to zero.

Estimates for the univariate marginal distributions of Y and Z show the relative popularities of the parties and of their respective candidates. Bergsma et al. [12] investigate further the hypothesis that these marginal distributions are identical, i.e., the candidates are just as popular as the parties nominating them. This hypothesis is called latent marginal homogeneity.

To formulate latent marginal homogeneity, it is easiest to use the following marginals: $Y, Z, YZ, YA, YB, ZC, ZD, YZABCD$. Here also, the path model is imposed by setting to zero all parameters which are calculated in the $YZABCD$ marginal, and latent marginal homogeneity is imposed by requiring that

$$\lambda_Y^Y = \lambda_Z^Z,$$

which is a marginal log-linear model.

Manifest variables are often considered indicators of the latent variables. The reliability of such an indicator is the extent to which the manifest variable is determined by the latent variable. This, of course, may be measured in many ways; one of these is based on the conditional distribution of the manifest variable, given the latent variable.

To consider a very simple model, let Y and Z be latent variables with manifest indicators A and C respectively. We are not interested now in how the latent or the manifest variables are related, rather only in to what extent Y determines A and to what extent Z determines C . By Theorem 3.2, if the YZ, YA and ZC marginals are used in a marginal log-linear parameterization, then λ_A^{YA} and λ_{YA}^{YA} determine the conditional distribution of A given Y , and λ_C^{ZC} and λ_{ZC}^{ZC} determine the conditional distribution of C given Z .

If, now, all the variables are assumed to have identical categories, like party and candidate preference in the example above, then the requirement that

$$\lambda_A^{YA} = \lambda_C^{ZC} \text{ and } \lambda_{YA}^{YA} = \lambda_{ZC}^{ZC},$$

means equal reliabilities of the two manifest variables.

The strength of this approach to analysing reliability is that it can be combined with any other modelling assumption, given that the relevant marginals may be written in a non-decreasing order. For details and applications see Bergsma et al. [12].

Marginal log-linear models with latent variables have also been considered in the context of capture-recapture models, see Bartolucci and Forcina [5] and Stanghellini and van der Heijden [81]. In this case also, observed variables are not necessarily independent conditionally on the latent variables.

3.9.4 Further Applications and Extensions

This section gives brief summaries of some further theoretical developments and interesting applications published in the literature.

Qaqish and Ivanova [67] consider multivariate logistic parameterizations, which are generalized by the marginal log-linear parameterizations defined above, and provide results for the strong compatibility of such parameters.

Bartolucci and Forcina [5] apply marginal log-linear parameterizations to develop models for the capture-recapture problem. For related work see also Turner [83].

Forcina [32] develops a marginal log-linear parameterization of latent class models with covariates and obtains identifiability results.

Bartolucci et al. [7] develop a Bayesian approach to selecting the model best supported by the data from among a wide class of marginal models defined by equality or inequality constraints on generalized logits or generalized odds ratios. They use the Bayes factor to govern model selection.

Dardanoni et al. [25] analyse intergenerational socioeconomic mobility tables for many countries, to test the monotonicity hypothesis stating that a higher socioeconomic class is never less advantageous than a lower one. They formulate this monotonicity as a marginal model, using the parameterization proposed by Bartolucci et al. [6].

Shpitser et al. [80] develop marginal log-linear parameterizations for nested Markov models and impose sparsity similar to the idea described in Sect. 3.9.2.

Kuijpers et al. [44] propose methods to formulate and test hypotheses for the widely used measure of test score reliability, Cronbach's alpha, as marginal log-linear models. Kuijpers et al. [45] provide standard errors of scalability coefficients in the case when the items are not binary, and also for large numbers of items, using a marginal modelling approach.

Roverato et al. [70] and Lupparelli and Roverato [57] consider *log-mean linear* parameterizations of marginal models for binary data. These are alternative marginal

log-linear parameterizations to the ones considered in the present chapter. Log-mean linear parameterizations have the interesting advantage of a closed-form likelihood.

Bergsma et al. [13] showed how marginal modelling methods can be extended to deal with complex sampling designs; in particular, they analysed a data set collected via a rotating panel design. The analysis there is carried out on data that are partially dependent. Furthermore, they showed how marginal modelling can be used for complex statistical models, giving an example of a data analysis using latent variables and both log-linear and non-log-linear constraints on the cell probabilities.

Colombi and Giordano [20] parameterize the two components of a latent Markov model (the observed time series and the unobserved Markov chain) with marginal log-linear parameters and show that relevant hypotheses may be formulated by setting some to zero.

Colombi and Forcina [21] test inequality hypotheses for marginal log-linear parameters. They propose a likelihood-based procedure to test a set of equality constraints against positive departures from equality (the inequality constraints) and then the latter against the saturated model.

Ntzoufras et al. [65] discuss aspects of Bayesian inference for graphical marginal log-linear models. They provide a strategy to perform Markov chain Monte Carlo to obtain posterior densities. Their method also takes into account the requirement that the parameter values should be selected in a way which provides compatible marginal distributions.

Colombi et al. [23] model the latent behaviour of raters with a binary variable indicating either one of two possible strategies. A marginal parameterization is used to link responses to underlying explanatory factors.

Bon et al. [16] deal with disclosure limitation of sensitive or confidential data. They model the partial information provided by the data custodians as log-linear models on possibly overlapping marginals of a super-table and investigate methods of combining the available information. They also provide an application to Australian housing tenure transition data.

Nicolussi and Cazzaro [63] analyse context specific independences, that is, independences which only hold in certain but not all category combinations of the variables involved, and show that hierarchical multinomial marginal models may be used to model such relationships. For the relationship between context specific independence and marginal modeling also see Colombi and Forcina [19].

References

1. Agresti, A.: *Categorical Data Analysis*, 3rd edn. Wiley, London (2013)
2. Andersson, S.A., Madigan, D., Perlman, M.D.: Alternative Markov properties for chain graphs. *Scand. J. Stat.* **28**, 33–85 (2001)
3. Aitchison, J., Silvey, S.D.: Maximum likelihood estimation of parameters subject to restraints. *Ann. Math. Stat.* **29**, 813–828 (1958)
4. Barndorff-Nielsen, O.: *Information and Exponential Families*. Wiley, New York (1978)

5. Bartolucci, F., Forcina, A.: A class of latent marginal models for capture–recapture data with continuous covariates. *J. Am. Stat. Assoc.* **101**, 786–794 (2006)
6. Bartolucci, F., Colombi, R., Forcina, A.: An extended class of marginal link functions for modelling contingency tables by equality and inequality constraints. *Stat. Sin.* 691–711 (2007)
7. Bartolucci, F., Scaccia, L., Farcomeni, A.: Bayesian inference through encompassing priors and importance sampling for a class of marginal models for categorical data. *Comput. Stat. Data Anal.* **56**, 4067–4080 (2012)
8. Bergsma, W.P. (1997). *Marginal models for categorical data*. Tilburg: Tilburg University Press
9. Bergsma, W., Rudas, T.: Marginal models for categorical data. *Ann. Stat.* **30**, 140–159 (2002)
10. Bergsma, W., Rudas, T.: Variation independent parameterizations of multivariate categorical distributions. In: Cuadras, C.M., Fortiana, J., Rodriguez-Lallena, J.A. (eds.) *Distributions with Given Marginals and Statistical Modelling*, pp. 21–27. Kluwer, Dordrecht (2002)
11. Bergsma, W., Rudas, T.: On conditional and marginal association. *Annales de la Faculte des Sciences de Toulouse* **6**(11), 455–468 (2003)
12. Bergsma, W., Croon, M., Hagenaars, J.A.: *Marginal Models For Dependent, Clustered and Longitudinal Categorical Data*. Springer, New York (2009)
13. Bergsma, W.P., Croon, M.A., Hagenaars, J.A. (2013). Advancements in marginal modeling for categorical data. *Sociol. Methodol.* **43**(1), 141
14. Bergsma, W.P., Rapcsák, T.: An exact penalty method for smooth equality constrained optimization with application to maximum likelihood estimation. *Eurandom Technical Report* (2005)
15. Bishop, Y.M.M., Fienberg, S.E., Holland, P.W.: *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge (1975)
16. Bon, J., Baffour, B., Spallek, M., Haynes, M.: Analysing sensitive data from dynamically-generated overlapping contingency tables. *J. Off. Stat.* **36**, 275–296 (2020)
17. Cocchi, M. (ed.): *Data Fusion Methodology and Applications*. Elsevier, Amsterdam (2019)
18. Colombi, R., Forcina, A.: Marginal regression models for the analysis of positive association of ordinal response variables. *Biometrika* **88**, 1007–1019 (2001)
19. Colombi, R., Forcina, A.: A class of smooth models satisfying marginal and context specific conditional independencies. *J. Multivariate Anal.* **126**, 75–85 (2014)
20. Colombi, R., Giordano, S.: Multiple hidden Markov models for categorical time series. *J. Multivariate Anal.* **140**, 19–30 (2015)
21. Colombi, R., Forcina, A.: Testing order restrictions in contingency tables. *Metrika* **79**, 73–90 (2016)
22. Colombi, R., Giordano, S., Cazzaro, M.: hmmm: an R package for hierarchical multinomial marginal models. *J. Stat. Softw.* **59**(11), 1–25 (2014)
23. Colombi, R., Giordano, S., Gottard, A., Iannario, M.: Hierarchical marginal models with latent uncertainty. *Scand. J. Stat.* **46**, 595–620 (2019)
24. Cox, D.R., Wermuth, N.: *Multivariate Dependencies*. Chapman and Hall, London (1996)
25. Dardanoni, V., Fiorini, M., Forcina, A.: Stochastic monotonicity in intergenerational mobility tables. *J. Appl. Econom.* **27**, 85–107 (2012)
26. D’Orazio, M., Di Zio, M., Scanu, M.: Statistical matching for categorical data: displaying uncertainty and using logical constraints. *J. Off. Stat.* **22**, 137–157 (2006)
27. Drton, M.: Discrete chain graph models. *Bernoulli* **15**(3), 736–753 (2009)
28. Eppmann, H., Krügener, S., Schäfer, J.: First German register based census in 2011. *Allgemeines Statistisches Archiv* **90**(3), 465–482 (2006)
29. Evans, R.J.: Smoothness of marginal log-linear parameterization. *Electron. J. Stat.* **9**, 475–491 (2015)
30. Evans, R.J., Forcina, A.: Two algorithms for fitting constrained marginal models. *Comput. Stat. Data Anal.* **66**, 1–7 (2013)
31. Evans, R.J., Richardson, T.S.: Marginal log-linear parameters for graphical Markov models. *J. R. Stat. Soc. B: Stat. Methodol.* **75**(4), 743 (2013)
32. Forcina, A.: Identifiability of extended latent class models with individual covariates. *Comput. Stat. Anal.* **52**, 5263–5268 (2008)

33. Forcina, A.: Smoothness of conditional independence. *J. Multivariate Anal.* **106**, 49–56 (2012)
34. Forcina, A., Kateri, M.: A new general class of RC association models: estimation and main properties. *J. Multivariate Anal.* **184**, 1–16 (2021)
35. Forcina, A., Lupporelli, M., Marchetti, G.M.: Marginal parameterizations of discrete models defined by a set of conditional independencies. *J. Multivariate Anal.* **101**, 2519–2527 (2010)
36. Frees, E.W., Kim, J.-S.: Panel studies. In Rudas, T. (ed.) *Handbook of Probability: Theory and Applications*, pp. 205–224. Sage, Thousand Oaks (2008)
37. Frydenberg, M.: The chain graph Markov property. *Scand. J. Stat.* **17**, 333–353 (1990)
38. Ghosh, S., Vellaisamy, P.: Marginal log-linear parameters and their collapsibility for categorical data (2019). arXiv 1711.00680v4
39. Glonek, G.F.V.: A class of regression models for multivariate categorical responses. *Biometrika* **83**, 15–28 (1996)
40. Glonek, G.F.V., McCullagh, P.: Multivariate logistic models. *J. R. Stat. Soc. B* **57**, 533–546 (1995)
41. Goodman, L.A.: The analysis of multidimensional contingency tables when some variables are posterior to others: a modified path analysis approach. *Biometrika* **60**, 179–192 (1973)
42. Hagenaaars, J.A., Bergsma, W., Croon, M.: Nonloglinear marginal latent class models. In: *Advances in Latent Class Analysis: A Festschrift in Honor of C. Mitchell Dayton*, vol. 61 (2019)
43. Huber, P.J.: The behavior of maximum likelihood estimates under nonstandard conditions. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **1** 221–233 (1967)
44. Kuijpers, R.E., Ark, L.A., Croon, M.A.: Testing hypotheses involving Cronbach’s alpha using marginal models. *Br. J. Math. Stat. Psychol.* **66**, 503–520 (2013)
45. Kuijpers, R.E., Ark, L.A., Croon, M.A.: Standard errors and confidence intervals for scalability coefficients in Mokken scale analysis using marginal models. *Sociol. Methodol.* **43**, 42–69 (2013)
46. Lang, J.B.: Maximum likelihood methods for a generalized class of log-linear models. *Ann. Stat.* **24**, 726–752 (1996)
47. Lang, J.B.: On the comparison of multinomial and Poisson log-linear models. *J. R. Stat. Soc. B (Methodological)* **58**(1), 253–266 (1996b)
48. Lang, J.B.: Multinomial-Poisson homogeneous models for contingency tables. *Ann. Stat.* **32**, 340–383 (2004)
49. Lang, J.B., Agresti, A.: Simultaneously modelling the joint and marginal distributions of multivariate categorical responses. *J. Am. Stat. Assoc.* **89**, 625–632 (1994)
50. Lauritzen, S.L.: *Graphical Models*. Clarendon Press, Oxford (1996)
51. Lauritzen, S.L., Dawid, A.P., Larsen, B.N., Leimer, H.-G.: Independence properties of directed Markov fields. *Networks* **20**(5), 491–505 (1990)
52. Lauritzen, S.L., Wermuth, N.L.: Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann. Stat.* **17**, 31–57 (1989)
53. Liang, K.Y., Zeger, S.L.: Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22 (1996)
54. Lipsitz, S.R., Laird, N.M., Harrington, D.P.: Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association. *Biometrika* **78**(1), 153–160 (1991)
55. Little, R., Rubin, D.: *Statistical Analysis with Missing Data*, 3rd. edn. Wiley, New York (2019)
56. Lupporelli, M., Marchetti, G.M., Bergsma, W.P.: Parameterizations and fitting of bi-directed graph models to categorical data. *Scand. J. Stat.* **36**(3), 559–576 (2009)
57. Lupporelli, M., Roverato, A.: Log-mean linear regression models for binary responses with an application to multimorbidity. *J. R. Stat. Soc. C (Applied Statistics)* **66**(2), 227–252 (2017)
58. Marchetti, G.M., Lupporelli, M.: Chain graph models of multivariate regression type for categorical data. *Bernoulli* **17**(3), 827–844 (2011)
59. McCullagh, P., Nelder, J.A.: *Generalized Linear Models*, 2nd edn. Chapman and Hall, London (1989)
60. Molenberghs, G., Verbeke, G.: *Models for Discrete Longitudinal Data* (2005)

61. Németh, R., Rudas, T.: On the application of discrete marginal graphical models. *Soc. Methodol.* **43**, 70–100 (2013)
62. Németh, R., Rudas, T.: Discrete graphical models in social mobility research—a comparative analysis of American, Czechoslovakian and Hungarian mobility before the collapse of state socialism. *Bull. Soc. Methodol.* **118**, 5–21 (2013)
63. Nicolussi, F., Cazzaro, M.: Context-specific independencies in hierarchical multinomial marginal models. *Stat. Methods Appl.* **29**, 767–786 (2020)
64. Nicolussi, F., Colombi, R.: Type II chain graph models for categorical data: a smooth subclass. *Bernoulli* **23**, 863–883 (2017)
65. Ntzoufras, I., Tarantola, C., Lupparelli, M.: Probability based independence sampler for Bayesian quantitative learning in graphical log-linear marginal models. *Bayesian Anal.* **14**, 777–803 (2019)
66. Pan, W.: Akaike’s information criterion in generalized estimating equations. *Biometrics* **57**(1), 120–125 (2001)
67. Qaqish, B.F., Ivanova, T.: Multivariate logistic models. *Biometrika* **93**, 1011–1017 (2006)
68. Rhemtulla, M., Little, T.: Tool of the trade: planned missing data designs for research in cognitive development. *J. Cogn. Dev.* **13**(4), 10 (2012). <https://doi.org/1080/15248372.2012.717340>
69. Rotnitzky, A., Jewell, N.P.: Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika* **77**(3), 485–497 (1990)
70. Roverato, A., Lupparelli, M., La Rocca, L.: Log-mean linear models for binary data. *Biometrika* **100**(2), 485–494 (2013)
71. Rudas, T.: Prescribed conditional interaction structure models with application to the analysis of mobility tables. *Q. Quantity* **25**, 345–358 (1991)
72. Rudas, T.: Odds Ratios in the Analysis of Contingency Tables. No 119, Quantitative Applications in the Social Sciences. Sage, Thousand Oaks (1998)
73. Rudas, T.: Informative allocation and consistent treatment selection. *Stat. Methodol.* **7**, 323–337 (2010)
74. Rudas, T.: Directionally collapsible parameterizations of multivariate binary distributions. *Stat. Methodol.* **27**, 132–145 (2015)
75. Rudas, T.: Lectures on Categorical Data Analysis. Springer, New York (2018)
76. Rudas, T., Bergsma, W.: On applications of marginal models to categorical data. *Metron* **42**, 15–37 (2004)
77. Rudas, T., Bergsma, W., Németh, R.: Parameterization and estimation of path models for categorical data. In: Rizzi, A., Vichi, M. (eds.) COMPSTAT 2006, pp. 383–394. Physica Verlag, Heidelberg (2006)
78. Rudas, T., Bergsma, W., Németh, R.: Marginal log-linear parameterization of conditional independence models. *Biometrika* **97**, 1006–1012 (2010)
79. Rudas, T., Leimer, H.-G.: Analysis of contingency tables with known conditional odds ratios or known log-linear parameters. In: Francis, B., Seeberg, G.U.H., van der Heijden, P.G.M., Jansen, W. (eds.) *Statistical Modelling*, pp. 313–322. Elsevier, Amsterdam (1992)
80. Shpitser, I., Evans, R.J., Richardson, T.S., Robins, J.M.: Sparse nested Markov models with loglinear parameters. In: Twenty-ninth Conference on Uncertainty in Artificial Intelligence, pp. 576–585 (2013)
81. Stanghellini, E., van der Heijden, P.G.: A multiple-record systems estimation method that takes observed and unobserved heterogeneity into account. *Biometrics* **60**(2), 510–516 (2004)
82. Touloumis, A., Agresti, A., Kateri, M.: GEE for multinomial responses using a local odds ratios parameterization. *Biometrics* **69**(3), 633–640 (2013)
83. Turner, E.L.: Marginal Modelling of Capture-Recapture Data. Ph.D. Thesis. McGill University Montreal (2007)
84. Wedderburn, R.W.: Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika*, **61**(3), 439–447 (1974)

Chapter 4

Bayesian Inference for Multivariate Categorical Data



Jonathan J. Forster and Mark E. Grigsby

4.1 Introduction

4.1.1 Contingency Tables

Suppose we have a set of multivariate categorical data, where n units have been cross-classified by a number of categorical variables and the counts of the resulting cross-classification presented in a contingency table. Let the set of categorical variables or factors be Γ , resulting in a $|\Gamma|$ -way contingency table.

Following the notation introduced by Darroch et al. [5], the set of cells in the table is the set $I = \prod_{\gamma \in \Gamma} I_{\gamma}$, where I_{γ} is the set of levels of factor γ . A particular cell will be denoted by $\mathbf{i} = (i_{\gamma} : \gamma \in \Gamma)$, the corresponding cell count by $n(\mathbf{i})$, and the cell probability by $p(\mathbf{i})$, where this represents the probability that a particular unit lies in cell \mathbf{i} . The vector of all the cell probabilities will be written \mathbf{p} , and the cell counts \mathbf{n} . The total cell count will be denoted n , where $n = \sum_{\mathbf{i}} n(\mathbf{i})$. The number of cells in the table is $|I| = \prod_{\gamma} |I_{\gamma}|$. This notation is best illustrated by an example:

Suppose we have three variables A , B , and C , where A is binary and B and C have 3 levels, and that these variables cross-classify some data in a 3-way table. In this case, $\Gamma = \{A, B, C\}$, and a cell in the table is therefore $\mathbf{i} = (i_A, i_B, i_C)$ where i_A can take values 1 and 2, and i_B and i_C take values 1, 2, or 3. Hence the cell

J. J. Forster (✉)
Department of Statistics, University of Warwick, Coventry, UK
e-mail: J.J.Forster@warwick.ac.uk

M. E. Grigsby
Proctor and Gamble, The Heights Weybridge, Surrey, UK

which contains the data for variables A and B at level 1 and variable C at level 3 is $\mathbf{i} = (1, 1, 3)$, and the cell probability is $p(\mathbf{i})$.

The typical model for data in a contingency table assumes that a known number of individual units n are assigned at random to a particular cell \mathbf{i} with probability $p(\mathbf{i})$. Therefore the vector of cell counts \mathbf{n} has a *multinomial* distribution, which has probability function

$$f(\mathbf{n}|\mathbf{p}) = n! \prod_i \frac{p(\mathbf{i})^{n(\mathbf{i})}}{n(\mathbf{i})!}. \quad (4.1)$$

4.1.2 Log-Linear Models

One motivation for analysing contingency table data is modelling the associations between classifying variables. Such considerations typically include how variables are conditionally independent or independent of one another. The standard way of doing this is by representing the underlying statistical model as a *log-linear interaction model*. Different association structures, including independence and conditional independence, result from models with different forms, and from varying parameter values within a particular model.

We assume that $n(\mathbf{i})$ is an observation of a multinomial random variable with corresponding vector of cell probabilities $p(\mathbf{i})$. Then, again following [5], we denote the log-linear interaction model

$$\log p(\mathbf{i}) = \sum_{a \subseteq \Gamma} \xi_a(\mathbf{i}_a) \quad \mathbf{i} \in I \quad (4.2)$$

where \mathbf{i}_a is the marginal cell $\mathbf{i}_a = (i_\gamma, \gamma \in a)$. As $p(\mathbf{i})$ is a vector of cell probabilities which sum to 1, a normalising constant ξ_\emptyset is necessary in (4.2).

A saturated model is parameterised by a full set of interaction terms, whereas setting certain ξ_a terms to zero defines a particular non-saturated log-linear model. Hence the non-zero terms define the model, and may take arbitrary values. It is straightforward to write down the number of possible distinct log-linear models for a set of factors Γ ; there are $2^{|\Gamma|}$ possible $a \subseteq \Gamma$, giving rise to $2^{2^{|\Gamma|}}$ different log-linear models. We use m to denote a model corresponding to a set of interaction terms, so each m is a subset of $\mathcal{P}(\Gamma)$, the power set of Γ .

In order to admit more straightforward analyses and calculations involving the log-linear models described above, it is usual to consider a parameterisation of (4.2) where the parameters are identifiable and linearly independent by restricting each $\xi_a = \{\xi_a(\mathbf{i}_a)\}$ to a d_a -dimensional subspace where

$$d_a = \prod_{\gamma \in a} (|I_\gamma| - 1).$$

Then the model is parameterised using a selection of d_a components of each ξ_a , typically

$$\beta_a = \{\xi_a(i_a), i_\gamma > 1 \text{ for all } \gamma \in a\} \quad (4.3)$$

and, in a slight abuse of notation, we let $\beta_m = \{\beta_a, a \in m\}$, with dimensionality $d_m = \sum_{a \in m} d_a$ represent the parameter vector for model m . For ease of notation, we drop the subscript m on β_m and d_m while we consider analysis under a single model.

The log-linear model is expressed in terms of $\log p$, but since these cell probabilities lie in a simplex space, where each $p(i)$ satisfies $0 < p(i) < 1$ and $\sum p(i) = 1$, it is useful to consider a multivariate logit transformation

$$\theta(i) = \log p(i) - \sum_{\ell=1}^r a(\ell) \log p(\ell) \quad (4.4)$$

where $\sum a(\ell) = 1$, $a_\ell \geq 0$. Typically $a_\ell = I[\ell = i_0]$ for some reference cell i_0 (usually a ‘corner-point’ of the contingency table) though $a_\ell = 1/|I|$ (centred logit) can sometimes be more useful. The multivariate logit defined by \mathbf{a} leads to the simple linear constraint $\sum_i a(i)\theta(i) = 0$ and the same inversion to obtain the cell probabilities in terms as a function of the logits,

$$p(i) = \frac{\exp \theta(i)}{\sum \exp \theta(i)}. \quad (4.5)$$

For model m , the model matrix \mathbf{X} of a non-saturated log-linear model relates the vector of multivariate logits $\boldsymbol{\theta}$ to the d_m log-linear model parameters $\boldsymbol{\beta}$. In standard matrix notation, the log-linear model may be expressed as

$$\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}. \quad (4.6)$$

The form of this matrix depends on the logit chosen, and must satisfy $\mathbf{a}^T \mathbf{X} = 0$. Log-linear interaction models described by (4.2) with (4.3) imply model matrices with a particular structure.

4.1.3 Hierarchical, Graphical, and Decomposable Log-Linear Models

Commonly, we do not consider the full set of log-linear interaction models, and instead restrict attention to a smaller subset of these called the *hierarchical log-linear models*. To obtain these, we impose restrictions on the $\xi_a(i_a)$, namely that setting ξ_a equal to zero means we must also set ξ_b to be zero for all $b \supseteq a$.

For example, suppose that $\Gamma = \{A, B, C\}$, and that $\xi_{AB} = 0$. In this case, we require $\xi_{ABC} = 0$ in a hierarchical model. It is not possible to write an explicit expression for the number of such models, but this number is much smaller than the total number of log-linear models. The *generators* of a hierarchical log-linear interaction model are the maximal sets a such that ξ_a is non-zero; a hierarchical model is determined uniquely by its generators.

The set of graphical models form a highly attractive subset of the hierarchical models, both for ease of analysis and their obvious interpretation in terms of conditional independence (an interpretation which is immediately obvious from the graph). Graphical models may be either directed or undirected. A graphical log-linear model may be represented by a graph, with a set of vertices \mathcal{V} corresponding to the variables, and a set \mathcal{E} of edges representing the independence structure. The notation (A, B) is used to represent the edge between variables A and B . The absence of an edge between two vertices A and B means that A is conditionally independent of B given all other variables. This is equivalently written as: if $(A, B) \notin \mathcal{E}$, then $A \perp\!\!\!\perp B \mid \mathcal{V} \setminus \{A, B\}$. Variables A and B are (marginally) independent if no path of edges exists between vertices A and B , in which case $A \perp\!\!\!\perp B$.

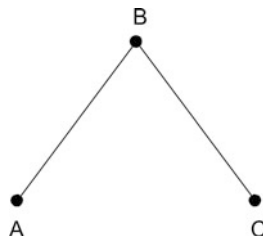
A subset C of Γ is called a *clique* if the subgraph containing only elements of C has an edge connecting each element (i.e. is complete), and the inclusion of another vertex from \mathcal{V} in C would result in at least one pair of unconnected vertices. A graph is *triangulated* if it contains no chordless cycles of length greater than three, and the subset D is said to *separate* subsets A and B if every path from any vertex in A to one in B must pass through a vertex in D . In such a case, variables in A are conditionally independent from those in B , given D .

A hierarchical model is *graphical* if its generators correspond to the cliques of its (undirected) conditional independence graph. These models form a subset of the log-linear models. We will assume throughout that all models include the intercept term ξ_{\emptyset} and all main effect terms (ξ_a where $|a| = 1$), since those without are of little interest. Then the $\binom{|\Gamma|}{2}$ possible edges in each graph gives the total number of possible graphical models as $2^{\binom{|\Gamma|}{2}}$.

Note that any hierarchical model can be represented by an (undirected) conditional independence graph, although such a graph does not necessarily represent a single hierarchical model. However, these models will not be excluded from our analyses, as they form a rich collection of models with many applications. An example of such a model is the model containing the three variables A , B , and C with interaction terms AB , AC , and BC though no 3-way interaction term ABC . In this case, the 2-way interactions are homogeneous with respect to the third variable; for example, the interaction between A and B does not depend on the value of C . Although this model is clearly not graphical, real data may be found to follow this pattern of association, so this model should not be excluded from our analyses.

An important subset of graphical models are *decomposable models*. These are defined as models whose joint cell probabilities may be directly expressed as a

Fig. 4.1 Independence graph
for the graphical
(hierarchical) model
 $\{(A, B), (B, C)\}$



function of the marginal probabilities of the cliques of the model. A model is decomposable if its graph is triangulated.

For example, consider the model represented in Fig. 4.1. This graph is clearly triangulated (with cliques $\{A, B\}$ and $\{B, C\}$), so the model is decomposable and the joint cell probabilities \mathbf{p} may be written as a product of the marginal and conditional probabilities \mathbf{p}^A , $\mathbf{p}^{B|A}$, $\mathbf{p}^{C|B}$. Equivalently, the cell probabilities may be directly expressed in terms of marginal probabilities of cliques and separators as $p(\mathbf{i}) = p(\mathbf{i}_{AB})p(\mathbf{i}_{BC})/p(\mathbf{i}_B)$.

Decomposable models admit the most straightforward analyses, but clearly exclude many potential (and useful) models, and there is often little justification to restrict attention to these models other than computational considerations. One key reason that decomposable models are analytically more straightforward is that, if the model is decomposable, then it can be represented as a *Directed Acyclic Graphical* (DAG) model.

A *directed graph* contains edges *from* one vertex *to* another, for example $A \rightarrow B$ denotes the presence on an edge from A to B , and we call A a parent of B and B a child of A . The edge from A to B will be written $\langle A, B \rangle$. The set of parents of B is denoted by $pa(B)$. A path of length $n \geq 0$ from A to B is a sequence $A = X_0, \dots, X_n = B$ of distinct vertices such that $\langle X_{i-1}, X_i \rangle \in \mathcal{E}$ for all $i = 1, \dots, n$. If there is a path from A to B we write $A \succ B$. The set of vertices A such that $A \succ B$ are the ancestors $an(B)$ of B and the descendants $de(A)$ of A are the vertices B such that $A \succ B$. The nondescendants of A are $nd(A) = V \setminus (de(A) \cup \{A\})$. A path which starts and ends at the same point is known as a *cycle*, and a directed graph is *acyclic* if it contains no cycles. A Directed Acyclic Graphical (DAG) model implies the conditional independences $A \perp\!\!\!\perp nd(A) | pa(A)$. A DAG admits a *perfect numbering* of the variables in the graph, by numbering the vertices (variables) such that edges are necessarily directed from vertices with lower numbers to those with higher numbers. For DAG models, the joint probability may be expressed as

$$p(\mathbf{i}) = \prod_{\gamma=1}^{|\Gamma|} P(\gamma = i_\gamma | pa(\gamma) = \mathbf{i}_{pa(\gamma)}) \quad (4.7)$$

where the probabilities are unconstrained (except for the requirement for probabilities to be non-negative and sum to one) and hence form a convenient parameterisation of the model.

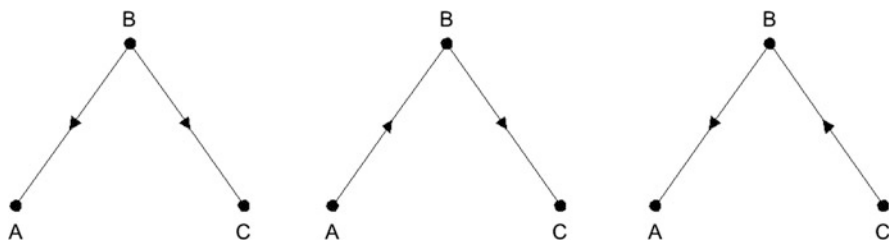


Fig. 4.2 Equivalent DAGs for the undirected graphical model in Fig. 4.1

As an example of directed graphical representations, consider the model represented by the undirected graph in Fig. 4.1. Several possible directed versions of this graph are possible and displayed in Fig. 4.2. Each of these graphs admit different perfect orderings of variables, and one admits two orderings. Working from left to right, the first admits orderings BCA and BAC , and the second and third admit orderings CBA and ABC respectively.

The use of graphs to represent the pattern of associations in statistical models dates back to Wright [40], but it was [5] who used graphs in contingency table analysis, defining the subset of the hierarchical log-linear models known as graphical models.

Early adopters of methods within a Bayesian framework were [36] and [6]. Madigan and York [29] presented a comprehensive discussion on Bayesian graphical models for a variety of discrete data problems. Graphical modelling was shown to allow prior information to be effectively incorporated into the analysis, and model uncertainty properly accounted for.

4.2 Bayesian Inference for Contingency Tables

Bayesian inference for Contingency Tables under the multinomial model (4.1) is obtained via the posterior distribution for \mathbf{p} given \mathbf{n}

$$f(\mathbf{p}|\mathbf{n}) = \frac{f(\mathbf{n}|\mathbf{p})f(\mathbf{p})}{\int f(\mathbf{n}|\mathbf{p})f(\mathbf{p})d\mathbf{p}} \quad (4.8)$$

where the *prior distribution* for \mathbf{p} , $f(\mathbf{p})$, represents the uncertainty about \mathbf{p} prior to observing the cell counts \mathbf{n} .

An important choice in the analysis of log-linear models is that of the prior distribution $f(\mathbf{p})$. For the saturated (unconstrained) multinomial model the Dirichlet

distribution is a natural choice of prior distribution for cell probabilities \mathbf{p} (which are positive and sum to one). Its density has the form

$$f(\mathbf{p}) = \frac{\Gamma(\boldsymbol{\alpha})}{\prod_i \Gamma(\alpha(i))} \prod_{i \in I} p(i)^{\alpha(i)-1} \quad (4.9)$$

where $\boldsymbol{\alpha}$ are parameters which control the location and dispersion of the distribution, and $\alpha = \sum_i \alpha(i)$.

Under multinomial sampling, the likelihood function for a saturated log-linear model is given by (4.1) and hence the Dirichlet distribution is a conjugate prior distribution for a saturated log-linear model, as it leads to a Dirichlet posterior distribution with density of the form

$$f(\mathbf{p}|\mathbf{n}) \propto \prod_{i \in I} p(i)^{n(i)+\alpha(i)-1}. \quad (4.10)$$

Conjugacy is convenient in Bayesian statistical analysis as it may (as in this case) result in tractable computation. Furthermore, prior specification may be facilitated if conjugacy is a result of prior and likelihood having a similar form. In such cases the ‘information content’ of the prior may be straightforward to specify. As may be seen from expression (4.10), the parameters $\boldsymbol{\alpha}$ may be considered as a ‘prior cell count’. Hence, for reference analyses, where the prior is intended to be only weakly informative, small common values of $\alpha(i)$ are appropriate. In one of the first Bayesian approaches to contingency table modelling, [27] considered the limiting case where $\alpha(i) = 0$, producing an improper prior density (which does not integrate to 1). The problem with this approach is that it will lead to an improper posterior density if any cell counts $n(i)$ are zero.

Setting $\alpha(i) = 1$ results in a uniform prior [26], a conventional choice for a weakly informative prior density. Two other common choices for $\alpha(i)$ are common: $\alpha(i) = \frac{1}{2}$ [21], which is Jeffreys’ prior for a multinomial \mathbf{p} ; and $\alpha(i) = 1/|I|$ [33], which has the appealing interpretation of a single prior observation distributed throughout the table.

For nonsaturated log-linear models, such as hierarchical, graphical, and decomposable log-linear models, \mathbf{p} is constrained by the form of the model and it is more convenient to work with an unconstrained parameterisation. Typically, these are the log-linear parameters $\boldsymbol{\beta}$ as in (4.6), or possibly, for a decomposable model, the conditional probabilities in the decomposition (4.7). For the former (and dropping the model subscript m for the present, for ease of exposition) Bayes theorem in (4.8) becomes

$$f(\boldsymbol{\beta}|\mathbf{n}) = \frac{f(\mathbf{n}|\boldsymbol{\beta})f(\boldsymbol{\beta})}{\int f(\mathbf{n}|\boldsymbol{\beta})f(\mathbf{p})d\boldsymbol{\beta}} \quad (4.11)$$

where $f(\boldsymbol{\beta})$ is the prior distribution for the unconstrained log-linear model parameters $\boldsymbol{\beta}$ and $f(\mathbf{n}|\boldsymbol{\beta})$ is the multinomial likelihood obtained through (4.1), (4.5), and (4.6)

$$\begin{aligned} f(\mathbf{n}|\boldsymbol{\beta}) &= \frac{n!}{\prod_{\mathbf{i} \in I} n(\mathbf{i})!} \exp \left[\sum_{j \in M} \left(\sum_{\mathbf{i} \in I} n(\mathbf{i})x(\mathbf{i}, j) \right) \beta_j \right. \\ &\quad \left. - n \log \left(\sum_{\mathbf{i} \in I} \exp \left(\sum_{j \in M} x(\mathbf{i}, j)\beta_j \right) \right) \right] \\ &= \exp \left[\sum_{j \in M} t_j \beta_j - n \log \left(\sum_{\mathbf{i} \in I} \exp \left(\sum_{j \in M} x(\mathbf{i}, j)\beta_j \right) \right) \right] \end{aligned} \quad (4.12)$$

where $t_j = \sum_{\mathbf{i} \in I} \alpha(\mathbf{i})x(\mathbf{i}, j)$, $j = 1, \dots, d_m$ and $x(\mathbf{i}, j)$ are the elements of the model matrix \mathbf{X} in (4.6).

4.2.1 Distributions Based on the Normal Distribution

As defined in Sect. 4.1.2, the log-linear model parameters are unconstrained and allowed to take any real value, so that $\boldsymbol{\beta} \in \mathbb{R}^d$. A natural prior distribution for these parameters may therefore be multivariate normal, i.e. $\boldsymbol{\beta} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, for suitable mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$. The use of such a distribution was first investigated by Good [16], motivated by the desire to obtain smoothed estimates for cell probabilities with small observed frequencies, an idea further developed by Leonard [25] and Laird [24]. The purpose of [23] was to use a Normal distribution to effectively encapsulate prior information into the analysis of contingency tables. Their approach used a multivariate Normal prior for all parameters together, and as such allowed separate specification of prior information for each log-linear model main effect or interaction term. However, they found the use of such a prior resulted in a generally intractable posterior distribution, and so developed a measure of posterior dispersion based on the curvature of the log of the posterior density at its mode.

The prior on $\boldsymbol{\beta}$ induces a prior distribution on $\log \mathbf{p}$. Forster [13] showed that a multivariate normal prior for $\log \boldsymbol{\mu}$, the log-cell-means in a Poisson log-linear model must have a certain form in order for it to be invariant to permutations of the set of levels I_γ of each factor (a reasonable requirement for a reference prior). He also derived conditions on the prior to make Poisson and multinomial analysis equivalent using such a prior. The distribution takes the form

$$\log \boldsymbol{\mu} \sim N(\delta \mathbf{1}, \sum_{a \subseteq \Gamma} \alpha_a^2 \mathbf{T}_a)$$

where the T_a are projection matrices given by

$$T_a = \bigotimes_{\gamma \in \Gamma} \left\{ 1(\gamma \in a) \left(\mathbf{I}_{|I_\gamma|} - \frac{1}{|I_\gamma|} \mathbf{J}_{|I_\gamma|} \right) + 1(\gamma \notin a) \frac{1}{|I_\gamma|} \mathbf{J}_{|I_\gamma|} \right\} \quad (4.13)$$

and \mathbf{I}_d is a $d \times d$ identity matrix and \mathbf{J}_d a $d \times d$ matrix of 1s. The prior distributions for the model parameters β_a , with the exception of β_\emptyset (corresponding to the intercept term), are then given by

$$\beta_a \stackrel{ind}{\sim} N(\mathbf{0}, \alpha_a^2 \Sigma_a) \quad a \subseteq \Gamma$$

where

$$\Sigma_a = \frac{1}{|I|} \prod_{\gamma \in a} |I_\gamma| \bigotimes_{\gamma \in a} \left(\mathbf{I}_{(|I_\gamma|-1)} - \frac{1}{|I_\gamma|} \mathbf{J}_{(|I_\gamma|-1)} \right) \quad a \subseteq \Gamma.$$

The prior for β_\emptyset is

$$\beta_\emptyset \sim N(\tau, \alpha_\emptyset^2)$$

for a specified value of τ . It is necessary to assume independence of the model parameters, though this is not restrictive as it seems sensible to do so if we are to perform a reference analysis.

Posterior inference using Normal priors, based on Markov chain Monte Carlo sampling, is possible following results by Dellaportas and Smith [9]. They present a method for sampling from a wide range of generalised linear models using Gibbs sampling. Their Gibbs sampler is based on the adaptive rejection sampling method proposed by Gilks and Wild [15], which is a technique for sampling from any log-concave univariate probability density function.

4.2.2 Distributions Based on the Dirichlet Distribution

Although the Normal distribution is a natural choice for the log-linear model parameters β it lacks the ease of interpretability of the conjugate Dirichlet prior for the saturated model. Prior distributions based on the Dirichlet and suitable for the analysis of log-linear model analysis may instead be considered.

4.2.2.1 Conditional Dirichlet Distribution

The Dirichlet prior for the saturated model is given in (4.9) and, by straightforward transformation of variables, can be expressed as a density for any multivariate logit (4.4). To do this, we note that θ is overspecified (as is p) since $\mathbf{a}^T \theta = 0$ and so when we write θ we are effectively making a selection, $\theta_{\setminus i_0}$, of any $|I| - 1$ components of θ (provided $a(i_0) > 0$) in which case the Jacobian $|J|$ for the transformation from $p_{\setminus i_0}$ to $\theta_{\setminus i_0}$ is therefore given by

$$\left| \frac{\partial \theta}{\partial p} \right|^{-1} = \frac{1}{a(i_0)} \prod_j p(j).$$

Then the Dirichlet distribution for θ is

$$\begin{aligned} f(\theta) &= \frac{\Gamma(\alpha)}{\prod_i \alpha(i)} \prod_{i \in I} p(\theta(i))^{\alpha(i)-1} \left[\frac{1}{a(i_0)} \prod_{i \in I} p(\theta(i)) \right] \\ &= \frac{1}{a(i_0)} \frac{\Gamma(\alpha)}{\prod_i \alpha(i)} \prod_{i \in I} \frac{\exp(\theta(i)\alpha(i))}{(\sum \exp \theta(i))^{\alpha(i)}} \\ &= \frac{1}{a(i_0)} \frac{\Gamma(\alpha)}{\prod_i \alpha(i)} \exp \left[\sum_{i \in I} \theta(i)\alpha(i) - \alpha \log \left(\sum_{i \in I} \exp \theta(i) \right) \right] \end{aligned}$$

where $\alpha = \sum_i \alpha(i)$.

The saturated log-linear interaction model sets $\theta = \mathbf{X}\beta$, for a suitable full rank model matrix \mathbf{X} and full set of interaction parameters $\beta = \{\beta_a, a \subseteq I\}$. Therefore, a Dirichlet distribution for β in the saturated model is obtained by a simple linear transformation as

$$\begin{aligned} f(\beta) &= \frac{C}{a(i_0)} \frac{\Gamma(\alpha)}{\prod_i \alpha(i)} \exp \left[\sum_j \left(\sum_{i \in I} \alpha(i)x(i, j) \right) \beta_j \right. \\ &\quad \left. - \alpha \log \left(\sum_{i \in I} \exp \left(\sum_j x(i, j)\beta_j \right) \right) \right] \end{aligned} \quad (4.14)$$

where $x(i, j)$ are the elements of the saturated model matrix \mathbf{X} and C is a constant which will depend on the parametrisation adopted.

The *Conditional Dirichlet* distribution for a particular non-saturated log-linear interaction model as that distribution obtained from expression (4.14) by conditioning on certain β_j terms to be zero. More precisely, we partition β into those terms in the model β_m and those not in the model $\beta_{\bar{m}}$, and condition on $\beta_{\bar{m}} = \mathbf{0}$. We also partition \mathbf{X} into \mathbf{X}_m containing the columns corresponding to the parameters in β_m ,

and $\mathbf{X}_{\bar{m}}$ containing the others. Then the conditional Dirichlet distribution for $\boldsymbol{\beta}$ is obtained from expression (4.14) by summing over β_j for $j \in m$ only:

$$\begin{aligned} f(\boldsymbol{\beta}_m) &\propto \exp \left[\sum_{j \in M} \left(\sum_{\mathbf{i} \in I} \alpha(\mathbf{i}) x(\mathbf{i}, j) \right) \beta_j - \alpha \log \left(\sum_{\mathbf{i} \in I} \exp \left(\sum_{j \in M} x(\mathbf{i}, j) \beta_j \right) \right) \right] \\ &= \exp \left[\sum_{j \in M} s_j \beta_j - \alpha \log \left(\sum_{\mathbf{i} \in I} \exp \left(\sum_{j \in M} x(\mathbf{i}, j) \beta_j \right) \right) \right] \end{aligned} \quad (4.15)$$

where $s_j = \sum_{\mathbf{i} \in I} \alpha(\mathbf{i}) x(\mathbf{i}, j)$, $j = 1, \dots, d_m$.

Care must be taken with any conditioning such as this, as the prior distribution obtained through conditioning on a set of complex constraints is not invariant under general reparameterisation of those constraints. This is known as the Borel paradox. However, the prior distributions induced under various parameterisations may be shown to be related by the Borel-Kolmogorov dependence formula [19]. Furthermore, we argue that deriving a prior for a log-linear model by conditioning on a scale on which the constraints that determine the model are linear, is natural.

The correspondence of (4.12) and (4.15) makes this the Diaconis-Ylvisaker conjugate prior. Massam et al. [30] investigate this structure in detail for hierarchical log-linear models. It is clear that the conditional Dirichlet distribution satisfies the conditions stated in [30] for the prior to be proper, with a proper initial full Dirichlet prior leading to a proper prior on each submodel; see also [19]. The conditional Dirichlet conjugate priors are also compatible (in a conditional sense; see [7]).

4.2.3 Hyper-Dirichlet Distribution

A sub-class of models which admit straightforward analyses are decomposable log-linear models. In these models, we may parameterise directly in terms of the unconstrained conditional probabilities appearing in the decomposition (4.7). The *hyper-Dirichlet* distribution was proposed by Dawid and Lauritzen [6] as a conjugate prior distribution for the parameters of a decomposable log-linear model. Under a hyper-Dirichlet prior, each clique has a Dirichlet marginal distribution, and the realisations on overlapping portions of cliques must be consistent regardless of the clique from which they are derived.

For a DAG representation of a decomposable model, we know that a cell probability may be expressed using the decomposition (4.7). With this parameterisation, a natural (closed under sampling) prior family is

$$P(\gamma | pa(\gamma) = \mathbf{i}_{pa(\gamma)}) \sim \text{Dirichlet}(\boldsymbol{\alpha}_\gamma(\mathbf{i}_{pa(\gamma)})) \quad (4.16)$$

independently, for each variable $\gamma \in \Gamma$ and each parent combination $\mathbf{i}_{pa(\gamma)}$. This is the product Dirichlet described by Cowell et al. [4]. For this distribution to be a hyper-Dirichlet distribution the parameters $\{\boldsymbol{\alpha}_\gamma(\mathbf{i}_{pa(\gamma)})\}$ must be chosen so that any clique marginal prior is also Dirichlet distributed. A straightforward way of ensuring this is by deriving the prior distributions on the conditional probabilities (and hence also the cliques) as the marginal distributions from a Dirichlet distribution on the full set of probabilities.

Substituting (4.7) into (4.1) it is immediately obvious that, *a posteriori*,

$$P(\gamma | pa(\gamma) = \mathbf{i}_{pa(\gamma)}) \sim \text{Dirichlet}(\mathbf{n}_\gamma(\mathbf{i}_{pa(\gamma)}) + \boldsymbol{\alpha}_\gamma(\mathbf{i}_{pa(\gamma)}))$$

independently, where $\mathbf{n}_\gamma(\mathbf{i}_{pa(\gamma)}) = \{n(i_\gamma, \mathbf{i}_{pa(\gamma)}), i_\gamma \in I_\gamma\}$ so the posterior has the same form as the prior (and is hyper-Dirichlet if the prior is).

Marginal inference from a hyper-Dirichlet distribution is straightforward. Using the directed representation, we can write down the hyper-Dirichlet distribution as a product of independent Dirichlet distributions. Monte Carlo samples may then be obtained from each of these distributions in turn by sampling from independent gamma distributions and applying the result that if z_i are independent samples from Gamma (a_i, b) distributions, then $\mathbf{z} / \sum z_i$ is a sample from a Dirichlet(\mathbf{a}) distribution. Further convenient properties for computation are introduced later.

4.2.4 Relationship Between Conditional Dirichlet and Hyper-Dirichlet Distributions

Given that they are both closed under multinomial sampling and directly derived from a full Dirichlet prior for the saturated model, it is perhaps not surprising that the conditional Dirichlet distribution, for any log-linear model which is a decomposable graphical model, can be shown to be hyper-Dirichlet. This is proved by Massam et al. [30].

The conditional Dirichlet distribution is an attractive prior distribution as its parameters may be interpreted as prior data, and inference using this prior is straightforward for decomposable models by considering the equivalent hyper Dirichlet distribution, which is tractable. The relationship of the conditional Dirichlet distribution to the hyper-Dirichlet distribution allows it to be considered as a natural extension to non-decomposable models.

4.3 Posterior Inference

Except in the case of decomposable models, for which tractable posterior computation was described in Sect. 4.2.3, the conditional Dirichlet distribution is not

generally tractable. It does, however, have some properties which make it amenable to computational methods which allow accurate approximate posterior inference.

A straightforward Normal approximation can be obtained by computing the posterior mode and the curvature (negative second derivative of log-posterior) at the mode. Because the conditional Dirichlet density is ‘likelihood-like’ then these summaries can be straightforwardly obtained by any statistical software for fitting log-linear models (provided that fractional cell counts are not prohibited) simply by inputting the data as $\mathbf{n} + \boldsymbol{\alpha}$ rather than \mathbf{n} . This approximation can be reasonably accurate, but can be unreliable in examples where there are small cell counts. In such examples we typically resort to Markov chain Monte Carlo computation where we generate a (dependent) sample from (approximately) the posterior distribution, from which it is then straightforward to compute a sample from the marginal posterior distribution of any function of interest. Summaries in the form of posterior expectations are approximated by sample means.

The posterior distribution for $\boldsymbol{\beta}$ in a conditional Dirichlet distribution is amenable to computation by a Gibbs sampler. In particular, its posterior density is (globally) log-concave in the case where all of the components of \mathbf{s} are positive. This makes sampling from univariate conditional posterior distributions accessible using, for example, adaptive rejection sampling [15], which is available in standard Gibbs sampling software such as OpenBUGS and JAGS. Log-concavity can be established by noting that the Hessian of prior (or similarly posterior) density $f(\boldsymbol{\beta})$ can be written as

$$H = -(\alpha + n)X^T(\text{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^T)X$$

where \mathbf{p} is a function of $\boldsymbol{\beta}$ through (4.4) and (4.6). This Hessian is clearly negative definite as, for any $\mathbf{y} \neq \mathbf{0}$, $(X\mathbf{y})^T(\text{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^T)(X\mathbf{y})$ is the variance of a discrete distribution with sample space $\{[X\mathbf{y}]_i\}$ and probabilities $\{p_i\}$ and where the $[X\mathbf{y}]_i$ cannot all be equal (other than at zero, which only arises when $\mathbf{y} = \mathbf{0}$) because then $\mathbf{a}^T X = 0$ would imply that $\sum a(\ell) = 0$ rather than 1. Log-concavity of each univariate density is a direct consequence. An alternative Gibbs sampling approach for conditional Dirichlet distributions, based on Bayesian iterative proportional fitting, is presented by Dobra and Massam [11].

4.3.1 Example 1

Consider a 2^3 contingency table and the log-linear interaction model which may be represented graphically in Fig. 4.1. This model has cliques $\{A, B\}$ and $\{B, C\}$, and

may be parameterised as $P(B)P(A|B)P(C|B)$. A model matrix for this model (for the centred multivariate logit) is given by

$$\mathbf{X} = \frac{1}{\sqrt{8}} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ -1 & 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 & -1 \\ 1 & 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 & 1 \\ -1 & -1 & -1 & 1 & 1 \end{pmatrix}.$$

The prior chosen for this example is the diffuse prior with parameters $\alpha(i) = \frac{1}{8}$ for all i (Perks' prior). The hyper-Dirichlet distribution may be constructed from the 'full' Dirichlet distribution as follows. The distribution for $P(B)$ is obtained first, then the conditional distributions $P(A|B = 1)$, $P(A|B = 2)$, $P(C|B = 1)$, and $P(C|B = 2)$ are obtained in such a way that they are consistent with the distribution for $P(B)$. In this example, $P(B)$ is distributed as a $Beta(\frac{1}{2}, \frac{1}{2})$, and all the conditional densities follow $Beta(\frac{1}{4}, \frac{1}{4})$ distributions. As is clear from the model parameterisation, there are five independent components in this decomposition.

The Gibbs sampler was used to generate samples from this prior, and Fig. 4.3 shows kernel density estimates for the five independent distributions produced by the sampler, overlaid with the true Dirichlet density. All densities are on the logistic scale.

As can be seen in Fig. 4.3, there is excellent agreement between the kernel density estimates from the Gibbs samples and the true densities. This is to be expected, and validates the use of the Gibbs sampler in this example; the computation time for such a sample is negligible (a few seconds). Although sampling from the prior is not usually a requirement in Bayesian analysis, it will be a requirement of accurate posterior computation in Sect. 4.4.

4.3.2 Example 2

The Gibbs sampler was used to produce a posterior sample for some data concerning incidence of coronary heart disease. The data was presented by Edwards and Havranek [12], and analysed further by Madigan and Raftery [28] and Dellaportas and Forster [8].

The data (presented in Table 4.1) concerns 1841 men, who have been cross-classified in a 2^6 table by six factors for coronary heart disease. The six factors are: A —Smoking (no or yes); B —Strenuous mental work (no or yes); C —Strenuous physical work (no or yes); D —Systolic Blood pressure (<140 or ≥ 140); E —Ratio

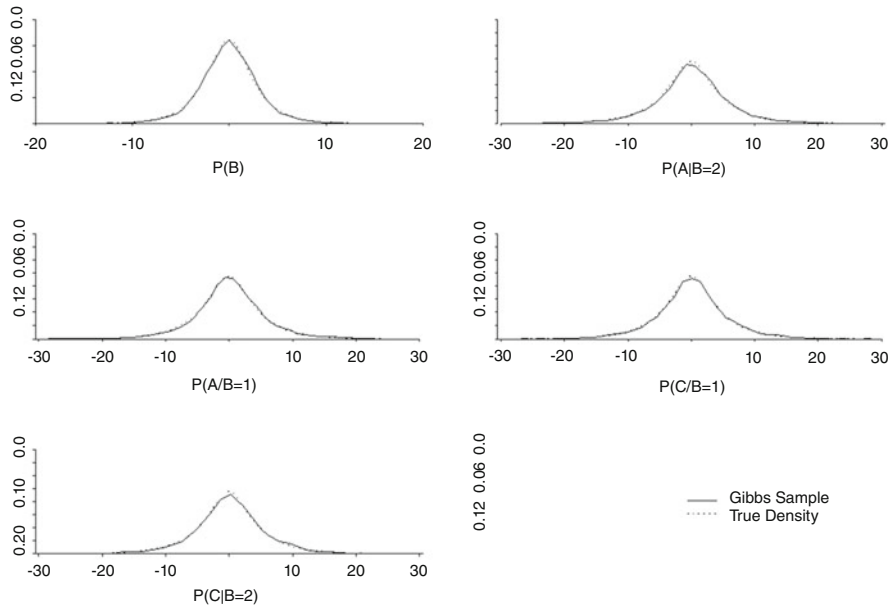


Fig. 4.3 Plots showing kernel density estimates from Gibbs samples overlaid with the true density functions

of α and β lipoproteins (<3 or ≥ 3); F —Family anamnesis of coronary heart disease (negative or positive).

Posterior samples were obtained for this data using the Gibbs sampler, for the most probable (hierarchical) models identified by Dellaportas and Forster [8]. These models have posterior probabilities of >0.05 . The prior parameters were set to $\alpha_i = 1/|I| = 0.015625$ for a diffuse prior. Figures 4.4, 4.5, 4.6, and 4.7 show the posterior distributions of the 2-way interaction parameters—each figure corresponds to a particular model, within which the posterior for each interaction parameter is presented.

4.3.3 Convergence of Gibbs Sampler

Repeated use of the Gibbs sampler leads to the conclusion that samples produced are not highly dependent, as the sampler appears to mix well. For the samples in Example 1, the autocorrelations at lag 1 are 0.2, and drop below 0.05 after lag 4. Fig. 4.8 presents time series plots for the data in Example 1. For the sake of clarity, only the first 2000 observations are plotted in each case. The time series show that the Gibbs sampler is mixing very well, and so the observations are not highly dependent. Scatterplots for each pair of variables are shown in Fig. 4.9. There is clearly no distinct correlation between parameters.

Table 4.1 Risk factors for coronary heart disease

<i>F</i>	<i>E</i>	<i>D</i>	<i>C</i>	<i>B</i>		Yes				
				<i>A</i>	No	Yes	No	Yes		
Negative	<3	<140	No		44	40	112	67		
			Yes		129	145	12	23		
		≥140	No		35	12	80	33		
			Yes		109	67	7	9		
		≥3	<140	No		23	32	70	66	
				Yes		50	80	7	13	
	≥140		No		24	25	73	57		
			Yes		51	63	7	16		
	Positive		<3	<140	No		5	7	21	9
					Yes		9	17	1	4
		≥140		No		4	3	11	8	
				Yes		14	17	5	2	
≥3		<140	No		7	3	14	14		
			Yes		9	16	2	3		
≥140	No		4	0	13	11				
	Yes		5	14	4	4				

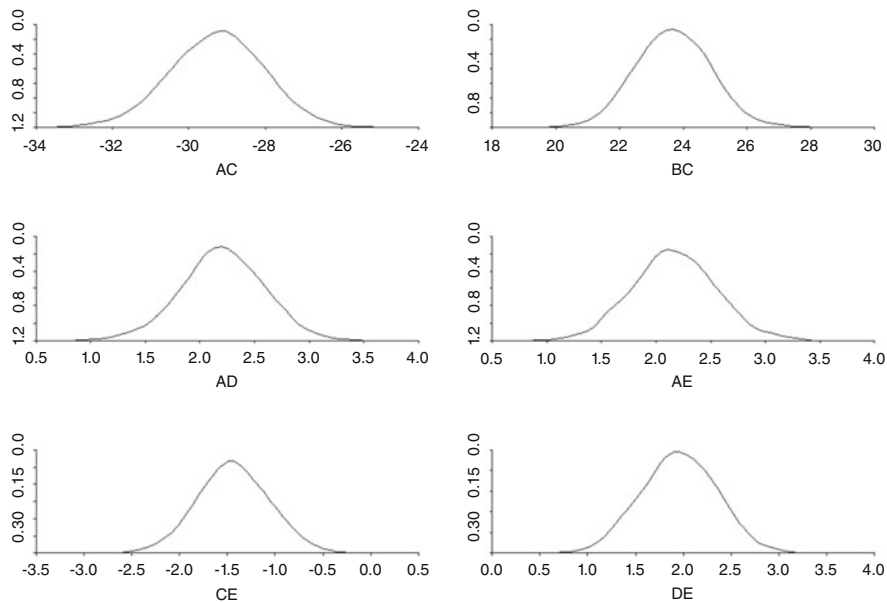


Fig. 4.4 Model $AC + BC + AD + AE + CE + DE + F$ (posterior probability 0.28)

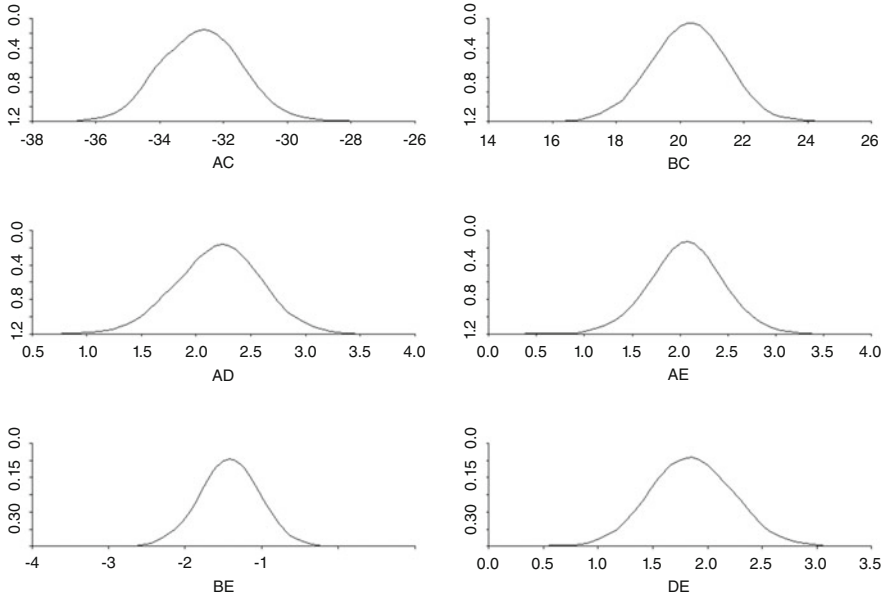


Fig. 4.5 Model $AC + BC + AD + AE + BE + DE + F$ (posterior probability 0.16)

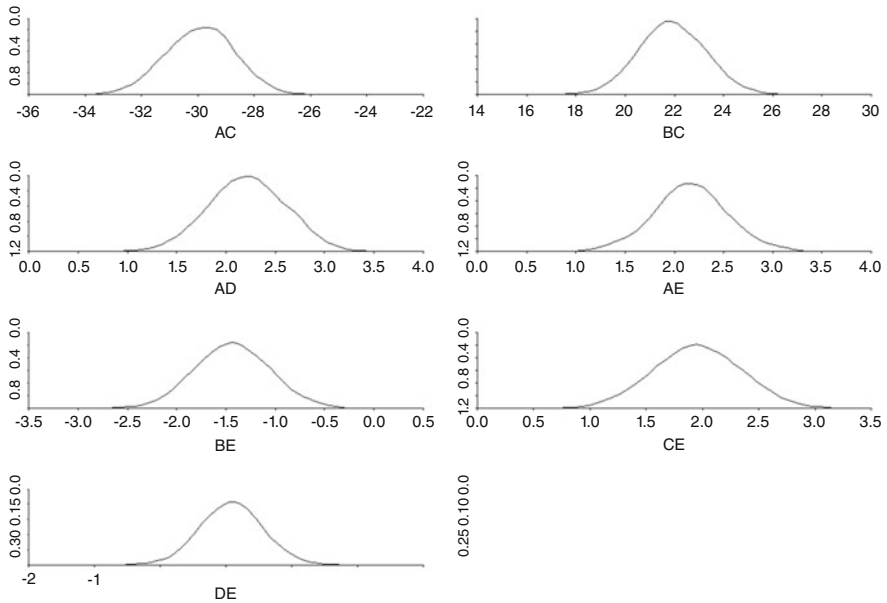


Fig. 4.6 Model $AC + BC + AD + AE + BE + CE + DE + F$ (posterior probability 0.07)

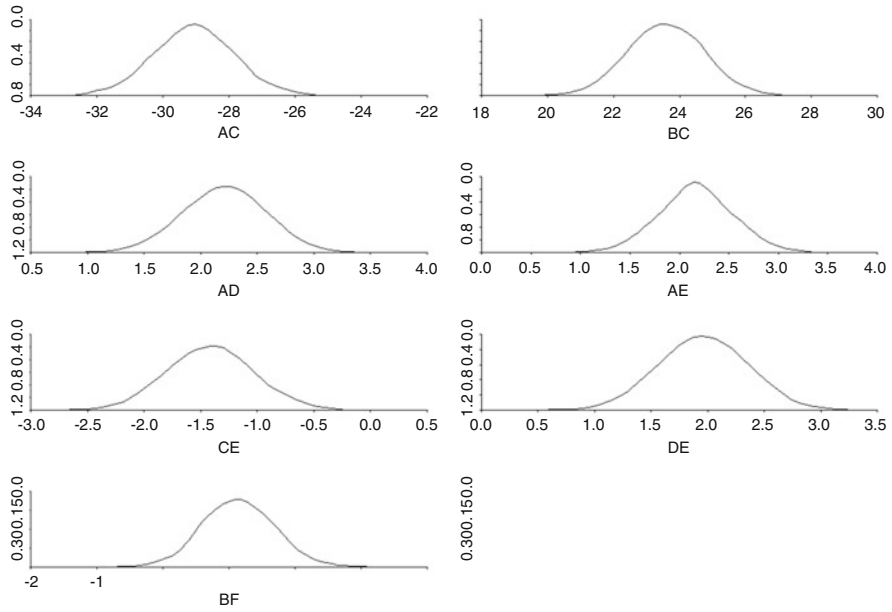


Fig. 4.7 Model $AC + BC + AD + AE + CE + DE + BF$ (posterior probability 0.07)

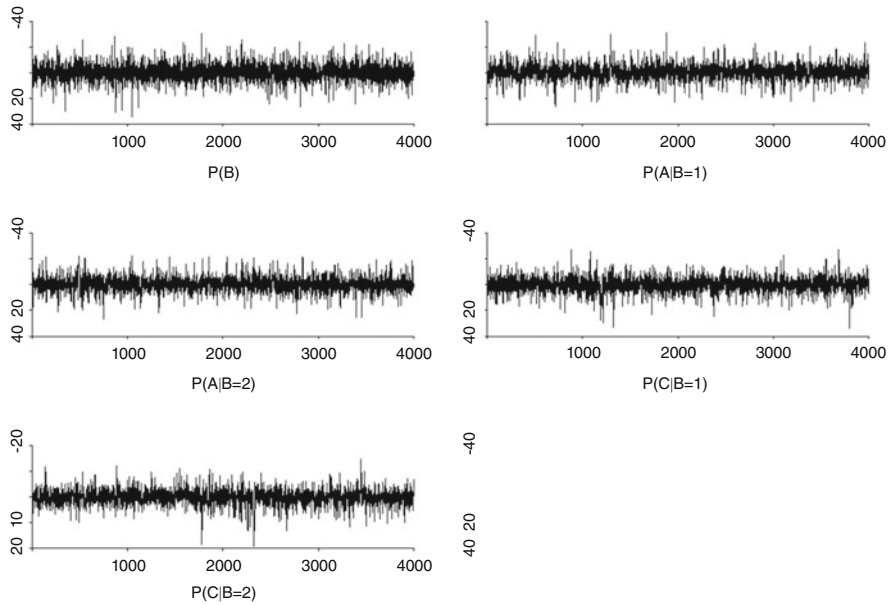


Fig. 4.8 Time series plots for Gibbs samples in Example 1

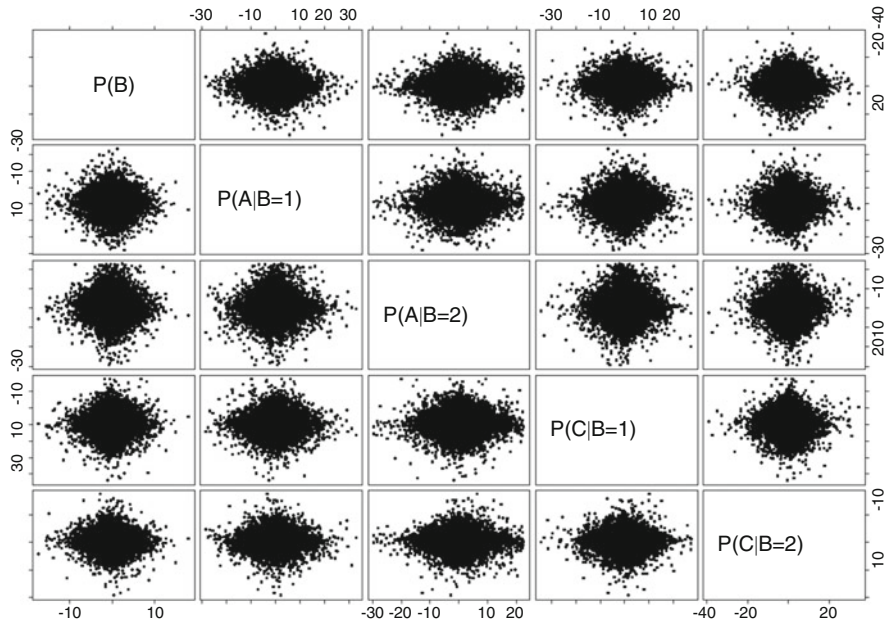


Fig. 4.9 Pairwise scatterplots for Gibbs samples in Example 1

Time series plots for the data presented in Example 2 are all similar. Figure 4.10 presents these for each interaction parameter for the most probable model, $AC + BC + AD + AE + CE + DE + F$, though again only the first 2000 observations are plotted. Again, these time series show the observations are not highly dependent, and that the sampler is mixing well.

4.4 Model Determination and Model Averaging

Until now, we have focussed on Bayesian inference for contingency table data using a single log-linear interaction model chosen *a priori*. Using a parsimonious non-saturated model in this way can bring considerable benefits for inference, providing smooth estimates, particularly when data are sparse, and potentially more reliable predictions. However, it is rarely the case that we will have prior knowledge of which model will be the most appropriate to use. This is an additional component of our prior uncertainty. One possible solution is to use a saturated model but use the prior on certain log-linear parameters to ‘smooth towards’ a more parsimonious structure. This is the spirit of several of the previous approaches mentioned in Sect. 4.2.1, particularly for a two-way contingency table where the target model is one of marginal independence. For tables of higher dimension, [1] extended this type of approach to the analysis of general multiway tables using mixtures of multivariate

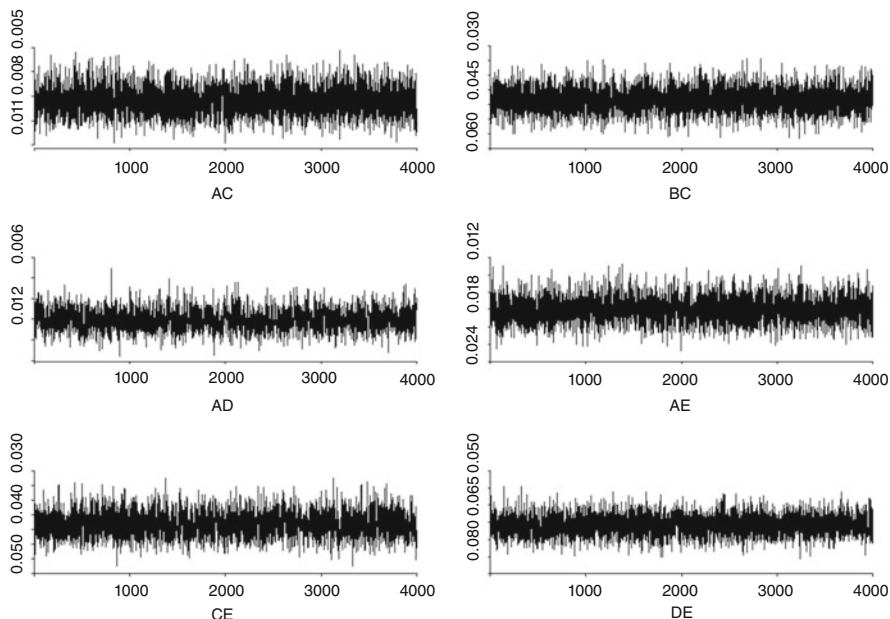


Fig. 4.10 Time series plots for Gibbs samples corresponding to model $AC + BC + AD + AE + CE + DE + F$ in Example 2

Normal distributions to model prior opinion, in the spirit of [25]. His method partitioned the saturated model parameter β into subsets $\beta = (\eta, \beta_1, \dots, \beta_s)$, where the elements of η are non-zero, but the elements of β_1, \dots, β_s may be zero. A Normal distribution was then assigned to β with mean $\mathbf{0}$ and variance Σ , where Σ^{-1} has a block-diagonal structure of multiples of identity matrices, with zeros corresponding to η and a single dispersion parameter P_i for each β_i . Such prior distributions model prior beliefs for each of the 2^s possible models. Hypotheses setting $\beta_i = 0$ correspond to letting P_i tend to infinity, whereas hypotheses for non-zero β_i values require a choice to be made for P_i . This choice is not arbitrary, as different values will have a pronounced effect on the Bayes factor. Albert's proposal was to place a prior on P_i , motivated by the approach of Good [17]. Following applications to examples involving two- and three-way contingency tables, he suggested that P_i should have a gamma($v_i/2, b_i^2 v_i/2$) distribution, where the choice of b_i depends on prior information, and v_i may vary, though his advocated choice $v_i = 1$ corresponds to a set of Cauchy distributions.

An alternative approach, which has gained in popularity since the 1990s, is to explicitly incorporate model uncertainty into the analysis, making the prior and posterior distributions discrete mixtures over competing models with prior model probabilities updated to posterior model probabilities favouring those models which are best supported by the observed data. Posterior inference and prediction naturally smooths over models, weighted by these posterior probabilities. For contingency

table modelling, this approach was adopted by Madigan and York [29] where the class of models under consideration was the decomposable graphical models for the table under consideration. This is the approach we shall consider throughout the rest of this chapter, with a focus on log-linear interaction models and the conditional Dirichlet prior.

4.4.1 Bayesian Inference Under Model Uncertainty

Suppose we have a set of models, M , one of which we believe provides a reasonable model for our data \mathbf{n} , the cell counts in a contingency table. For the full set of log-linear interaction models $m \subseteq \mathcal{P}(\Gamma)$ and hence $M = \mathcal{P}(\mathcal{P}(\Gamma))$. Each model m specifies a distribution for \mathbf{n} , $f(\mathbf{n}|m, \boldsymbol{\beta}_m)$, with the $\boldsymbol{\beta}_m$ an unknown vector of parameters for model m . We use Bayes' theorem to obtain the joint posterior distribution of m and $\boldsymbol{\beta}_m$

$$\begin{aligned} f(m, \boldsymbol{\beta}_m|\mathbf{n}) &\propto f(\mathbf{n}|m, \boldsymbol{\beta}_m)f(m, \boldsymbol{\beta}_m) \\ &\propto f(\mathbf{n}|m, \boldsymbol{\beta}_m)f(\boldsymbol{\beta}_m|m)f(m). \end{aligned}$$

Hence the posterior probability of model m may be found explicitly from

$$f(m|\mathbf{n}) = \frac{f(m) \int f(\mathbf{n}|m, \boldsymbol{\beta}_m)f(\boldsymbol{\beta}_m|m)d\boldsymbol{\beta}_m}{\sum_{m \in M} f(m) \int f(\mathbf{n}|m, \boldsymbol{\beta}_m)f(\boldsymbol{\beta}_m|m)d\boldsymbol{\beta}_m} \quad m \in M \quad (4.17)$$

where the integral in the numerator

$$\int f(\mathbf{n}|m, \boldsymbol{\beta}_m)f(\boldsymbol{\beta}_m|m)d\boldsymbol{\beta}_m \equiv f(\mathbf{n}|m) \quad (4.18)$$

is the *marginal likelihood*, sometimes called the evidence and interpreted as the prior predictive probability of observing the data computed before any data were observed.

If we have two competing models, m_1 and m_2 , the problem reduces to the calculation of the well-known *Bayes Factor*, which is the ratio of the posterior odds to the prior odds, and we have

$$\frac{f(m_1|\mathbf{n})}{f(m_2|\mathbf{n})} = \frac{f(m_1) \int f(\mathbf{n}|m_1, \boldsymbol{\beta}_{m_1})f(\boldsymbol{\beta}_{m_1}|m_1)d\boldsymbol{\beta}_{m_1}}{f(m_2) \int f(\mathbf{n}|m_2, \boldsymbol{\beta}_{m_2})f(\boldsymbol{\beta}_{m_2}|m_2)d\boldsymbol{\beta}_{m_2}}$$

where the second term on the right-hand side is the Bayes factor for model m_1 against model m_2 . Denoting the Bayes factor for comparing models m_1 and m_2 by B_{12} we have

$$B_{12} = \frac{\int f(\mathbf{n}|m_1, \boldsymbol{\beta}_{m_1})f(\boldsymbol{\beta}_{m_1}|m_1)d\boldsymbol{\beta}_{m_1}}{\int f(\mathbf{n}|m_2, \boldsymbol{\beta}_{m_2})f(\boldsymbol{\beta}_{m_2}|m_2)d\boldsymbol{\beta}_{m_2}}. \quad (4.19)$$

This notation can be extended to the case where we have multiple plausible models, by writing B_{jk} as the Bayes factor for model m_j against model m_k . The Bayes factor is the Bayesian analogue of the classical likelihood ratio, obtained by integration instead of maximisation.

Identifying those models which are best supported by the data is of interest in its own right, as it allows inferences about associations, independence, and conditional independence amongst the classifying variables. But it is also a very useful approach for obtaining smoothed estimates of cell probabilities and other functions of interest, particularly in the presence of sparse data. Model uncertainty, as encapsulated by the posterior model probabilities, is naturally propagated into inferences and predictions for any quantities which have a ‘model-independent’ interpretation (broadly speaking, anything which can be expressed as a predictand). Most significantly, for contingency table modelling, with log-linear model uncertainty, this applies to the cell probabilities. Suppose our quantity of interest is ϕ , which can be expressed as a function of model parameters $\boldsymbol{\beta}_m$ under every model, then we may obtain the posterior distribution for ϕ using the expression

$$f(\phi|\mathbf{n}) = \sum_{m \in M} f(\phi|m, \mathbf{n})f(m|\mathbf{n}) \quad (4.20)$$

where $f(m|\mathbf{n})$ is obtained from (4.17).

There is an important distinction between expression (4.17) for posterior model probabilities and expression (4.11) for the posterior distributions of the model parameters in the role played by prior normalising constants. In (4.11) any such constants can be factorised out of the numerator and denominator and cancel. Hence, for example, for conditional Dirichlet distributions, it is of no consequence that the prior density (4.15) is known only up to the constant of normalisation for general log-linear interaction models. On the other hand, no such factorisation is possible for (4.17) and therefore a means of computing the prior normalising constant is required if the conditional Dirichlet prior is to be used in this context. This has led some authors, as described in Sect. 4.2.1, to prefer the use of Normal priors for log-linear parameters.

This distinction between expression (4.17) and (4.11) also plays a role in the use of arbitrary diffuse prior distributions, which are commonly used to represent prior uncertainty. While they can be applied without consequence within a single model, extreme normalising constants cancelling in (4.11), the same is not true for inference under model uncertainty. Several solutions to this have been proposed in the literature, including by Spiegelhalter and Smith [37], Berger and Pericchi

[3], and O’Hagan [32]. In examples where inference is required under a vague or default prior specification, we prefer to use a diffuse, but properly defined (non-limiting), prior. We argue that a conjugate (conditional Dirichlet) prior is particularly helpful here, as it allows the user to consider the equivalent data content of the prior. Dellaportas and Forster, using (proper) normal priors, justified their choice of prior variance by drawing comparisons with the information provided by an equivalent prior sample. The conditional Dirichlet allows such arguments to be made more directly.

4.4.2 Computation Under Model Uncertainty

We require to compute the marginal likelihood $f(\mathbf{n}|m)$ defined in (4.18) for each $m \in M$, which presents two major issues. First, the integral in the marginal likelihood is often analytically intractable. For log-linear interaction models a straightforward expression for the marginal likelihood is only generally available for decomposable models with hyper-Dirichlet (or other similar product Dirichlet) prior distributions. For the hyper-Dirichlet which is equivalent to conditional Dirichlet, and derived from a Dirichlet(α) for the saturated model, then the resulting marginal likelihood is

$$P(\mathbf{n}|m) = \prod_{\gamma} \prod_{i_{\text{pa}(\gamma)}} \frac{\Gamma[\alpha(i_{\text{pa}(\gamma)})] \prod_{i_{\gamma}} \Gamma[\alpha(i_{\gamma}, i_{\text{pa}(\gamma)}) + n(i_{\gamma}, i_{\text{pa}(\gamma)})]}{\Gamma[\alpha(i_{\text{pa}(\gamma)}) + n(i_{\text{pa}(\gamma)})] \prod_{i_{\gamma}} \Gamma[\alpha(i_{\gamma}, i_{\text{pa}(\gamma)})]}. \quad (4.21)$$

The second issue which arises is the size of the set, M , of possible models. Even for relatively modest numbers of factors, this grows rapidly and computation of the marginal likelihood for all $m \in M$ becomes infeasible. A solution proposed by Madigan and Raftery [28], which they called Occam’s window, was to eliminate many of the models from (4.20). Their approach first eliminates any model with probability much smaller than the most probable model, then any model with probability lower than a model nested within it. They proposed search strategies for identifying a set of potentially acceptable models. Alternative search strategies have been proposed by Forster and Webb [14] for decomposable models and [11] for log-linear models with conjugate conditional Dirichlet priors.

An alternative approach to dealing with large numbers of competing models is to use Markov chain Monte Carlo (MCMC) methods. This effectively allows all possible models to be considered, A Markov chain is constructed so as to obtain a sample from $f(m, \theta_m | \mathbf{n})$, and the posterior model probabilities $f(m | \mathbf{n})$ are then estimated from this using the Monte Carlo sample proportions. The ‘reversible jump’ method of sampling was introduced by Green [18] and based on the Metropolis-Hastings method. Green presented a general description of the method, together with a particular implementation which may be adopted for log-linear models. This method was adapted by Dellaportas and Forster [8] and applied

to several classes of log-linear models, using normal priors. MCMC under model uncertainty is less attractive for analysis with conditional Dirichlet conjugate priors as it would remain limited by the size of the model space, M . This is because the prior normalising constants would be required to calculate the acceptance probability of proposed transitions in the chain, and these would still need to be calculated, either in advance or as each model was proposed for the first time.

Again, decomposable models are exempt from this difficulty as their prior normalising constants are analytically available. Madigan and York [29] proposed an MCMC approach to generating from the posterior distribution under decomposable model uncertainty which they called MC^3 . It is a Metropolis-Hastings method where, at each step, transition to a neighbouring model is proposed and the acceptance probability calculated using ratios of marginal likelihoods analogous to (4.21). This proved extremely effective for decomposable models. For more general conditional Dirichlet priors, though, we will focus on methods which approximate marginal likelihoods for individual models directly, combining with a search strategy where the size of the model space, M , makes computation for all $m \in M$ infeasible. The advantage for conjugate priors is that the same method of computation can be considered for computing the marginal likelihood and the prior normalising constant, as the integrand is of the same functional form in each case. If we denote the prior for model m as $f(\boldsymbol{\beta}_m) = C_m^{-1}g(\boldsymbol{\beta}_m|m)$, where g denotes the unnormalised expression given in (4.15), then we can write the marginal likelihood (4.18) as

$$\begin{aligned} f(\mathbf{n}|m) &= \int f(\mathbf{n}|m, \boldsymbol{\beta}_m)C_m^{-1}g(\boldsymbol{\beta}_m|m)d\boldsymbol{\beta}_m \\ &= \frac{\int f(\mathbf{n}|m, \boldsymbol{\beta}_m)g(\boldsymbol{\beta}_m|m)d\boldsymbol{\beta}_m}{\int g(\boldsymbol{\beta}_m|m)d\boldsymbol{\beta}_m} \end{aligned} \quad (4.22)$$

and then the integrands in the numerator and denominator can be seen to have the same functional form by comparing (4.12) and (4.15) and so s and α in the denominator are updated to $s + t$ and $\alpha + n$ in the numerator.

4.4.3 Laplace's Method

Arguably the most popular analytic approach to approximating integrals in Bayesian computation is *Laplace's Method*, developed in detail for this purpose by Tierney and Kadane [38]. Significantly [30] propose it for approximating both marginal likelihoods and posterior model probabilities for conjugate inference for log-linear models. Laplace's method is based on the principle that, provided L has a unique maximum θ , or is at least dominated by a single mode, then, for large n , the value of the integral strongly depends on the properties of L around the maximum;

specifically the log-integrand can be well-approximated by a quadratic function with the same mode and curvature at the mode. The approximation is

$$\int e^{nL(\boldsymbol{\theta})} d\boldsymbol{\theta} = \frac{(2\pi)^{\frac{d_m}{2}} e^{nL(\tilde{\boldsymbol{\theta}})}}{n^{\frac{d_m}{2}} |-\mathbf{H}(\tilde{\boldsymbol{\theta}})|^{\frac{1}{2}}} (1 + O(n^{-1}))$$

where $\tilde{\boldsymbol{\theta}}$ is the posterior mode, and \mathbf{H} is the Hessian (second derivative) matrix of L . Equivalently,

$$\begin{aligned} \log \int f(\mathbf{n}|\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta} &= \log f(\mathbf{n}|\tilde{\boldsymbol{\theta}}) + \log g(\tilde{\boldsymbol{\theta}}) \\ &+ \frac{d}{2} \log 2\pi - \frac{d}{2} \log n - \frac{1}{2} \log |-\mathbf{H}(\tilde{\boldsymbol{\theta}})| + O(n^{-1}) \end{aligned} \quad (4.23)$$

where \mathbf{H} is the Hessian for $\log(n^{-1}f(\mathbf{n}|\boldsymbol{\theta})g(\boldsymbol{\theta}))$. Kass and Wasserman [22] show that \mathbf{H} can be replaced in (4.23) by the Fisher (or observed) information matrix at the expense of the error increasing to $O(n^{-\frac{1}{2}})$.

Raftery [34] considered the problem of using Laplace's method to approximate Bayes factors for generalised linear models, utilising the output of standard statistical software in terms of the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_m$, the deviance, and the observed or expected Fisher information matrix to modify (4.23) at the expense of loss of accuracy. Nevertheless, on application to the calculation of Bayes factors for generalised linear models, the approximations were found to be of acceptable quality.

For a conditional Dirichlet prior (posterior), Laplace's method results in the approximation

$$\log \int g(\boldsymbol{\beta})d\boldsymbol{\beta} \approx \frac{d_m}{2} \log 2\pi - \frac{1}{2} \log |\alpha \mathbf{X}^T (\text{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^T) \mathbf{X}|^{1/2} + \log g(\tilde{\boldsymbol{\beta}}). \quad (4.24)$$

One key feature of a conditional Dirichlet prior is that, provided that the software allows non-integer cell counts, as does R for example, then the actual posterior mode and Hessian required in (4.24) can be obtained as the 'maximum likelihood estimate' and (inverse) variance-covariance matrix arising from 'cell counts' $\mathbf{n} + \boldsymbol{\alpha}$.

Diciccio et al. [10] compared several methods of estimating the Bayes factor when it is possible to obtain a sample from the posterior distribution. They presented a modified version of Laplace's method based on this, and a Bartlett adjustment to Laplace's method which improved the Laplace estimate by an order of magnitude. They also considered importance sampling and reciprocal importance sampling, two special cases of bridge sampling, which is described in detail in Sect. 4.4.5.

4.4.4 Evaluation of Laplace's Method for the Conditional Dirichlet

The aim of applying Laplace's method to the Conditional Dirichlet distribution is to obtain the normalising constant (marginal likelihood) for the (mostly) analytically intractable density function which results by conditioning on a particular log-linear model. Because, for decomposable models, the normalising constant is analytically available, this allows us to directly assess the fitness of Laplace's method for this purpose.

As Laplace's method approximates the log-integrand by a quadratic around the mode, an approximation which improves with increasing sample size n , it does raise a concern over whether it can accurately approximate the prior normalising constant for the modest values of α which are likely to be used. In particular, it is straightforward to observe that the tails of the conditional Dirichlet distribution are lighter than those of the Normal distribution, and so Laplace's method is likely to underestimate normalising constants when $\alpha(i)$ is small. The conditional Dirichlet distribution as expressed in (4.15) can be seen to have (conditional) tails which decay in each direction like $\exp(-|\beta_j|)$, more slowly than the quadratic assumed by Laplace's method. This means that Laplace's method is likely to produce approximations which underestimate the conditional Dirichlet normalising constants. In order to check how significant this underestimation is, and any dependence on the dimension and complexity of the log-linear model, we apply the method to certain conditional Dirichlet distributions resulting from several log-linear models which are decomposable, and hence of a tractable form, having known normalising constants. Similarly, as the approximation is of order $O(n^{-1})$ it is clear that the accuracy of the approximation will improve for large sample sizes. This is also investigated by obtaining Laplace approximations for increasing sample sizes, using a selection of models. The results are summarised in Figs. 4.11 and 4.12.

Figure 4.11 contains 8 plots representing 8 different log-linear models. Each plot is of the error in the log of the Laplace approximation (given as the log of the approximate value minus the log of the true value), against the value of the cell parameter $\alpha(i)$ (the hypothetical 'sample' in each cell). The parameters are equally distributed throughout the cells in each case. The cell parameter runs from 0.25 to 25 in each case. The 8 models are:

- | | |
|----------------------------------|--|
| (a) $A + B$ [2] | (e) $A + B + C + D$ [4] |
| (b) $AB + BC$ [5] | (f) $ABCD$ [16] |
| (c) ABC [8] | (g) $A^{(3)}B^{(3)}C^{(3)}D^{(3)}$ [81] |
| (d) $A^{(3)}B^{(3)}C^{(3)}$ [27] | (h) $A^{(4)}B^{(4)}C^{(4)}D^{(4)}$ [256] |

All variables have 2 levels, except where indicated, and the numbers in square brackets give the number of model parameters in each case.

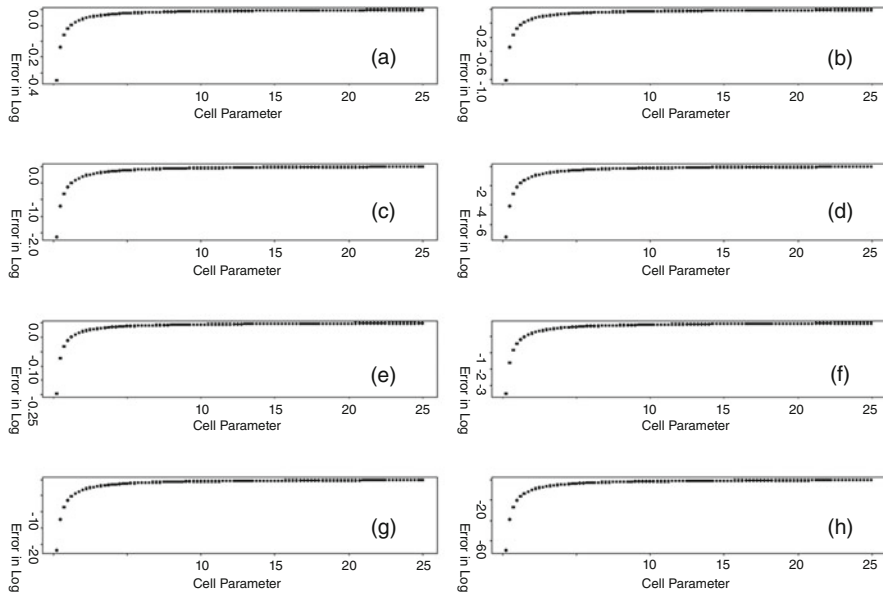


Fig. 4.11 Plots showing convergence of Laplace estimates for various models with equal samples in each cell

It is clear that for sample sizes greater than about 10 in each cell, the error of the approximation is negligible, and so the Laplace approximation is excellent. This is true for all the models. However it is also clear that, for certain models, the Laplace approximation for small values of cell parameters is poor, and so may not be reliably used to determine the normalising constant for (reference) prior distributions. Indeed, with all cell parameters equal to 0.5, the approximation for the 4-way saturated model where all variables have four levels has an error of -39 , which is huge. Examination of Fig. 4.11 shows that the error of the Laplace approximation increases significantly with increasing numbers of parameters in the model.

The approximations presented in Fig. 4.11 are all based on equal parameters in each cell. This is fine for prior distributions (where it seems that the Laplace approximation is of little use anyway), but is unrealistic for posterior distributions. In order to consider the unbalanced situation, Laplace approximations were obtained for posterior distributions where all the data was in a single cell. The results are presented graphically in Fig. 4.12. In each case, the ‘Cell Parameter’ refers to the data in the single cell. All other cells have a parameter of 0.25, representing a prior distribution based on a $Dirichlet(\frac{1}{4}\mathbf{1})$ distribution.

The graphs in Fig. 4.12 show that, when the data is distributed as described above, there is a considerable error in the Laplace approximation for all but the simplest model. It is therefore clear that the Laplace approximation to the normalising constant for conditional Dirichlet distributions is only reliable when there are at least

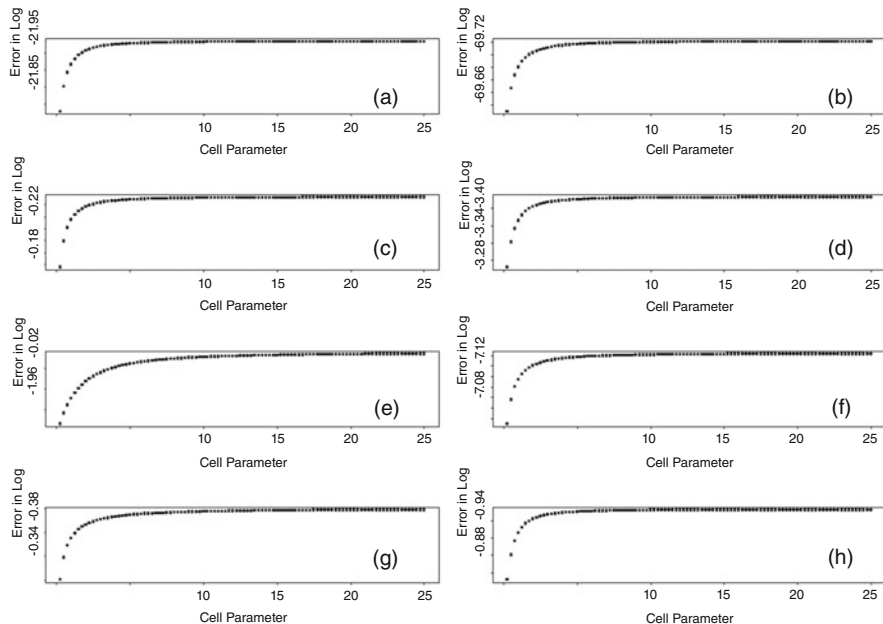


Fig. 4.12 Plots showing convergence of Laplace estimates for various models with unbalanced cell counts

a few observations in each cell. Exhaustive use of the Laplace approximation leads to the ‘rule of thumb’ that the approximation produced acceptable results when there were at least 5 observations in 80% of the cells, though note that the accuracy of the approximation always improves with greater total sample size and deteriorates with increasing numbers of model parameters.

In all the approximations presented, the error in the log normalising constants is negative, which implies that the approximation for the normalising constant is too small, as expected. In summary, it is clear that a more reliable method for obtaining prior normalising constants and marginal likelihoods for sparse tables is required.

4.4.5 Bridge Sampling

The class of techniques known as bridge sampling were introduced by Bennett [2], although they were studied in depth by Meng and Wong [31] and DiCiccio et al. [10]. The method allows the estimation of the ratio of two normalising constants, though it can be modified to allow the estimation of a single normalising constant.

Suppose we have two densities, f_1 and f_2 , and write these as

$$f_i = \frac{g_i}{C_i}$$

where $C_i = \int g_i$ for $i = 1, 2$. Now let γ be a function which satisfies

$$0 < \left| \int \gamma(\theta) f_1(\theta) f_2(\theta) d\theta \right| < \infty.$$

Then we may write

$$\frac{C_1}{C_2} = \frac{\int g_1(\theta) \gamma(\theta) f_2(\theta) d\theta}{\int g_2(\theta) \gamma(\theta) f_1(\theta) d\theta}. \quad (4.25)$$

Now let our unnormalised density of interest (prior or posterior) be denoted by $g(\theta)$, the associated normalising constant by C , and the normalised density by $f(\theta)$, so that $C = \int g(\theta) d\theta$ and $f(\theta) = \frac{g(\theta)}{C}$. Suppose we have a sample from f , and denote this by $\theta_1, \dots, \theta_m$. Let $q(\theta)$ be some density from which we may easily obtain a sample, and denote that sample by $\tilde{\theta}_1, \dots, \tilde{\theta}_M$. Now, in expression (4.25), let $g_1 = g$, $C_1 = C$, $g_2 = q$, and $C_2 = 1$. Then

$$C = \frac{\int g(\theta) \gamma(\theta) q(\theta) d\theta}{\int q(\theta) \gamma(\theta) f(\theta) d\theta} = \frac{E_{\theta \sim q}[g(\theta) \gamma(\theta)]}{E_{\theta \sim f}[q(\theta) \gamma(\theta)]}. \quad (4.26)$$

Using our samples, the bridge estimator of C introduced by Meng and Wong is given by

$$\hat{C} = \frac{\frac{1}{M} \sum_i g(\tilde{\theta}_i) \gamma(\tilde{\theta}_i)}{\frac{1}{m} \sum_i q(\theta_i) \gamma(\theta_i)}.$$

Clearly, a choice has to be made for the function γ . Several obvious choices are available—for example, $\gamma = \frac{1}{q}$ or $\gamma = \frac{1}{g}$. These reduce the bridge estimate to the commonly used estimates based on Importance Sampling and Reciprocal Importance Sampling. However, Meng and Wong found the optimal choice, in terms of minimising the mean squared error, is

$$\gamma(\theta) \propto \left[\frac{mg(\theta)}{C} + Mq(\theta) \right]^{-1}. \quad (4.27)$$

This would appear to be of little practical use, as it requires the normalising constant, C , in its calculation. However, it is possible to use an estimate of C produced by an alternative approximation method, and substitute this value in expression (4.27). For example, an estimate based on Laplace's method may be used, and indeed this is a technique which DiCiccio *et al.* found produced a discernible increase in the accuracy of the approximation compared to other bridge samplers (for example, importance sampling).

In practice, repeated applications of the bridge sampler may be used to iteratively update the approximation, using the previous value of C each time. This is the

method which will be applied in the next section to the conditional Dirichlet distribution.

We will apply bridge sampling to estimate the normalising constant of the conditional Dirichlet distribution. For this purpose, we choose q to be a Normal density with mean equal to the mode of the conditional Dirichlet distribution, and variance matrix equal to the inverse of the Hessian matrix of second derivatives. The bridge estimate (4.26) allows the size of the samples from densities q and f to differ, though for this application they will be equal, and denoted by m .

Let the (un-normalised) conditional Dirichlet density (as in (4.15)) be denoted by $g(\boldsymbol{\beta})$, the sample from this be denoted $\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, \dots, \boldsymbol{\beta}^{(m)}$, and the Normal sample generated from density q be denoted $\tilde{\boldsymbol{\beta}}^{(1)}, \tilde{\boldsymbol{\beta}}^{(2)}, \dots, \tilde{\boldsymbol{\beta}}^{(m)}$. The bridge sampler will be applied iteratively, with the j th iteration denoted C_j . Then the bridge estimate is given by the expression

$$C_j = \int g(\boldsymbol{\beta}) d\boldsymbol{\beta} \approx \frac{\sum_i g(\tilde{\boldsymbol{\beta}}^{(i)}) \gamma_j(\tilde{\boldsymbol{\beta}}^{(i)})}{\sum_i q(\boldsymbol{\beta}^{(i)}) \gamma_j(\boldsymbol{\beta}^{(i)})}$$

where

$$\gamma_j(\boldsymbol{\beta}) = \left[\frac{mg(\boldsymbol{\beta})}{C_{j-1}} + mq(\boldsymbol{\beta}) \right]^{-1}$$

and C_0 is the estimate for the normalising constant by Laplace's method.

4.4.6 Numerical Examples

In this section, the bridge sampler will be used to obtain prior and posterior normalising constants for a set of log-linear models where the true value is also available, as in Sect. 4.3.1 (Laplace approximations). Successive runs of the bridge sampler produce values which, after about 3 iterations, seem to fluctuate slightly about a common value. Hence, to produce the estimates below, the bridge sampler is run iteratively 10 times, taking the Laplace estimate as a starting value, and the result presented is the mean of the final 7 iterations.

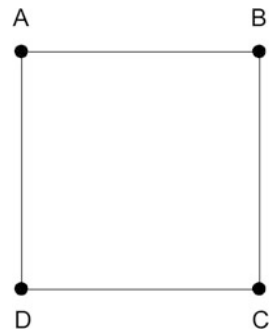
Table 4.2 gives the bridge sampling estimates for the log of the prior normalising constants, together with the error (expressed as the estimate minus the true value), and the value of the prior parameters, which are the same for each cell. All variables have 2 levels, except where indicated.

It is clear from the table that the bridge sampling approximation is extremely good, even for distributions where the prior parameter is small. It therefore represents a huge improvement over the Laplace estimates, where the errors were of a much higher magnitude. Such accuracy is also evident when the parameters in each cell are not equal (the unbalanced case).

Table 4.2 Bridge estimates, and their respective errors, of normalising constants for various models

Hierarchical log-linear model	Prior parameter	Bridge approximation	Error in bridge approximation
$A + B$	0.25	2.29	0
$AB + BC$	0.125	9.16	0
ABC	0.125	16.15	0.01
$A^{(3)}B^{(3)}C^{(3)}$	0.5	-5.86	-0.05
$A + B + C + D$	0.0625	4.57	-0.01
$ABCD$	0.0625	43.94	0.11
$A^{(3)}B^{(3)}C^{(3)}D^{(3)}$	0.5	-61.78	0.33

Fig. 4.13 Independence graph for the non-decomposable graphical (hierarchical) model $\{(A, B), (B, C), (C, D), (D, A)\}$



The approximations in Table 4.2 were all obtained using Gibbs sample sizes of 10,000. This choice was motivated by the desire for the bridge estimate to vary by less than 0.1 about its limit, and for the sample to be produced reasonably quickly using the Gibbs sampler. Smaller sample sizes are adequate for simpler models.

We have demonstrated the accuracy of the method of bridge sampling to determine the normalising constants for the conditional Dirichlet prior for several decomposable models (where exact results are possible). However, there is one graphical model with up to and including 4 variables which is not decomposable. This is the model represented by Fig. 4.13.

Table 4.3 gives the normalising constants for the conditional Dirichlet distributions for this model, with varying numbers of levels of the variables. The prior parameters in each case are symmetric, with a single observation split throughout the table (i.e. $\alpha(i) = \frac{1}{|I|}$).

Note that many other non-graphical log-linear models exist for which this approach is required; for example, the model $AB + BC + AC$.

Table 4.3 Normalising constants for model $AB + BC + CD + DA$

Levels of A, B, C, D	log(Normalising Constant)
2, 2, 2, 2	1.45
3, 2, 2, 2	2.75
2, 3, 2, 2	3.67
3, 3, 2, 2	6.64
3, 3, 3, 2	10.09
3, 3, 3, 3	12.78

4.4.7 Example: Risk Factors for Coronary Heart Disease

In Sect. 4.3.2, the Gibbs sampler was used to obtain posterior samples from a number of models fitted to a contingency table summarising incidence of coronary heart disease, originally presented by Edwards and Havranek [12], and analysed further by Madigan and Raftery [28] and Dellaportas and Forster [8].

The marginal likelihood for a log-linear model with conditional Dirichlet prior is given by (4.22) and therefore the Bayes factor, the second ratio of the right of (4.19), can be written as

$$B_{12} = \frac{\int f(\mathbf{n}|m_1, \boldsymbol{\theta}_{m_1})g(\boldsymbol{\theta}_{m_1}|m_1)d\boldsymbol{\theta}_{m_1} \int g(\boldsymbol{\beta}_{m_2}|m_2)d\boldsymbol{\beta}_{m_2}}{\int f(\mathbf{n}|m_2, \boldsymbol{\theta}_{m_2})g(\boldsymbol{\theta}_{m_2}|m_2)d\boldsymbol{\theta}_{m_2} \int g(\boldsymbol{\beta}_{m_1}|m_1)d\boldsymbol{\beta}_{m_1}}$$

where $f(\mathbf{n}|m, \boldsymbol{\theta}_m)$ is the likelihood under model m and $f(\boldsymbol{\theta}_m|m)$ is the prior under model m .

In this application, the prior approximation to $\log \int g(\boldsymbol{\theta}_m|m)d\boldsymbol{\theta}_m$ will be obtained using the bridge sampler, and the posterior approximation to $\log \int g(\mathbf{n}|m, \boldsymbol{\theta}_m)g(\boldsymbol{\theta}_m|m)$ obtained using Laplace’s method. This is sensible, as the sample size is large, with cell counts of at least 5 in 80% of the cells. The results are presented in Table 4.4, which gives the estimated log Bayes factors for several models, taken against the most probable hierarchical model $AC + BC + AD + AE + CE + DE + F$, for a prior where $a(i) = 1/64$. A sample size of 5000 was used for the prior estimates.

The top three models in the table are the most probable hierarchical models (identified by Dellaportas and Forster), and the fourth is the most probable decomposable model. There is a good deal of agreement between the bridge/Laplace estimates and

Table 4.4 Estimated Bayes factors for heart disease data

Hierarchical log-linear model	Log Bayes factor estimate	Log Bayes factor (D&F)
$AC + BC + AD + AE + BE + DE + F$	0.49	0.57
$AC + BC + AD + AE + BE + CE + DE + F$	1.35	1.34
$AC + BC + AD + AE + CE + DE + BF$	1.79	1.42
$BC + ACE + ADE + F$	8.25	>6

those obtained by Dellaportas and Forster. Note that we are using different prior densities here, so don't expect exact agreement with their results.

4.5 Further Examples

The data from [12] have been used throughout this chapter to illustrate Bayesian inference for log-linear models. Here we present two further illustrative examples.

4.5.1 Example 1: Lymphoma and Chemotherapy

This example concerns 30 patients suffering from lymphocytic lymphoma, and cross-classifies their type of lymphoma L (nodular or diffuse) against their response to combination chemotherapy R and their sex S . The data are presented in Table 4.5 and were originally analysed by Skarin et al. [35].

Three different diffuse prior distributions are investigated, in order to determine the posterior model probabilities for the 8 potential graphical models. The first is the conditional Dirichlet distribution with parameters $\alpha(i) = 1/2$, corresponding to conditioning on Jeffreys' prior for the saturated model. The second is the conditional Dirichlet distribution with parameters $\alpha(i) = 1/8$, which corresponds to a single observation distributed evenly between all cells (Perks' prior). Note that the equivalence of the hyper-Dirichlet and conditional Dirichlet distributions is used to ease the calculations. Elsewhere the methods described above, such as bridge sampling using Monte Carlo samples, are used. The third prior distribution is a log-Normal prior with parameters chosen using the same considerations as [8]. The posterior model probabilities are presented in Table 4.6 (models with probability less than 0.01 are excluded).

All priors identify the most probable model, namely $RC + CS$. Similar probabilities are also obtained for the model $RC + RS$. However, the various priors differ with respect to models $RC + S$ and RCS . As expected, the conditional Dirichlet distribution with $\alpha(i) = 1/8$ tends to favour the simpler model $RC + S$.

Table 4.5 Chemotherapy and lymphoma

Cell type	Sex	Remission	
		No	Yes
Nodular	Male	1	4
	Female	2	6
Diffuse	Male	12	1
	Female	3	1

Table 4.6 Posterior model probabilities for cancer data using various prior distributions

Model	Conditional Dirichlet $\alpha(i) = 1/2$	Conditional Dirichlet $\alpha(i) = 1/8$	Log-normal prior
<i>RC + CS</i>	0.48	0.42	0.48
<i>RC + S</i>	0.19	0.38	0.30
<i>RC + RS</i>	0.22	0.18	0.17
<i>RCS</i>	0.09	0.01	0.05

4.5.2 Example 2: Toxaemia in Pregnancy

The data in Table 2 are presented in [20] and is a cross-classification of record 13,384 pregnant women by their socio-economic class (*C*—5 levels), their smoking habit (*S*—none, light, or heavy), and whether or not they suffer from two toxaemic signs, hypertension (*H*) and proteinuria (*P*). The data was collected in England between 1968 and 1977, and the aim of the analysis of the $2 \times 2 \times 3 \times 5$ contingency Table 4.7 is to determine relationships between the variables, via the posterior model probabilities for all possible graphical models.

As in the first example, three prior distributions are used. These are the conditional Dirichlet distribution with parameters $\alpha(i) = 1/2$, the conditional Dirichlet distribution with parameters $\alpha(i) = 1/60$, and a log-Normal prior. Under each of the distributions, a maximum of two models were identified as having posterior probabilities greater than 0.001. These are the models *HP + PS + SC* and

Table 4.7 Toxaemia in pregnancy

Social class	Smoking	Proteinuria			
		Yes		No	
		Hypertension		Hypertension	
		Yes	No	Yes	No
1	None	28	82	21	286
	Light	5	24	5	71
	Heavy	1	3	0	13
2	None	50	266	34	785
	Light	13	92	17	284
	Heavy	0	15	3	34
3	None	278	1101	164	3160
	Light	120	492	142	2300
	Heavy	16	92	32	383
4	None	63	213	52	656
	Light	35	129	46	649
	Heavy	7	40	12	163
5	None	20	78	23	245
	Light	22	74	34	321
	Heavy	7	14	4	65

Table 4.8 Posterior model probabilities for Toxaemia data using various prior distributions

Model	Conditional Dirichlet $\alpha(i) = \frac{1}{2}$	Conditional Dirichlet $\alpha(i) = \frac{1}{60}$	Log-normal prior
<i>HP + PS + SC</i>	0.9950	1.0000	1.0000
<i>HPS + SC</i>	0.0050	0.0000	0.0000

HPS + SC, and their respective probabilities are shown in Table 4.8. These results conflict somewhat with the classical approach based on stepwise model selection using analysis of deviance. That approach prefers model *HP + PS + SC + CH*. However, each of the priors used here gives a posterior model probability $< 10^{-6}$ to this model. The reason for this is that the five levels of factor *C* mean that to include any log-linear term involving this factor requires an increase in dimensionality of the log-linear parameter of at least 4. The Bayesian approach adopted here is much more cautious about adding extra complexity to the model unless it is justified by a commensurate improvement in fit. One way to address this is to consider models which allow extra interactions but parameterised more efficiently, which is possible for ordinal factors (but outside the scope of this chapter; see [39] for further details).

4.6 Summary

In this chapter we have described a Bayesian methodology for log-linear models, with particular attention paid to the use of conjugate conditional Dirichlet distribution, which has the attractive property that its parameters may be interpreted as prior cell counts. This makes it useful for both reference analyses, where small prior values are used, and as an informative prior, where (hypothetical) prior cell counts may be available. Our treatment relies heavily on the pioneering work of Dawid and Lauritzen [6] and [30], but focussing on practical computational methodology. Our main focus has been on inference under model uncertainty where computation of the marginal likelihood is the key. The requirement to obtain normalising constants for the unnormalised forms of both the prior and posterior conditional Dirichlet leads us to recommend the use of a bridge sampler for approximating the required integrals, particularly for the prior.

References

1. Albert, J.H.: Bayesian selection of log-linear models. *Can. J. Stat.* **24**, 327–347 (1996)
2. Bennett, C.: Efficient estimation of free energy differences from Monte Carlo data. *J. Comput. Phys.* **22**, 245–268 (1976)
3. Berger, J.O., Pericchi, L.R.: The intrinsic Bayes factor for model selection and prediction. *J. Am. Stat. Assoc.* **91**, 109–122 (1993)

4. Cowell, R.G., Dawid, A.P., Lauritzen, S.L., Spiegelhalter, D.J.: Probabilistic Networks and Expert Systems. Springer-Verlag, New York. (1996)
5. Darroch, J.N., Lauritzen, S.L., Speed, T.P.: Markov fields and log-linear interaction models for contingency tables. *Ann. Stat.* **8**, 522–539 (1980)
6. Dawid, A.P., Lauritzen, S.L.: Hyper Markov laws in the statistical analysis of decomposable graphical models. *Ann. Stat.* **21**, 1272–1317 (1993)
7. Dawid, A.P., Lauritzen, S.L.: Compatible prior distributions. In: Bayesian Methods with Applications to Science, Policy and Official Statistics: Proceedings of the 6th World Meeting of the International Society for Bayesian Analysis, pp. 109–118. Office for Official Publications of the European Communities, Luxembourg (2001)
8. Dellaportas, P., Forster, J.J.: Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika* **86**, 615–633 (1999)
9. Dellaportas, P., Smith, A.F.M.: Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling. *Appl. Stat.* **42**, 443–459 (1993)
10. DiCiccio, T.J., Kass, R.E., Raftery, A., Wasserman, L.: Computing Bayes factors by combining simulation and asymptotic approximations. *J. Am. Stat. Assoc.* **92**, 903–915 (1997)
11. Dobra, A., Massam, H.: The mode oriented stochastic search (MOSS) algorithm for log-linear models with conjugate priors. *Stat. Methodol.* **7**, 204–253 (2010)
12. Edwards, D., Havranek, T.: A fast procedure for model search in multidimensional contingency tables. *Biometrika* **72**, 339–351 (1985)
13. Forster, J.J.: Symmetric models and prior distributions for multiway contingency tables. *Stat. Methodol.* **7**, 210–224 (2010)
14. Forster, J.J., Webb, E.L.: Bayesian disclosure risk assessment: predicting small frequencies in contingency tables. *Appl. Stat.* **56**, 551–570 (2007)
15. Gilks, W.R., Wild, P.: Adaptive rejection sampling for Gibbs sampling. *Appl. Stat.* **41**, 337–348 (1992)
16. Good, I.J.: On the estimation of small frequencies in contingency tables. *J. R. Stat. Soc. Ser. B* **18**, 113–124 (1956)
17. Good, I.J.: On the application of symmetric Dirichlet distributions and their mixtures to contingency tables. *The Annals of Statistics* **4**, 1159–1189 (1976)
18. Green, P.: Reversible jump MCMC computation and Bayesian model determination. *Biometrika* **82**, 711–732 (1995)
19. Günel, E., Dickey, J.: Bayes factors for independence in contingency tables. *Biometrika* **61**, 545–557 (1974)
20. Hand, D.J., Daly, F., McConway, K., Lunn, D., Ostrowski, E.: *A Handbook of Small Data Sets*. CRC Press, Boca Raton (1993)
21. Jeffreys, H.: An invariant form for the prior probability in estimation problems. *Proced. R. Soc. A* **186**, 453–461 (1946)
22. Kass, R.E., Wasserman, L.: A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Am. Stat. Assoc.* **90**, 928–934 (1995)
23. Knuiman, M.W., Speed, T.P.: Incorporating prior information into the analysis of contingency tables. *Biometrics* **44**, 1061–1071 (1988)
24. Laird, N.M.: Empirical Bayes methods for two-way contingency tables. *Biometrika* **65**, 581–590 (1978)
25. Leonard, T.: Bayesian estimation methods for two-way contingency tables. *J. R. Stat. Soc. B* **37**, 23–37 (1975)
26. Lidstone, G.J.: A note on the general case of the Bayes-Laplace formula for inductive or a posteriori probabilities. *Trans. Fac. Actuaries Scotl.* **8**, 182–192 (1920)
27. Lindley, D.V.: The Bayesian analysis of contingency tables. *Ann. Math. Stat.* **35**, 1622–1643 (1964)
28. Madigan, D., Raftery, A.E.: Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Am. Stat. Assoc.* **89**, 1535–1545 (1994)
29. Madigan, D., York, J.: Bayesian graphical models for discrete data. *Int. Stat. Rev.* **63**, 215–232 (1995)

30. Massam, H., Liu, J., Dobra, A.: A conjugate prior for discrete hierarchical log-linear models. *Ann. Stat.* **37**, 3431–3467 (1995)
31. Meng, X.-L., Wong, W.H.: Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Stat. Sin.* **6**, 831–860 (1996)
32. O’Hagan, A.: Fractional Bayes factors for model comparison. *J. R. Stat. Soc. B* **57**, 99–138 (1995)
33. Perks, W.: Some observations on inverse probability including a new indifference rule. *J. Inst. Actuaries* **73**, 285–334 (1947)
34. Raftery, A.E.: Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika* **83**, 251–266 (1996)
35. Skarin, A.T., Pinkus, G.S., Myerowitz, R.L., Bishop, Y.M., Moloney W.C.: Combination chemotherapy of advanced lymphocytic lymphoma: importance of histologic classification in evaluating response. *Cancer* **34**, 1023–1029 (1974)
36. Spiegelhalter, D.J., Lauritzen, S.L.: Sequential updating of conditional probabilities on directed graphical structures. *Networks* **20**, 579–605 (1990)
37. Spiegelhalter, D.J., Smith, A.F.M.: Bayes factors for linear and log-linear models with vague prior information. *J. R. Stat. Soc. B* **44**, 377–387 (1982)
38. Tierney, L., Kadane, J.B.: Accurate approximations for posterior moments and marginal densities. *J. Am. Stat. Assoc.* **81**, 82–86 (1986)
39. Webb, E.L., Forster, J.J.: Bayesian model determination for multivariate ordinal and binary data. *Comput. Stat. Data Anal.* **52**, 2632–2649 (2008)
40. Wright, S.: Correlation and causation. *Journal of Agricultural Research* **20**, 557–585 (1921)

Chapter 5

Simple Ways to Interpret Effects in Modeling Binary Data



Alan Agresti, Claudia Tarantola, and Roberta Varriale

5.1 Introduction

Suppose you are consulting with a non-statistician colleague in academia, government, or industry, for a study that has a binary response variable. If you use a standard binary regression model such as logistic or probit regression, is your colleague able to understand its natural effect measures, such as odds ratios or probit differences? In our consulting experiences as well as teaching such methods to students in various disciplines, interpretation can be challenging.

Models for binary responses that apply link functions to the probability of “success,” such as logistic regression models, are generalized linear models that employ non-linear link functions. With such models, effect parameters are not as simple to interpret as slopes and correlations for ordinary linear regression. This article surveys simple measures that can supplement the ordinary model-based measures, being easier to interpret. Our intention is not to present new methodology but rather to show ways of using existing approaches to supplement the most popular model-based analyses as well as more complex models for binary data.

We consider a binary response variable y taking values 0 and 1 and a set of explanatory variables (x_1, \dots, x_p) , which may be a mixture of quantitative

A. Agresti (✉)

Department of Statistics, University of Florida, Gainesville, FL, USA

e-mail: aa@stat.ufl.edu

C. Tarantola

Department of Economics and Management, University of Pavia, Pavia, Italy

e-mail: claudia.tarantola@unipv.it

R. Varriale

Istat and La Sapienza University, Rome, Italy

e-mail: varriale@istat.it

© Springer Nature Switzerland AG 2023

M. Kateri, I. Moustaki (eds.), *Trends and Challenges in Categorical Data Analysis*,

Statistics for Social and Behavioral Sciences,

https://doi.org/10.1007/978-3-031-31186-4_5

and categorical. In describing effect summaries for comparing two groups of a categorical explanatory variable, we sometimes use a separate indicator variable z to distinguish between the groups. The logistic regression model, defined by

$$\log \left[\frac{P(Y = 1)}{1 - P(Y = 1)} \right] = \alpha + \beta z + \beta_1 x_1 + \cdots + \beta_p x_p,$$

is a generalized linear model (GLM) with link function $\text{logit}[P(Y = 1)] = \log[P(Y = 1)/(1 - P(Y = 1))]$. This model has effects most naturally interpreted using odds ratios. For example, adjusting for the other explanatory variables, the odds that $y = 1$ for the group having $z = 1$ divided by the odds that $y = 1$ for the group having $z = 0$ are

$$\frac{P(Y = 1 \mid z = 1, x_1, \dots, x_p)/P(Y = 0 \mid z = 1, x_1, \dots, x_p)}{P(Y = 1 \mid z = 0, x_1, \dots, x_p)/P(Y = 0 \mid z = 0, x_1, \dots, x_p)} = \exp(\beta).$$

The coefficient β_k of x_k is the change in the log odds per each 1-unit increase in x_k , adjusting for the other explanatory variables, so $\exp(\beta_k)$ is a multiplicative effect of each 1-unit increase in x_k on the odds of response $y = 1$ versus response $y = 0$.

To compare two levels of an explanatory variable such as two groups, however, it is easier for methodologists or practitioners to understand a difference or a ratio of *probabilities* than a ratio of *odds*. In our experience, many (even some statisticians) misinterpret the odds ratio as if it were a ratio of probabilities. When two groups have probabilities close to 0, the ratio of odds is similar to the ratio of probabilities, but this is not true otherwise. In fact, [22] noted that when the probabilities exceed 0.2, the odds ratio is better approximated by the *square* of the ratio of probabilities. For example, if an odds ratio is 9, one group may have success probability merely about 3 times the success probability for the other group.

Other aspects of logistic regression that are due to its nonlinear link function are not as well known to users. For instance, suppose explanatory variables x_1 and x_2 are uncorrelated, such as in many experimental designs. In ordinary linear models, the estimated effect of x_1 is the same when x_1 is the sole predictor as when x_1 and x_2 are joint predictors. For logistic regression, this is not the case with model-based odds effect measures. For instance, the effect β_1^* when x_1 is the sole predictor relates approximately to the effect β_1 when x_2 is also in the model by $\beta_1^* \approx \beta_1 \sqrt{3.29/[3.29 + \beta_2^2 \text{var}(x_2)]}$, where $3.29 = \pi^2/3$ is the variance of the standard logistic distribution [18]. For the model with probit link, $\beta_1^* = \beta_1 \sqrt{1/[1 + \beta_2^2 \text{var}(x_2)]}$. Equality of the effects in the two cases is, however, approximately true for the simpler measures discussed in this article.

The structure of this paper is as follows. In Sect. 5.2, we show that generalized linear models using the identity link function and the log link function, although not as natural for binary data, have simpler summaries and can sometimes supplement logistic and probit models. We illustrate these summary measures with an Italian study to model an employment response variable. In Sect. 5.3, we focus on probit

and logit models and we present alternative probability-based summaries that can be used to study the effect of an explanatory variable, while adjusting for other explanatory variables in the model. For group comparisons, these include average differences and average log-ratios of probabilities and comparisons that result directly from corresponding latent variable models. In this section we also show the correspondence between these effect measures obtained for logistic and probit models and the model-based effect measures obtained with the identity and log link functions. We conclude this section illustrating the proposed measures with the Italian study. Section 5.4 uses the measures of Sect. 5.3 to aid in interpreting effects for more complex models, such as generalized additive models. We illustrate with an example about horseshoe crab mating, generalizing existing results for a logistic model.

5.2 Alternative Models for Binary Data

Standard models for binary response variables are special cases of the GLM

$$\text{link}[P(Y = 1)] = \alpha + \beta z + \beta_1 x_1 + \cdots + \beta_p x_p, \quad (5.1)$$

for link functions such as the logit and probit. For describing effects, we find it useful to refer to the model expressed as

$$F^{-1}[P(Y = 1)] = \alpha + \beta z + \beta_1 x_1 + \cdots + \beta_p x_p, \quad (5.2)$$

where the link function F^{-1} is the inverse of a standard cumulative distribution function (cdf). For logistic regression, $F(z) = \exp(z)/[1 + \exp(z)]$ is the standard logistic cdf. For probit regression, F is the standard normal cdf, which we denote by Φ . The nonlinear link function naturally produces effects on the link scale. For example, with the probit link, β is the difference between $F^{-1}[P(Y = 1)]$ when $z = 1$ and when $z = 0$, and β_k is the change in $F^{-1}[P(Y = 1)]$ per each 1-unit increase in x_k , adjusting for the other explanatory variables. Such effect measures are not easy to interpret by those who need to understand the effects in more real-world terms. Although the probit model was the first model for binary data to receive much attention (pre-dating logistic regression by nearly 10 years), its use by methodologists has undoubtedly been hampered by the difficulty of interpretation unless one uses a corresponding latent variable model. The same applies to other link functions that are potentially very useful, such as those with log-log and complementary log-log link functions.

In addition, effects often behave in a way that is counterintuitive to those mainly familiar with ordinary linear models. For example, as mentioned in the introductory section, if an explanatory variable uncorrelated with x_1 is added to a logistic regression model, the partial effect of x_1 is typically different than in the model without the other explanatory variable; it would be identical in an ordinary linear

model. For contingency tables, this relates to standard collapsibility results [e.g., 1, pp. 53–54]. For example, consider several 2×2 tables relating binary y to binary x_1 at different levels for x_2 . If the difference or the ratio of proportions is the same in each table, then when x_1 and x_2 are marginally independent, the marginal table collapsing over x_2 has the same value for that measure. For the odds ratio, however, collapsibility occurs when x_1 and x_2 are *conditionally* independent, given y , rather than marginally independent. Because of this, regardless of correlation structure among explanatory variables, it can be challenging to compare the effect of an explanatory variable to its effect when other variables are added to the model. Generally, the relation between conditional and marginal effect measures depends on the model and measure considered. For related literature, particularly for logistic regression, see [4, 7, 8, 10, 19], and [20]. Related remarks also occur in comparing effects in marginal models for multivariate responses with effects in corresponding models that add a random effect to the model [1, pp. 495–497].

5.2.1 Identity and Log Link Models for Binary Data

For comparing groups, simple difference and ratio measures on the proportion scale result from alternative link functions in model (5.1). For the identity link function, the coefficient β of an indicator variable in that model is the difference between $P(Y = 1)$ for two groups, adjusting for other variables. The corresponding model is called the *linear probability model*. For the log link function, β is the log ratio of probabilities.

Generalized linear models with identity and log link functions are relatively rarely used for binary data. The link values for the linear probability model are restricted to the $[0, 1]$ range, rather than the entire real line that is the range of linear predictor values in the model. The log-link values are restricted to negative values. Because of these restrictions, ordinary maximum likelihood (ML) fitting of such models, assuming a binomial distribution for the response, may fail. One can always fit the linear probability model using least squares, as in fitting ordinary linear models, but the fitted values may be outside $[0, 1]$ for some values of explanatory variables. When ML works for such a model and it fits the data decently, however, one obtains the advantage of simpler interpretation of effects than in the logistic model.

The appearance of the linear probability model is similar to the logistic and probit models for probabilities between about 0.2 and 0.8. To illustrate, the first panel in Fig. 5.1 shows 500 observations in which X was uniformly distributed over $(0, 100)$, and conditional on $X = x$, $P(Y = 1)$ follows a logistic model with $P(Y = 1)$ increasing from 0.2 to 0.8 over the range of x values. (For clarity of showing the data, the binary observations are jittered slightly.) The figure also shows the ML fits of the logistic and linear probability models. The appearance of the log-link model is similar to the logistic and probit models when probabilities are uniformly less than about 0.25 over the ranges of explanatory-variable values and similar to those

models with link applied to $P(Y = 0)$ when probabilities are uniformly above about 0.75. To illustrate, the second panel of Fig. 5.1 shows 500 observations in which X was uniformly distributed over $(0, 100)$, and conditional on $X = x$, $P(Y = 1)$ follows a logistic model with $P(Y = 1)$ increasing from 0.01 to 0.25 over the range of x values. The figure also shows the ML fits of the logistic model and the model with log link.

When we have reason to expect probabilities to fall in the previously specified ranges, we believe that it can be helpful in summarizing the size of an effect to use the models with identity and log link functions, even if only to supplement ordinary logistic and probit models. In addition, the binary models with identity and log links share the property with ordinary linear models that effects remain stable when explanatory variables are added to the model that are uncorrelated with ones already in the model.

5.2.2 Example: Models for Italian Survey Data

In this section, we fit generalized linear models with logit, log and identity link functions to some data from a simple random sample of about 100,000 Italians from the Toscana region in December 2015. The information comes from administrative sources collected and organized by Istituto Nazionale di Statistica (Istat). Administrative data relevant for the labor statistics derive mainly from social security and fiscal authority and are organized in an information system having a linked employer-employees structure. From this data structure it is possible to obtain information on the statistical unit of interest, i.e., the worker. The response variable y indicates whether the subject is present in any administrative source ($1 = \text{yes}$, $0 = \text{no}$). Assuming there are no measurement errors, a person not present in an administrative labor source either is not working or is doing so illegally, so in the following we refer to y as whether employed ($1 = \text{yes}$, $0 = \text{no}$). The examined explanatory variables are $x_1 = \text{gender}$ ($1 = \text{female}$, $0 = \text{male}$), $x_2 = \text{Italian}$ (1 if the individual is an Italian citizen, 0 otherwise), and $x_3 = \text{pension}$ (1 if the individual is receiving a pension, 0 otherwise). For Istat confidentiality reasons, we cannot report the exact data, but we provide in tables the approximate cross-classified sampled proportions.

We first restrict attention to the 27,775 subjects having age over 65. The sample proportions that were employed ($y = 1$) in the eight cases that cross classify the three explanatory variables were small, so we fitted models both with logit and log links, as shown in Table A1 of the Appendix. The main-effects model fits are

$$\text{logit}[\hat{P}(Y = 1)] = -1.8686 - 1.3236x_1 - 0.4295x_2 + 0.2162x_3$$

and

$$\log[\hat{P}(Y = 1)] = -2.0374 - 1.2388x_1 - 0.3619x_2 + 0.2003x_3$$

Fig. 5.1 Data sets showing jittered binary data and fits of logistic regression model and (1) linear probability model when $P(Y = 1)$ varies from 0.2 to 0.8, (2) log-link model when $P(Y = 1)$ is less than 0.25

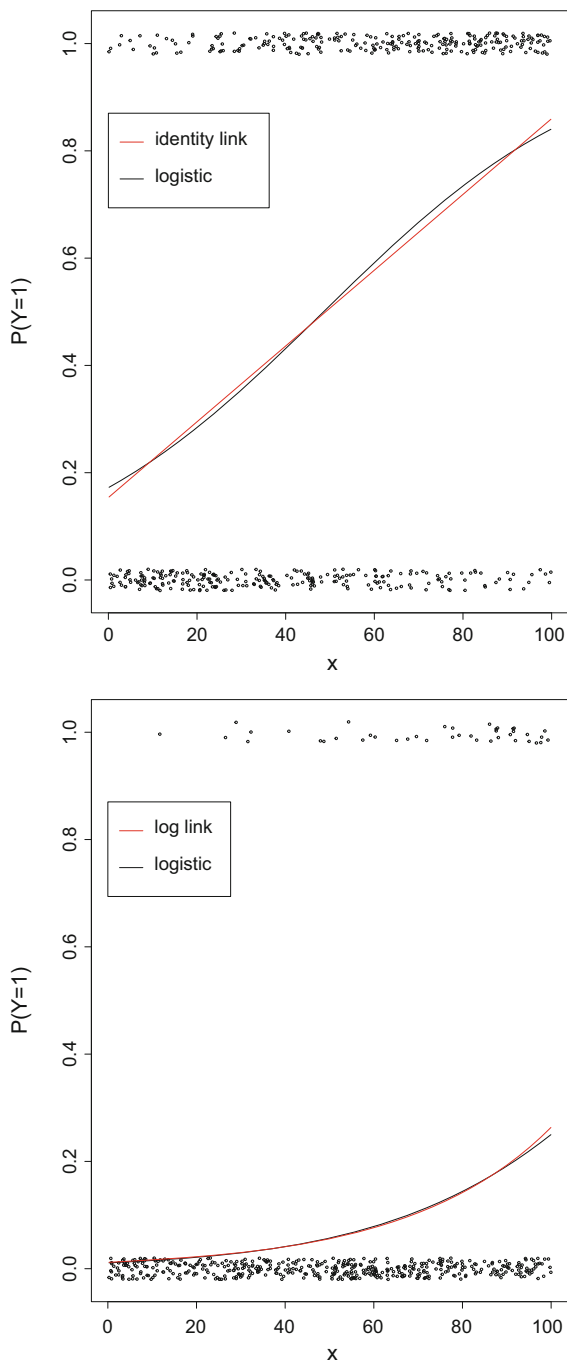


Table 5.1 Fitted values for Istat sample of older subjects, for models with logit and log links for predicting employment using gender (G), Italian (I), and pension (P)

G	I	P	Main effects		Gender/Italian interaction		Sample proportion (Sample size)
			Logit link	Log link	Logit link	Log link	
			$\hat{P}(Y = 1)$	$\hat{P}(Y = 1)$	$\hat{P}(Y = 1)$	$\hat{P}(Y = 1)$	
1	1	1	0.0321	0.0321	0.0314	0.0314	0.0316 (13,300)
1	1	0	0.0260	0.0263	0.0255	0.0257	0.0244 (2300)
1	0	1	0.0485	0.0461	0.0865	0.0864	0.0690 (100)
1	0	0	0.0395	0.0378	0.0709	0.0708	0.0812 (200)
0	1	1	0.1109	0.1109	0.1118	0.1118	0.1116 (11,000)
0	1	0	0.0913	0.0908	0.0920	0.0915	0.0949 (600)
0	0	1	0.1608	0.1593	0.1106	0.1111	0.1238 (100)
0	0	0	0.1337	0.1304	0.0911	0.0909	0.0800 (100)

Note: The sample sizes for the sample proportions were not the actual ones used but are rounded to the nearest hundredth, for Istat confidentiality reasons

Table 5.1 shows the fitted values for the two models. They are uniformly very close, with the absolute difference averaged over the 27,775 cases being only 0.000097. The residual deviances (for the grouped data files) are 13.17 and 13.85, with $df = 4$.

The log-link model has the advantage of simplicity of interpretation, the exponentiated coefficients estimating ratios of probabilities instead of ratios of odds. For instance, adjusting for whether an Italian citizen and whether receiving a pension, the probability that a woman is employed is estimated to be $\exp(-1.2388) = 0.2897$ times the probability that a man is employed.

Table 5.1 also shows sample proportions for the eight cases. Both models show clear lack of fit for the non-Italians, although the sample sizes for those cases are relatively small. In fact, for non-Italians, fitted and sampled values are quite different. Improved fits result from adding an interaction term between gender and whether an Italian citizen to reflect that the gender effect seems to be larger for Italian citizens than for non-citizens. Table 5.1 also shows fitted values for the model with this interaction term, with logit and log links. For this model, fitted values are again uniformly very close, with the absolute difference averaged over the 27,775 cases being only 0.000062. The residual deviances are 1.35 and 1.42 with $df = 3$.

We next consider the 72,225 subjects having age under 65. The sample proportions that were employed ($y = 1$) in the eight cases that cross classify the three explanatory variables fell between 0.20 and 0.75, so we fitted models both with logit and identity links, as shown in Appendix Table A2. The main-effects model fits are

$$\text{logit}[\hat{P}(Y = 1)] = 0.3502 - 0.6440x_1 + 0.7017x_2 - 1.8737x_3$$

and

$$\hat{P}(Y = 1) = 0.5876 - 0.1386x_1 + 0.1513x_2 - 0.4078x_3$$

Again, the identity-link model has the advantage of simplicity of interpretation. For instance, adjusting for whether an Italian citizen and whether receiving a pension, the probability that a woman is employed is estimated to be 0.1386 lower than the probability that a man is employed. (Interestingly, the estimated effects of x_2 and x_3 have reverse sign from the estimated effects for the older sample, and the gender effect in the logit model is about half the size.)

Table 5.2 shows the fitted values for the two models and sample proportions for the eight cases. The fits are quite close, with the absolute difference averaged over the 72,225 cases being only 0.00430. These two models show lack of fit for the non-Italians with a pension, although these are only 195 of the 72,225 cases. Improved fits result from adding an interaction term between the Italian citizen and pension variables. The gender main-effect estimate in the identity-link model changes only from -0.1386 to -0.1397 . Table 5.2 also shows fitted values for the interaction models with logit and identity links. They are quite close, with the average absolute difference being only 0.00286. The residual deviances are 15.80 and 30.32 with $df = 3$, not particularly large for this enormous sample size.

We do not wish to suggest by these examples that one should *not* use logistic regression. Indeed, an obvious advantage of it compared to the models with log and identity links is that it is relevant regardless of the range of values for $P(Y = 1)$. However, we believe that the log-link model and identity-link model can sometimes supplement the logit-link model, in particular by providing effect interpretations that are simpler for many to understand.

Table 5.2 Fitted values for Istat sample for younger subjects, for models with logit and identity links for predicting employment using gender (G), Italian (I), and pension (P)

G	I	P	Main effects		Italian/Pension interaction		Sample proportion (Sample size)
			Logit	Identity	Logit	Identity	
			$\hat{P}(Y = 1)$	$\hat{P}(Y = 1)$	$\hat{P}(Y = 1)$	$\hat{P}(Y = 1)$	
1	1	1	0.1876	0.1924	0.1845	0.1775	0.1991 (3400)
1	1	0	0.6006	0.6002	0.6011	0.6020	0.5974 (27,700)
1	0	1	0.1027	0.0410	0.2153	0.2119	0.2202 (100)
1	0	0	0.4271	0.4489	0.4243	0.4334	0.4339 (5200)
0	1	1	0.3054	0.3310	0.3012	0.3171	0.2879 (3800)
0	1	0	0.7411	0.7389	0.7416	0.7416	0.7453 (27,500)
0	0	1	0.1789	0.1797	0.3433	0.3516	0.3372 (100)
0	0	0	0.5867	0.5875	0.5840	0.5731	0.5725 (4400)

Note: The sample sizes for the sample proportions were not the actual ones but are rounded to the nearest hundred, for Istat confidentiality reasons

5.3 Alternative Effect Measures for Explanatory Variables

Because of the range restrictions for probabilities, the identity and log links are often not appropriate. But even in fitting a model such as logistic or probit regression, one can construct summary measures based on differences and ratios of probabilities to help others understand the size of the effects. In this section, we describe two types of interpretation that supplement estimated model-parameter effects with simpler effects reported on the probability scale rather than on the scale of the link function. Such effects also exhibit greater stability in terms of the impact of uncorrelated explanatory variables.

When a binary regression model of generalized linear model form contains solely main effects, $P(Y = 1)$ changes monotonically as a quantitative explanatory variable increases, with others at fixed values. This is the situation that we assume in forming these supplementary summary measures.

5.3.1 Probability Effect Measures

A simple summary for the effect of an explanatory variable x_k averages the rate of change in $P(Y = 1)$, as a function of x_k . For this, we consider the expression (5.2) of the model, namely $F^{-1}[P(Y = 1)] = \alpha + \beta z + \beta_1 x_1 + \cdots + \beta_p x_p$. Let $f(y) = \partial F(y)/\partial y$ denote the corresponding probability density function. For a quantitative explanatory variable x_k , the rate of change in $P(Y = 1)$ when other explanatory variables are fixed at certain values \mathbf{x}^* is

$$\partial P(Y = 1 | \mathbf{x} = \mathbf{x}^*) / \partial x_k = f(\alpha + \beta z^* + \beta_1 x_1^* + \cdots + \beta_p x_p^*) \beta_k.$$

These measures are denoted in different ways depending on the context; for example, the econometric literature [6] uses the term *elasticity*, while the statistics literature calls them either *marginal effects* or *partial effects*. Long and Mustillo [16] and many others refer to such an instantaneous effect as a *marginal effect*. This terminology is a bit misleading, as this partial derivative refers to a *conditional* effect of x_k rather than its *marginal* effect as the term *marginal* is commonly used (i.e., for a sole predictor, collapsing over the other explanatory variables). Some authors, e.g. [14], instead use the term *partial effect*, which we use in this paper.

For the logit link, the partial effect for x_k on $P(y = 1)$ has the expression

$$\partial P(Y = 1 | \mathbf{x} = \mathbf{x}^*) / \partial x_k = \beta_k P(y = 1 | \mathbf{x} = \mathbf{x}^*) [1 - P(y = 1 | \mathbf{x} = \mathbf{x}^*)].$$

This takes values bounded above by its highest value of $\beta_k/4$ that occurs when $P(Y = 1 | \mathbf{x} = \mathbf{x}^*) = 1/2$. For probit models, the highest value of this instantaneous change is $\beta_k/\sqrt{2\pi}$, also when $P(Y = 1 | \mathbf{x} = \mathbf{x}^*) = 1/2$. These maximum values need not be relevant, as $P(Y = 1)$ need not be near 1/2 for most or all the data.

Any particular way of fixing values of the explanatory variables has its corresponding partial effect value for x_k . Long and Mustillo [16] summarize various versions. Here, we mainly consider the *average partial effect*, which estimates the partial effect of x_k at each of the n sample values of the explanatory variables, and then averages them. We could instead estimate the partial effect with every explanatory variable, including x_k , set at its mean, which is the *partial effect at the mean*. Or, we could set all explanatory variables at values considered to be of particular interest. This might be more appropriate if the sample is not random or not representative of the population of interest, in which case it is sometimes referred to as a *partial effect at a representative value*. For each version, the summary value obtained still reflects the effect of x_k adjusting for the explanatory variables, unlike an averaging of the effect over values of a random effect in a generalized linear mixed model, in which case the effect changes nature to being population-averaged and can have quite different magnitude.

For a categorical explanatory variable, for each version one would instead use a *discrete change*, estimating the change in $P(Y = 1)$ for a change in an indicator variable. To compare two groups, for instance, for the n sample observations, we could find the difference between estimates of $P(Y = 1)$ when $z = 1$ and when $z = 0$ at the sample values for the other predictors and average the obtained values. When the number of possible values of the categorical explanatory variable is greater than two, the discrete change is computed as the difference in the predicted probabilities for cases in one category relative to the reference level.

Discrete changes are also relevant for quantitative explanatory variables, to summarize estimated changes in $P(Y = 1)$ over a particular range of x_k values. For example, to summarize the effect of a quantitative variable x_k on y , it can be useful to report the difference between the model-fitted estimate of $P(Y = 1)$ at the maximum and minimum values of x_k , when other explanatory variables are set at particular values such as their means. A caveat for such measures is that their relevance depends on the plausibility of x_k taking extreme values when all other explanatory variables fall at their means. Also, this summary can be misleading when outliers exist on x_k , in which case one can instead report the estimated probabilities at more resistant quantiles. Reporting them at the upper and lower quartiles of x_k summarizes the estimated change in $P(Y = 1)$ over the range of the middle half of the observations on x_k , with other explanatory variables fixed. Such a measure has greater scope for reflecting reality.

A useful and easy-to-obtain measure that we've not seen proposed for the two-group comparison focuses on *ratios* of estimated probabilities for the two groups. For example, we could average the n log-ratios of probability estimates, to obtain a measure comparable to the effect in the log-link GLM, and then exponentiate that average for interpretive purposes. Again, other versions are possible, such as finding the ratio at the mean of the other explanatory variables. Such measures would seem to be especially useful when fitted probabilities are near 0 for the groups being compared.

Greene [9, pp. 775–785] showed how to obtain standard errors for the maximum likelihood estimators of some effect measures based on instantaneous rates of

change and differences of probabilities. We have used the bootstrap to obtain a standard error (SE) for the log-ratio measure just proposed. Mood [18] pointed out that the average partial effect has behavior reminiscent of effects in ordinary linear models, in the sense that it is roughly stable when we add an explanatory variable to the model that is uncorrelated with the variable for which we are describing the effect. Such behavior is expected, as such an average partial effect typically takes similar value as the effect using the linear probability model discussed in Sect. 5.2. The effect measures are available in software, such as presented by Leeper [13], Long and Freese [15, pp. 341–351], and Sun [21, pp. 527–531]. Agresti and Tarantola [3] and Iannario and Tarantola [12] proposed analogous measures for modeling ordinal data.

5.3.2 A Probability Summary for Ordered Comparison of Groups

It is sometimes realistic to regard a categorical variable as crude measurement of an underlying continuous latent variable y^* that, if we could observe it, would be the response variable for an ordinary linear model. In fact, model (5.2) is implied by a model in which a latent response has conditional distribution with standard cdf given by the inverse of the link function [1, p. 252]. We next use this connection to suggest an alternative way to summarize an effect, in the context of comparing two groups ($z = 0$ and $z = 1$). Let y_1^* and y_2^* denote independent underlying latent variables for the binary response, representing the underlying distributions when $z = 1$ and when $z = 0$ respectively. At a particular setting \mathbf{x} for other explanatory variables, $P(Y_1^* > Y_2^*; \mathbf{x})$ is a summary measure of relative size, suggested by Agresti and Kateri [2] for ordinal response variables.

The normal latent variable model with $y^* \sim N(\beta z + \beta_1 x_1 + \dots + \beta_p x_p, 1)$ implies the probit model

$$\Phi^{-1}[P(Y = 1)] = \alpha + \beta z + \beta_1 x_1 + \dots + \beta_p x_p,$$

with α the cutpoint on the underlying scale between y^* values for which $y = 1$ and for which $y = 0$. For this model,

$$P(Y_1^* > Y_2^*; \mathbf{x}) = P\left[\frac{(y_1^* - y_2^*) - \beta}{\sqrt{2}} > \frac{-\beta}{\sqrt{2}}\right] = \Phi\left(\frac{\beta}{\sqrt{2}}\right). \tag{5.3}$$

This is true regardless of the \mathbf{x} value, so we denote it by $P(Y_1^* > Y_2^*)$. For the logit link,

$$P(Y_1^* > Y_2^*) \approx \frac{\exp(\beta/\sqrt{2})}{[1 + \exp(\beta/\sqrt{2})]}, \tag{5.4}$$

for the β coefficient of z in the logistic model.

When the latent variable model for binary data is realistic, this type of probability comparison of the groups supplements the ordinary interpretation of the effect coefficient β . As β increases from 0, the probability increases from 0.5 toward 1. In addition, a natural way to construct a summary measure of predictive power is to estimate R^2 for the linear model that is specified for the underlying latent response variable. McKelvey and Zavoina [17] suggested this measure for a probit model for ordinal responses, for which the underlying latent variable model is the ordinary normal linear model, but it applies also for binary data and for other link functions for ordinal data [3].

5.3.3 Example: Measures for Italian Survey Data

We illustrate these measures for the example from earlier in this article of modeling Italian employment status. For simplicity, here we consider only the main effects models.

The average partial effect for a logistic model approximates the corresponding effect from the binary model with identity link. For the younger age group, we obtained the gender effect of -0.1386 ($SE = 0.0035$) with the identity-link model. The estimated average partial effect for the model with logit link is -0.1409 ($SE = 0.0035$). This can be easily found with an existing package in R applied to the ungrouped data file, with code such shown in Table A2 in the Appendix.

The average partial effect for log-ratios that we suggested for a logistic model approximates the corresponding effect from the binary model with log link. For the older age group the gender effect estimate is equal to -1.2388 ($SE = 0.0516$) with the log-link model. The estimated average log-ratio partial effect for the model with logit link is -1.2398 ($SE = 0.0517$). Table A3 in the Appendix presents edited R code for obtaining the estimated average log-ratio partial effect and for using the bootstrap with 1000 resamplings of the data to obtain its SE . As one can do with the log-link model parameter estimate, one could utilize the asymptotic normality of the sample measure to obtain a corresponding confidence interval for the population value, such as the 95% confidence interval $-1.2398 \pm 1.96(0.0517)$, which is $(-1.341, -1.138)$. The exponentiated endpoints of the interval, that is $(0.26, 0.32)$, are a confidence interval on the probability-ratio scale. (Recall that the log-link model provided ML estimate 0.2897.) Alternatively, one can find a bootstrap confidence interval, such as shown with the percentile method in Table A3.

Whether a latent variable is sensible for measuring propensity toward employment is debatable. But if so, from Eq. (5.4) with the estimated gender effect $\hat{\beta} = -1.3236$ for the older sample, the estimated probability that a randomly selected female would be higher on the latent variable than a randomly selected male is $\exp(-1.3236/\sqrt{2})/[1 + \exp(-1.3236/\sqrt{2})] = 0.282$. For the younger sample, the effect is $\exp(-0.6440/\sqrt{2})/[1 + \exp(-0.6440/\sqrt{2})] = 0.388$.

5.4 Generalized Additive Model for Binary Data

A generalized additive model (GAM) replaces the linear predictor in a binary generalized linear model (GLM) by additive unspecified smooth functions. Its basic version has the form

$$\text{link}[P(y = 1)] = s_1(x_1) + \cdots + s_j(x_j) + \cdots + s_p(x_p),$$

where the smooth function s_j is typically based on cubic splines [11] and more generally uses basis expansions of low rank with complexity controlled by ridge penalties on regression coefficients [e.g. 23]. The name *additive* derives from the additive structure of the predictor. GAMs have the advantage over GLMs of greater flexibility, with an ordinary GLM with $s_k(x_k)$ replaced by $\beta_k x_k$. In practical application, it is often helpful to use both smooth and linear terms in a model. Using a graphical portrayal of a GAM fit, we may discover patterns that we would miss with ordinary GLMs, and we may obtain potentially better estimates of mean responses. A disadvantage of GAMs and other smoothing methods, compared with GLMs, is that interpretability is even more difficult. It can be more difficult to summarize an effect and judge when it has substantive importance.

Fasiolo et al. [5] described an efficient visual method for interpreting GAMs, using the `mcgViz` package in R. The proposed methods include ones to bin the data and summarize them in a form that can be displayed effectively, interactive Q-Q plots, portrayals of conditional residuals, and visualizations of the uncertainty of the fitted smooth effects. To supplement these with simple numerical summaries, we believe that measures that aid in interpreting binary GLMs can also be useful for GAMs. When an effect of a quantitative explanatory variable seems to be monotonic and not highly variable in the degree of non-linearity, useful measures include measures of average partial rates of change of probabilities and comparisons of the fitted probability at extreme values or quartiles (or other quantiles) of the explanatory variable.

How does one describe quantitatively the effect of an explanatory variable or obtain a confidence intervals for the true effect? Here, we suggest a way to construct an estimated average partial effect using the fit of a GAM. For explanatory variable k , let $\mathbf{x}_{i(k)}$ denote the values of the other explanatory variables for observation i . The fitted rate of change for explanatory variable x_k for observation i can be approximated by

$$[\hat{P}(y = 1 \mid \mathbf{x}_{i(k)}, x_{ik} + \epsilon) - \hat{P}(y = 1 \mid \mathbf{x}_{i(k)}, x_{ik} - \epsilon)]/2\epsilon$$

for a very small ϵ . Finding the mean of these values for the n observations yields an approximate average partial effect for that predictor. We suggest starting with a trial value such as $\epsilon = 0.000001$ and then using a smaller value yet to ensure that results are stable to several decimal places. For comparing two groups, one could find an

average difference or average ratio of estimated probabilities at the n values of the explanatory variables.

5.4.1 Example: GAM for Horseshoe Crab Study

We illustrate the use of the average partial effect in the context of GAMs with a data set analyzed extensively with logistic regression in [1], from a study of nesting horseshoe crabs. During spawning season, a female migrates to the shore to breed. With a male attached to her spine, she lays clusters of eggs, which are fertilized externally. During spawning, other male crabs, called *satellites*, may cluster around the pair and fertilize the eggs. The response outcome for each female crab is whether she had any satellites (1 = yes, 0 = no). Explanatory variables associated with this response were the female crab's carapace (shell) width, which is a summary of her size, and her color (four categories from light to dark), which is a surrogate for the crab's age, older crabs being darker. In the sample, width had a mean of 26.3 cm and a standard deviation of 2.1 cm. Logistic modeling showed that width had a positive effect on the presence of a satellite, and color being dark (category 4) had a negative effect.

The logistic ML fit with predictors width and an indicator for color that is 1 for dark-colored crabs and 0 for others is

$$\text{logit}[\hat{P}(Y = 1)] = -11.6790 + 0.4782(\text{width}) - 1.3005(\text{color}),$$

with standard errors of 0.104 for width and 0.526 for color. Table A4 in the appendix shows edited results for the logistic regression and a GAM fit with these data, which is

$$\text{logit}[\hat{P}(Y = 1)] = -11.2470 + s(\text{width}) - 1.2805(\text{color}),$$

$s(\text{width})$ being a smoothing spline. Figure 5.2 shows the GAM fit, with jittered observations. Adding an interaction term does not provide a significantly improved fit.

Table A5 in the Appendix shows edited R code for finding the average partial effects for width and for color for this GAM as well as for the corresponding logistic model. Interpretation is relatively simple. For the logistic fit at the $n = 173$ observed width values, the average rate of change is 0.087 in the estimated probability of a satellite per 1 cm increase in width, adjusting for color. At those width values, the estimated probability of a satellite averages 0.261 lower if the crab has dark color than if it has a lighter color. For the GAM, the corresponding values are 0.085 (standard error = 0.015) and 0.254 (standard error = 0.112), quite similar because the logistic model fits relatively well.

Table A6 in the Appendix shows edited R code for using the bootstrap with 1000 resamplings of the data to obtain standard errors and confidence intervals for the

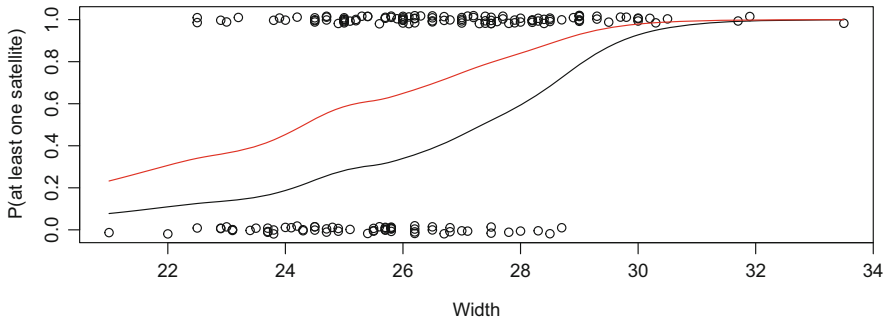


Fig. 5.2 Portrayal of GAM fit for the effects of width and color (black for dark, red for other colors) on jittered responses for whether a female horseshoe crab has at least one satellite

average partial effects in the GAM. For example, the bias-corrected and accelerated (BC_a) confidence interval of $(-0.466, -0.028)$ for the average partial effect for color indicates that at the sampled width values, the probability of a satellite is estimated to average between 0.028 and 0.466 lower if the crab has dark color than if it has a lighter color. The relatively wide interval reflects partly that the sample had only 22 dark-colored crabs.

5.5 Discussion and Future Research

Future research could apply the methods of this paper to other models for binary responses. In particular, using alternative link functions to aid in interpretation would be useful for marginal models, whether fitted by GEE methods or maximum likelihood. The binary and log links are more challenging for random effects models, as the usual assumption of normally-distributed random effects adds another restriction to models with bounded range values. Effect measures such as average partial effects are also relevant for models for multi-category responses. See [3] for their use with cumulative link models for ordinal responses.

Perhaps more challenging for future research is the development of effect measures for generalized additive models. The average partial effect measure presented in this article is of use when relationships are monotone, but often that is not the case. Even when it is the case, difference or ratio effects are sometimes highly variable across the range of an explanatory variable, and a single summary may be too simplistic. Also for the binary generalized linear models considered here, we assumed that $P(Y = 1)$ is monotone in quantitative explanatory variables, and alternative measures are needed when this is not the case.

In summary, in these days in which statistical science is ever more visible, partly because of the emergence of data science and methods for “big data,” it is increasingly important for statisticians to develop ways to present relatively simple

summaries of complex methods that will be understandable by a relatively wide audience. We hope that this paper is a step in that direction.

Acknowledgments The authors appreciate helpful comments from two referees and from Pablo Inchausti and Maria Kateri.

Appendix

This appendix provides the source code for the R analyses described in the text.

Table A1 R code for fitting logistic and log-link models to the older-age Istat sample

```
-----
>Italian1 <- read.csv("http://www.stat.ufl.edu/~aa/cat/data/Italian_older.csv",
+                    header=TRUE)
>mod.logit <- glm(empl ~ female + italian + pension, family=binomial,
+                data=Italian1)
>summary(mod.logit) # fit of logistic model; default link is logit
      Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.8686     0.1631  -11.46  <2e-16
female      -1.3236     0.0546  -24.26  <2e-16
italian      -0.4295     0.1632   -2.63  0.0085
pension      0.2162     0.0948    2.28  0.0225
---
>mod.log <- glm(empl ~ female + italian + pension, family=binomial(link=log),
+              data=Italian1)
>summary(mod.log) # fit of model with log link function
      Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.0374     0.1465  -13.91  <2e-16
female      -1.2388     0.0516  -24.00  <2e-16
italian      -0.3619     0.1460   -2.48  0.013
pension      0.2003     0.0885    2.26  0.024
-----
```

Table A2 R code for fitting logistic and linear probability models to the younger-age Istat sample and finding the average partial effect for the logistic regression model

```
-----
>Italian2 <- read.csv("http://www.stat.ufl.edu/~aa/cat/data/Italian_younger.csv",
+                    header=TRUE)
>mod.logit <- glm(empl ~ female + italian + pension, family=binomial,
+                data=Italian2)
>summary(mod.logit)
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.3502     0.0224   15.6 <2e-16
female       -0.6440     0.0161  -39.9 <2e-16
italian       0.7017     0.0225   31.2 <2e-16
pension      -1.8737     0.0288  -65.1 <2e-16
---
> mod.linprob <- glm(empl ~ female + italian + pension,
+                   family=quasi(link=identity, variance="mu(1-mu)"), data=Italian2)
>summary(mod.linprob) # fit of linear probability model
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.5876     0.0052  112.4 <2e-16
female       -0.1386     0.0035  -40.1 <2e-16
italian       0.1513     0.0052   29.1 <2e-16
pension      -0.4078     0.0052  -78.4 <2e-16
---
>library(mfx)
>logitmfx(mod.logit, atmean=FALSE, data=Italian2)
Marginal Effects:
      dF/dx Std. Err.      z    P>|z|
female -0.14062  0.00346 -40.6 <2e-16
italian  0.15820  0.00512  30.9 <2e-16
pension -0.41602  0.00508 -81.9 <2e-16
-----
```

Table A3 R code for finding average log-ratio partial effect and bootstrap *SE* and bootstrap CI for the logistic regression model applied to the older-age Istat sample

```

-----
>library(plyr)
>library(boot)
>attach(Italian1)
---
APER.log<-function(formula, data, indices, fam, var_exp)
{
  dat<-data[indices,]
  mod <- glm(formula, family=fam, data=dat)
  pred.prob <- (predict(mod,type="response"))
  var_exp_ind<-var_exp[indices,]
  r_new<- as.data.frame(cbind(var_exp_ind, pred.prob))
  r1_new<-count(r_new, vars = c(names(r_new)))
  row_r1<-nrow(r1_new)
  pred.prob.Male_new<-r1_new$pred.prob[1:(row_r1/2)]
  pred.prob.Female_new<-r1_new$pred.prob[((row_r1/2)+1):row_r1]
  r2_new <- count(var_exp_ind[,-1],vars = c(names(var_exp_ind)[-1]))
  APER.log_new <- ((log(pred.prob.Female_new/pred.prob.Male_new)%*%r2_new$freq)
  +
    /sum(r2_new$freq))
  return(APER.log_new)
}
APER.log(formula=empl ~ female + italian + pension, data=Italian1, indices=
+ c(1:nrow(Italian1)), fam=binomial,var_exp=cbind(female, italian, pension))
[1,] -1.2398
---
APER.log_boot <- boot(data=Italian1, statistic=APER.log, R=1000,
  formula=empl ~ female + italian + pension, fam=binomial,
  var_exp=cbind(female, italian, pension) )
> APER.log_boot
Bootstrap Statistics :
  original      bias    std. error
t1*  -1.2398 -0.00039268  0.051689
---
> boot.ci(APER.log_boot,type="perc")
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates
Intervals :
Level      Percentile
95%      (-1.344, -1.141 )
-----

```

Table A4 R code for GAM fit for using width and color as predictors of whether a female horseshoe crab has any satellites

```
-----
>Crabs <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Crabs.dat",
+                   header=TRUE)
>Crabs$c4 <- ifelse(Crabs$color == 4, 1, 0) # indicator for color cat. 4
>fit.glm <- glm(y ~ width + c4, family=binomial, data=Crabs)
>summary(fit.glm)
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.6790     2.6925  -4.338 1.44e-05
width        0.4782     0.1041   4.592 4.39e-06
c4          -1.3005     0.5259  -2.473 0.0134
Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 187.96  on 170  degrees of freedom
---
>library(gam)
>fit.gam <- gam(y ~ s(width) + c4, family=binomial, data=Crabs)
>summary(fit.gam)
Null Deviance: 225.7585 on 172 degrees of freedom
Residual Deviance: 185.4678 on 167.0001 degrees of freedom
Anova for Parametric Effects
              Df Sum Sq Mean Sq F value Pr(>F)
s(width)     1  17.774  17.7736  18.3127 3.15e-05
c4           1   5.928   5.9278   6.1076 0.01446
Residuals 167 162.084   0.9706
>fit.gam$coefficients
(Intercept)      c4
-11.2470      -1.2805
-----
```

Table A5 R code for finding the average partial effects for width and for color for the logistic regression model and for the generalized additive model for the presence of horseshoe crab satellites

```
-----
> Crabs <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Crabs.dat",
                      header=TRUE)
> Crabs$c4 <- ifelse(Crabs$color == 4, 1, 0) # indicator for dark color
> fit <- glm(y ~ width + c4, family=binomial, data=Crabs)
> library(mfx)
> logitmfx(fit, atmean=FALSE, data=Crabs) # with atmean=TRUE, finds
Marginal Effects:                                     # effect only at the mean
      dF/dx Std. Err.      z    P>|z|
width  0.08748   0.02447   3.5748  0.00035
c4     -0.26142   0.10569  -2.4735  0.01338
---
dF/dx is for discrete change for the following variables: "c4"

#Function to obtain Average Partial Effects in a GAM model
APE_GAM<-function(formula,data,indices, pvar1,pvar2, fam,epsilon){
  d <- data[indices,]
  fit <- gam(formula,family=fam, data=d)
  data_plus <- data
  data_minus <- data
  data_plus[,pvar1] <- data[,pvar1] + epsilon
  data_minus[,pvar1] <- data[,pvar1] - epsilon
  data1_plus <- data_plus[,c(pvar1, pvar2)]
  data1_minus <- data_minus[,c(pvar1, pvar2)]
  tvec <- (predict(fit, data1_plus, type="response")
           - predict(fit,data1_minus, type="response"))/(2*epsilon)
  APE <- mean(tvec)
  return(APE)
}

# APE for width
> APE_GAM(formula = y ~ s(width) + c4, data=Crabs, indices=c(1:nrow(Crabs)),
+         pvar1=5, pvar2=8, fam=binomial, epsilon=0.000001)
[1] 0.0850665

# Function to obtain a discrete change in a GAM model
dchange_GAM <-function(formula, data, indices, pvar1, pvar2, fam,epsilon){
  d <- data[indices,]
  fit <- gam(formula,family=fam, data=d)
  data_plus <- data
  data_minus <- data
  data_plus[,pvar2] <- 1
  data_minus[,pvar2] <- 0
  data1_plus <- data_plus[,c(pvar1, pvar2)]
  data1_minus <- data_minus[,c(pvar1, pvar2)]
  tvec <- (predict(fit, data1_plus, type="response")
           - predict(fit, data1_minus, type="response"))
  APE <- mean(tvec)
  return(APE)
}

dchange_GAM (formula = y ~ s(width) + c4,data=Crabs,indices=c(1:173),
             pvar1=5, pvar2=8, fam=binomial,epsilon=0.000001)
[1] -0.2539021
-----
```

Table A6 R code for using a bootstrap to find confidence intervals for the average partial effects for width and for color for the generalized additive model for the presence of horseshoe crab satellites

```
-----
#width variable
> APE_boot <- boot(data=Crabs, statistic=APE_GAM, R=1000, formula =
+ y ~ s(width) + c4, pvar1=5, pvar2=8, fam=binomial, epsilon=0.000001)
> APE_boot
Bootstrap Statistics :
      original      bias    std. error
t1* 0.0850665 -0.0007855886 0.01513634
> boot.ci(APE_boot, type="bca")
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates
Level           BCa
95%            ( 0.0536,  0.1122 )

#color variable
> disc_boot_1 <- boot(data=Crabs, statistic=dchange_GAM, R=1000, formula =
+ y ~ s(width) + c4, pvar1=5, pvar2=8, fam=binomial, epsilon=0.000001)
> disc_boot_1
Bootstrap Statistics :
      original      bias    std. error
t1* -0.2539021 -0.002005479 0.1118384

      boot.ci(disc_boot_1, type="bca")
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates
Level           BCa
95%            (-0.4658, -0.0285 )
-----
```

References

1. Agresti, A.: *Categorical Data Analysis*, 3rd edn. Wiley, Hoboken (2013)
2. Agresti, A., Kateri, M.: Ordinal probability effect measures for group comparisons in multinomial cumulative link models. *Biometrics* **73**, 214–219 (2017)
3. Agresti, A., Tarantola, C.: Simple ways to interpret effects in modeling ordinal categorical data. *Statistica Neerlandica* **72**, 210–223 (2018)
4. Brumback, B., Berg, A.: On effect-measure modification: relationships among changes in the relative risk, odds ratio, and risk difference. *Stat. Med.* **27**, 3453–3465 (2008)
5. Fasiolo, M., Nedellec, R., Goude, Y., Wood, S.N.: Scalable visualisation methods for modern generalized additive models. *J. Comput. Graph. Stat.* **29**, 78–86 (2020)
6. Franses P.H., Paap, R.: *Quantitative Models in Marketing Research*. Cambridge University Press, Cambridge (2004)
7. Gail, M.H.: Adjusting for covariates that have the same distribution in exposed and unexposed cohorts. In: Moolgavkar, S.H., Prentice, R.L. (eds.) *Modern Statistical Methods in Chronic Disease Epidemiology*, pp. 3–18. Wiley, New York (1986)
8. Gail, M.H., Wieand, S., Piantadosi, S.: Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* **71**, 431–444 (1984)

9. Greene, W.H.: *Econometric Analysis*, 6th edn. Pearson Prentice Hall, Upper Saddle River (2008)
10. Greenland, S., Robins, J.M., Pearl, J. Confounding and collapsibility in causal inference. *Stat. Sci.* **14**, 29–46 (1999)
11. Hastie, T., Tibshirani, R.: *Generalized Additive Models*. Chapman and Hall, London (1990)
12. Iannario, M., Tarantola, C.: How to interpret the effect of covariates on the extreme categories in ordinal data models. *Sociol. Methods Res.* (2021). <https://doi.org/10.1177/0049124120986179>
13. Leeper, T.J.: *Margins: An R Port of Stata's Margins Command*. R package version 0.3.23 (2018)
14. Long, J.S.: *Regression Models for Categorical and Limited Dependent Variables*. Sage, Thousand Oaks (1997)
15. Long, J.S., Freese, J.: *Regression Models for Categorical Dependent Variables Using Stata*, 3rd edn. Stata Press, College Station (2014)
16. Long, J.S., Mustillo, S.A.: Using predictions and marginal effects to compare groups in regression models for binary outcomes. *Sociol. Methods Res.* **50**, 1284–1320 (2021)
17. McKelvey, R.D., Zavoina, W.: A statistical model for the analysis of ordinal level dependent variables. *J. Math. Sociol.* **4**, 103–120 (1975)
18. Mood, C.: Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *Eur. Sociol. Rev.* **26**, 67–82 (2010)
19. Robinson, L., Jewell, N.P.: Some surprising results about covariate adjustment in logistic regression models. *Int. Stat. Rev.* **59**, 227–240 (1991)
20. Stanghellini, E., Doretti, M.: On marginal and conditional parameters in logistic regression models. *Biometrika* **106**, 732–739 (2019)
21. Sun, C.: *Empirical Research in Economics: Growing up with R*. Pine Square LLC, Baton Rouge (2015)
22. VanderWeele, T.J.: Optimal approximate conversions of odds ratios and hazard ratios to risk ratios. *Biometrics* **76**, 746–752 (2020)
23. Wood, S.N.: *Generalized Additive Models: An Introduction with R*, 2nd edn. CRC Press, Boca Raton (2017)

Chapter 6

Mean and Median Bias Reduction: A Concise Review and Application to Adjacent-Categories Logit Models



Ioannis Kosmidis

6.1 Overview

The first part of this chapter provides an example-driven, concise review of the developments in a fast growing body of literature about mean and median bias reduction (BR) in parametric estimation via adjusted score equations; see [10] for mean BR (mBR) and [15] for median BR (mdBR). Particular focus is placed on how these methods can be used as a remedy for the numerical and inferential consequences of boundary maximum likelihood (ML) estimates in categorical response models, which are illustrated in Sect. 6.2. Sections 6.3 and 6.4 describe how the mean and median bias of the ML estimator can be reduced in general parametric models through the appropriate adjustment of the gradient of the log-likelihood. Section 6.5 discusses the validity of inference when the ML estimates are replaced by mBR or mdBR estimates in standard first-order procedures. Section 6.6 takes a close look at the equivariance properties of mBR and mdBR estimators under transformation of the model parameters. We also present an approximation of the bias of general transformations of mBR estimators, which can be used to correct for the bias of transformations of the model parameters using only the mBR estimates, the second derivatives of the transformation, and the expected information matrix. The bias approximation is used to get mBR estimates of odds ratios from mBR estimates of regression coefficients in logistic regression models.

The second part of this chapter uses the results from the first to develop, for the first time, mBR and mdBR procedures for adjacent-categories logit (ACL) models for ordinal responses (see, for example, [2], Chapter 4 for an introduction). Section 6.7 reviews the proportional odds (PO) and non-proportional odds (NPO)

I. Kosmidis (✉)

Department of Statistics, University of Warwick, Coventry, UK
e-mail: ioannis.kosmidis@warwick.ac.uk

© Springer Nature Switzerland AG 2023

M. Kateri, I. Moustaki (eds.), *Trends and Challenges in Categorical Data Analysis*,
Statistics for Social and Behavioral Sciences,
https://doi.org/10.1007/978-3-031-31186-4_6

177

versions of the ACL models, and their key properties, including their equivalence to baseline-category logit (BCL) models, and discusses how that equivalence can be exploited for ML estimation. A real-data case study is used to illustrate that boundary estimates can also cause numerical and inferential issues for ACL models. Section 6.8 then details how and when the equivariance properties of mBR and mdBR, and implementations of the latter for BCL models, can be used for mBR and mdBR for the PO and NPO versions of ACL models. Finally, Sect. 6.9 details how the mBR estimates can be used for the explicit correction of the estimates of ordinal superiority summaries.

6.2 Boundary Estimates in Categorical Response Models

It is well known that ML estimation of regression models with categorical responses may result in estimates on the boundary of the parameter space. The data patterns that result in boundary estimates in general multinomial logistic regression models (also known as baseline category models; see [1, Section 7.1]) have been studied extensively and are completely characterized. For a range of binomial regression models, [38] proves that a certain degree of “overlap” on the data is a necessary and sufficient condition for the ML estimates to have finite values. Albert and Anderson [4] enrich the arguments in [38] generalizing the results in the case of baseline-category logit (BCL) models for nominal responses. In particular, [4] categorize the possible configurations for the sample points into complete separation, quasi-complete separation, and overlap, and then show that separation is necessary and sufficient for the ML estimate to have at least one infinite-valued component. Geometric representations of (quasi-)complete separation for binomial logistic regression –when the ACL and BCL models reduce to exactly the same form– are given in [4, Figure 1], and for multinomial responses in [27, Figure 1].

Example 6.1 (Separation in Logistic Regression) A simple illustration of a completely separated data set is shown in Fig. 6.1. The data consists of 100 realizations of two continuous covariates x_2 and x_3 , and a response y that ends up being 0 whenever $x_2 + 2x_3 > 0$. ML estimation of the logistic regression model with $\log\{\pi/(1 - \pi)\} = \beta_1 + \beta_2x_2 + \beta_3x_3$, where π is the probability of observing $y = 1$ given x , results in the estimated logistic discriminant line in Fig. 6.1, with the log-likelihood attaining its global maximum value of 0, and the fitted value 0 being assigned to all observations with $y = 0$, and 1 to the rest. The `detectseparation` R package [25] that implements the methods in the unpublished PhD thesis by Konis [17] can be used to show that the ML estimates of β_1 , β_2 and β_3 are $-\infty$, $+\infty$ and $+\infty$ respectively.

While there is no ambiguity in reporting infinite estimates, estimates on the boundary of the parameter space can (i) cause numerical instabilities to fitting procedures, (ii) lead to misleading output when estimation is based on iterative procedures with a stopping criterion, and more importantly, (iii) cause havoc

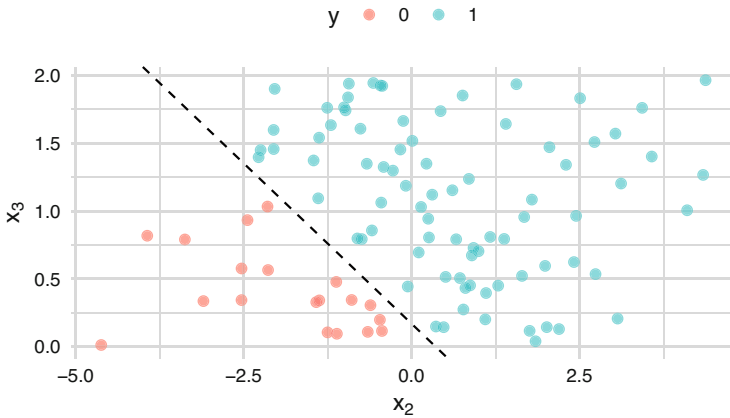


Fig. 6.1 The data described in Example 6.1. The dashed line is the line $0 = \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$, where the fitted probabilities are all 0.5

Table 6.1 ML, mBR and mdBR estimates for the logistic regression model in Example 6.1. The estimated standard errors (S.E.) are based on the expected information matrix at the estimates. The z -statistic is computed as estimate over estimated S.E., and the p -value is computed as $2 \min(\Phi(z), 1 - \Phi(z))$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard Normal distribution

Parameter	Estimate	Estimated S.E.	z -statistic	p -value
Maximum likelihood				
β_1	-22.397	13879.616	-0.002	0.999
β_2	62.578	20968.761	0.003	0.998
β_3	132.228	44964.541	0.003	0.998
Mean bias reduction				
β_1	-2.001	1.552	-1.289	0.197
β_2	5.266	1.997	2.637	0.008
β_3	11.166	3.984	2.803	0.005
Median bias reduction				
β_1	-2.583	1.984	-1.302	0.193
β_2	6.325	2.628	2.406	0.016
β_3	13.321	5.293	2.517	0.012

to asymptotic inferential procedures, and especially to the ones that depend on estimates of the standard error of the estimators (for example, Wald tests and related confidence intervals), oftentimes leading to wrong inferences. For example, the ML estimates in Table 6.1 have been obtained using the `glm()` function in R [37]. Despite the fact that the ML estimates for β_1 , β_2 and β_3 are in reality infinite, the stopping criteria of the fitting procedure that `glm()` implements are met for finite values of the parameters, which are returned. The reported estimated standard errors are also finite and substantially larger than the estimates. This results in small, in absolute value, z -statistics, and hence no evidence against the individual hypotheses $\beta_2 = 0$ and $\beta_3 = 0$; one would expect at least some evidence against the hypotheses given that the value of the response has been fully determined by the values of x_2 and x_3 .

One way to circumvent the numerical and inferential issues associated with boundary ML estimates is to replace ML with an alternative estimation method that (i) has comparable or sometimes better asymptotic properties than the ML estimator generally does, and (ii) tends to result or results in estimates away from the boundary of the parameter space. Popular examples of such alternative estimation methods are the mean bias-reducing adjusted score functions approach in [10], and the median bias-reducing adjusted score functions approach in [15], which we briefly review in Sects. 6.3, 6.4, and 6.6.

6.3 Mean Bias Reduction

Let $\ell(\theta)$ be the log-likelihood about a parameter vector θ with $\theta \in \mathfrak{R}^v$. Assuming that the model at hand is appropriate, then under fairly general regularity conditions about the model, the ML estimator $\hat{\theta} = \arg \max \ell(\theta)$ has mean bias $E_\theta(\hat{\theta} - \theta) = O(N^{-1})$, where N is a measure of information about θ , usually –but not necessarily– the sample size.

If $S(\theta) = \nabla \ell(\theta)$, [10] shows that we can define an alternative estimator θ^* with mean bias $E_\theta(\theta^* - \theta) = O(N^{-2})$, which is asymptotically smaller than the bias of $\hat{\theta}$, as the solution of

$$S(\theta) + A(\theta) = \mathbf{0}_v, \quad (6.1)$$

where

$$A_t(\theta) = \frac{1}{2} \text{trace} \left[i(\theta)^{-1} \{ P_t(\theta) + Q_t(\theta) \} \right] \quad (t = 1, \dots, v).$$

In the above expression, $P_t(\theta) = E_\theta(S(\theta)S(\theta)^\top S_t(\theta))$ and $Q_t(\theta) = -E_\theta(j(\theta)S_t(\theta))$, and $j(\theta) = -\nabla \nabla^\top \ell(\theta)$ and $i(\theta) = E_\theta(S(\theta)S(\theta)^\top)$ are the observed and expected information matrix about θ , respectively, with all expectations taken with respect to the model.

Mean bias reduction has been found to result in estimates away from the boundary of the parameter space in a range of categorical data models; see, for example, [10] and [14] for binomial logistic regression; [31] for the estimation of simple complementary log–log-models; [22, Section 6] for row-column association models; [6, 23], and [26, Section 6] for BCL models; and [19] for cumulative link models.

If θ is the canonical parameter of a full exponential family [see 33, Chapter 5], like in binomial and multinomial logistic regression, then $j(\theta) = i(\theta)$ and $j(\theta)$ does not depend on the stochastic part of the model. Hence, $Q_t(\theta) = \mathbf{0}_{v \times v}$, where $\mathbf{0}_{v \times v}$ is a $v \times v$ matrix of zeros, and some algebra [see 10, Section 3] gives that the

solution of the mean bias-reducing adjusted score equations (6.1) is equivalent to the maximization of the penalized log-likelihood

$$\ell(\theta) + \frac{1}{2} \log \det\{i(\theta)\}, \quad (6.2)$$

where the penalty is the logarithm of the Jeffreys prior. Recent work by [24] considers the impact of penalized likelihoods like (6.2) in the estimation of many well-used binomial-response generalized linear models, including logistic, probit, complementary log-log, and cauchit regression. Among other results, [24] prove that maximizing the likelihood after penalizing it by arbitrary positive powers of the Jeffreys prior always results in finite estimates, and derive the shrinkage directions implied by the penalty.

6.4 Median Bias Reduction

The median bias-reducing adjusted score functions of [15] is another method that has been found to result in finite estimates in extensive simulation studies with logistic regression and BCL models (see [26], Section 6) and with cumulative link models [12].

The ML estimator generally has median bias $P(\hat{\theta}_t \leq \theta_t) = 1/2 + O(N^{-1/2})$. [15] show that we can define an alternative estimator θ_t^\dagger with $P(\theta_t^\dagger \leq \theta_t) = 1/2 + O(N^{-3/2})$, which is asymptotically closer to $1/2$ than the median bias of $\hat{\theta}$, as the solution of

$$S(\theta) + A(\theta) - i(\theta)F(\theta) = 0_v. \quad (6.3)$$

In the above expression, $F_t(\theta) = [i(\theta)^{-1}]_t^\top \tilde{F}_t(\theta)$, with

$$\tilde{F}_{tu}(\theta) = \text{trace} \left[\tilde{i}_u(\theta) \left\{ \frac{1}{3} P_t(\theta) + \frac{1}{2} Q_t(\theta) \right\} \right] \quad (t = 1, \dots, v),$$

and $\tilde{i}_u(\theta) = [i(\theta)^{-1}]_u [i(\theta)^{-1}]_u^\top / [i(\theta)^{-1}]_{uu}$ ($u = 1, \dots, v$), where A_u and A_{tu} denote the u th column and (t, u) th element of a matrix A .

When $j(\theta) = i(\theta)$, expression (6.3) simplifies in a similar manner as expression (6.1) does. In fact, for one-parameter models ($v = 1$) that are exponential families in canonical parameterization, it can be shown that mdBR is formally equivalent to the maximization of $\ell(\theta) + \log \det\{i(\theta)\}/6$ (see [15], Section 2.1). However, mdBR has no penalized likelihood interpretation for $v > 1$.

6.5 Inference with Mean and Median Bias Reduction

According to the results in [10] and [15], both θ^* and θ^\dagger have the same asymptotic distribution as the ML estimator generally does, and hence are asymptotically efficient. Therefore, the distribution of those estimators for finite samples can be approximated by a Normal with mean θ and variance-covariance matrix $\{i(\theta)\}^{-1}$. The derivation of this result relies on the fact that both the adjustments $A(\theta)$ and $A(\theta) - i(\theta)F(\theta)$ to the score functions for mBR and mdBR in (6.1) and (6.3), respectively, are of order $O(1)$ as $N \rightarrow \infty$. Hence, the score function $S(\theta)$, which is $O_p(\sqrt{N})$, dominates the adjustments as information increases. The implication is that standard errors for the components of θ^* and θ^\dagger can be computed exactly as for the ML estimator, using the square roots of the diagonal elements of $\{i(\theta)\}^{-1}$ of $\{j(\theta)\}^{-1}$ at the estimates. Furthermore, first-order inferences, like standard Wald tests and Wald-type confidence intervals and regions are constructed in a plugin fashion, by replacing the ML estimates with the mBR or mdBR estimates in the usual procedures in standard software.

Example 6.2 (Separation in Logistic Regression (Continued.)) Continuing from Example 6.1, Table 6.1 provides the estimates of β_1 , β_2 and β_3 from mBR and mdBR. The estimates have been computed using the default arguments of the `brglm_fit()` method of the `brglm2` R package [21]. `brglm_fit()` implements a variant of the quasi-Fisher scoring procedure

$$\theta^{(k+1)} = \theta^{(k)} + \{i(\theta^{(k)})\}^{-1}U(\theta^{(k)}), \quad (6.4)$$

where $U(\theta) := S(\theta) + A(\theta)$ if the intention is to compute the mean BR estimates, and $U(\theta) := S(\theta) + A(\theta) - i(\theta)F(\theta)$ if the intention is to compute the mdBR estimates; see [26] for details on the quasi-Fisher iterations and the form of the adjusted scores for mBR and mdBR in generalized linear models. Convergence has been rapid and `brglm_fit()` reported no issues for either mBR or mdBR. Furthermore, the estimates and estimated standard errors appear to be finite. Note that the estimates and estimated standard errors from mBR are typically closer in absolute value to zero than those from mdBR. Importantly, the z -statistics for β_2 and β_3 are all away from zero, and, in contrast to ML, both mBR and mdBR suggest at least some evidence against the individual hypothesis $\beta_2 = 0$ and $\beta_3 = 0$, which agrees with the fact that the value of the response has been fully determined from the values of x_2 and x_3 .

6.6 Bias Reduction and Parameter Transformation

6.6.1 Maximum Likelihood Estimation and General Parameter Transformations

The ML estimator is equivariant in the sense that the ML estimator of $g(\theta)$ is exactly $g(\hat{\theta})$ for any one-to-one transformation $g(\cdot)$. Hence, there is no need to maximize the log-likelihood about $g(\theta)$ if the ML estimator of θ has already been computed. In contrast, the mBR and mdBR estimators are equivariant only for specific transformations $g(\cdot)$.

6.6.2 Mean Bias Reduction and Linear Parameter Transformations

The mBR estimator is equivariant under linear transformations for the parameters, in the sense that the mBR estimator of $C\theta$ for a known matrix C is exactly $C\theta^*$. The same is not true for the mdBR estimator.

For example, using Table 6.1, the mBR estimate of $\beta_2 - \beta_3$ in Example 6.2 is simply $5.266 - 11.166 = -5.9$. The mdBR estimate, however, is not $6.325 - 13.321 = 6.996$, but rather -7.227 , which is obtained by reparameterizing the model in terms of $\beta_2 - \beta_3$ and computing the mdBR estimate by solving (6.3) in the new parameterization.

6.6.3 Median Bias Reduction and Component-Wise Parameter Transformations

On the other hand, the mdBR estimator of $(g_1(\theta_1), \dots, g_v(\theta_v))^T$ is $(g_1(\theta_1^\dagger), \dots, g_v(\theta_v^\dagger))^T$ for any set of one-to-one functions $g_1(\cdot), \dots, g_v(\cdot)$. In other words, the mdBR estimator is equivariant under component-wise transformations. The same is not true for the mBR estimator. For example, the mdBR estimate of the odds-ratio $\exp(\beta_2)$ in Example 6.2 is exactly $\exp(6.325)$, but $\exp(5.266)$ is not an mBR estimate of $\exp(\beta_2)$.

6.6.4 Mean Bias Reduction and General Parameter Transformations

Di Caterina and Kosmidis [9] show that there is a simple way to derive the mean bias of $h(\theta^*)$ for any three-times differentiable function $h : C \rightarrow D$, with $C \subset \mathfrak{R}^p$ and $D \subset \mathfrak{R}$, where θ^* is an mBR estimator of θ with $O(N^{-2})$ bias. In particular, [9] show that the estimator $h(\theta^*)$ of $\zeta = h(\theta)$ has mean bias

$$E(h(\theta^*) - h(\theta)) = \frac{1}{2} \text{trace} \left\{ i(\theta)^{-1} \nabla \nabla^\top h(\theta) \right\} + O(N^{-2}), \quad (6.5)$$

where $\nabla \nabla^\top h(\theta)$ is the hessian of $h(\cdot)$ at θ . Note that for linear transformations, $\nabla \nabla^\top h(\theta) = 0_{v \times v}$, and hence $E(h(\theta^*) - h(\theta)) = O(N^{-2})$, which confirms the discussion in Sect. 6.6.2 that the mBR estimator is exactly equivariant for linear transformations of the parameters. The first term in the right-hand side of (6.5) can be evaluated at θ^* and be used to derive mean BR estimators of $h(\theta)$, based only on $\hat{\theta}^*$, $i(\hat{\theta}^*)$, and $\nabla \nabla^\top h(\theta^*)$. An obvious mean BR estimator resulting from (6.5) is $h(\theta^*) - \text{trace} \{ i(\theta^*)^{-1} \nabla \nabla^\top h(\theta^*) \} / 2$.

For example, consider the special case of estimation of the odds-ratio $\exp(\beta_j)$ in Example 6.1, which was estimated using the equivariance properties of mBR in Sect. 6.6.3. Expression (6.5) gives that the odds-ratio at the mBR estimator has

$$E(\exp(\beta_j^*)) = \exp(\beta_j) \left[1 + \frac{1}{2} v_{jj}(\theta) \right] + O(N^{-2}), \quad (6.6)$$

where $v_{jj}(\theta) = [i(\theta)^{-1}]_{jj}$. Hence, two mean BR estimators of $\zeta_j = \exp(\beta_j)$ with $O(N^{-2})$ bias are

$$\zeta_j^* = \exp(\beta_j^*) \left[1 - \frac{1}{2} v_{jj}(\theta^*) \right] \quad \text{and} \quad \zeta_j^{**} = \frac{\exp(\beta_j^*)}{1 + v_{jj}(\theta^*)/2},$$

arising from subtracting an estimate of the bias at $\theta := \theta^*$ from $\exp(\beta_j^*)$, and dividing $\exp(\beta_j^*)$ by the correction factor $1 + v_{jj}(\theta^*)/2$ from the right-hand side of (6.6), respectively. The estimator ζ_j^{**} for the odds-ratio ζ_j has the advantage of being always positive, while ζ_j^* takes negative values if $v_{jj}(\theta^*) > 2$. For example, to the accuracy reported in Table 6.1, $\zeta_2^* = \exp(5.266)(1 - 1.997^2/2) = -192.48$, which is clearly nonsensical as an odds-ratio estimate. In contrast, $\zeta_2^{**} = \exp(5.266)/(1 + 1.997^2/2) = 64.66$. The approximation $\exp\{v_{jj}(\theta)/2\} \approx 1 + v_{jj}(\theta)/2$ for small $v_{jj}(\theta)$ can be used to show that the mean BR estimator ζ_j^{**} closely relates to the mean BR estimator $\zeta_j^{***} = \exp\{\beta_j^* - v_{jj}(\theta^*)/2\}$ derived in [28].

The discussion in Sect. 6.5 implies that estimated standard errors for mBR estimators of transformed parameters constructed on the basis of (6.6) can be computed using the delta method, as for the ML estimator.

6.7 Adjacent-Categories Logit Models

6.7.1 Proportional and Non-proportional Odds Models

We now turn our attention in applying mBR and mdBR from Sects. 6.3 and 6.4 to ACL models.

Adjacent-categories logit models (see, for example, [2], Chapter 4 for an introduction) are a prominent family of regression models for ordinal responses, where the local odds ratios of consecutive categories of an ordinal response variable are linked with linear combinations of parameters and explanatory variables. Suppose that we observe realizations of n independent random vectors of frequencies Y_1, \dots, Y_n , where $Y_i = (Y_{i1}, \dots, Y_{ik})^\top$ has a k -category multinomial distribution with ordered categories $1 < 2 < \dots < k$, total $m_i = \sum_{j=1}^k Y_{ij}$ and probability vector $(\pi_1(x_i), \dots, \pi_k(x_i))^\top$ with $\sum_{j=1}^k \pi_j(x_i) = 1$, where $x_i = (x_{i1}, \dots, x_{ip})^\top$ is a p -vector of covariate values. An ACL model has

$$\log \frac{\pi_j(x)}{\pi_{j+1}(x)} = \eta_j(x) \quad (j = 1, \dots, k-1), \quad (6.7)$$

where $\eta_j(x)$ is typically a linear combination of unknown model parameters and a covariate vector x .

The specification of $\eta_j(x)$ results in ACL models with particular properties. The PO version of the ACL model has

$$\eta_j(x) = \alpha_j + \beta^\top x, \quad (6.8)$$

and $p + k - 1$ scalar model parameters $\theta = (\alpha_1, \dots, \alpha_{k-1}, \beta_1, \dots, \beta_p)^\top$. Straightforward algebra starting from (6.7) gives that

$$\frac{\pi_j(x_2)}{\pi_{j+1}(x_2)} = \exp\{\beta^\top (x_2 - x_1)\} \frac{\pi_j(x_1)}{\pi_{j+1}(x_1)} \quad \text{for any } x_1, x_2 \in \mathfrak{N}^p$$

and $j \in \{1, \dots, k-1\}$. (6.9)

As a result, adjacent-categories odds are indeed proportional with a constant of proportionality that does not depend on the category. The NPO version of the ACL model has

$$\eta_j(x) = \alpha_j + \beta_j^\top x, \quad (6.10)$$

with $(k-1)(p+1)$ scalar model parameters $\theta = (\alpha_1, \dots, \alpha_{k-1}, \beta_1^\top, \dots, \beta_{k-1}^\top)^\top$, where $\beta_j = (\beta_{j1}, \dots, \beta_{jp})^\top$. Figure 6.2 shows the adjacent-categories log-odds for two distinct categories under the PO and the NPO versions of the ACL model, for $x \in \mathfrak{N}$. Note that under the PO version of the model the log-odds for distinct

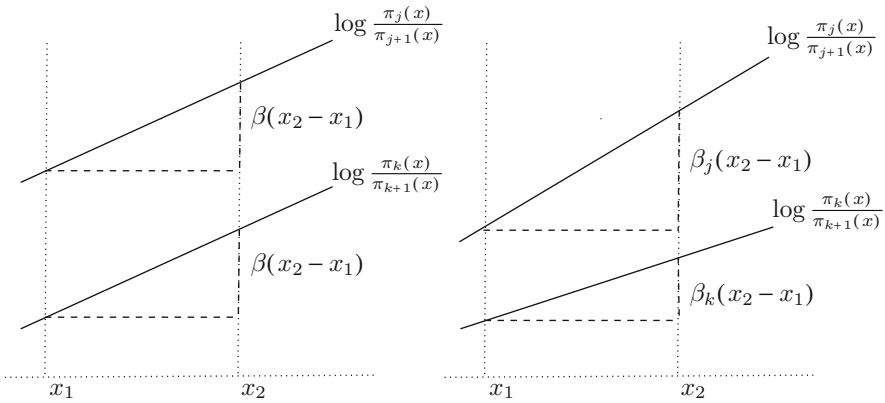


Fig. 6.2 The adjacent-categories log-odds for categories j and k , $j \neq k$, under the proportional odds (left) and the non-proportional odds (right) versions of the model, for $x \in \mathfrak{R}$. The probability for category j at covariate value x is denoted $\pi_j(x)$

categories are parallel lines, which in turn implies (6.9) for any pair of categories. On the other hand, under the NPO version of the model the log-odds are not parallel lines, so (6.9) is not generally satisfied.

The most general version of the ACL model is the partial proportional odds model with

$$\eta_j(x) = \alpha_j + \xi_j^\top x^{(np)} + \rho^\top x^{(p)},$$

where $x^{(np)}$ and $x^{(p)}$ are sub-vectors of x with distinct components characterizing the PO and NPO effects, respectively. All subsequent derivations, results, and discussions can be written in terms of the more general partial proportional odds version, and then PO and NPO can be presented as special cases. Nevertheless, we focus on the PO and NPO versions separately, to keep the notation concise, and because some of the following results are specific to PO and not to NPO.

Expressions (6.8) and (6.10) immediately imply that the ACL model provides valid category probabilities across the parameter space and regardless of whether the local odds $\pi_j(x)/\pi_{j+1}(x)$ are modelled as proportional or non-proportional. This is in contrast to other popular ordinal-response regression models, like cumulative-logit models [29], whose NPO versions [35] may provide invalid category probabilities in subsets of the parameter space and covariate space, and, hence, result in hard-to-circumvent issues with estimation, inference, and prediction.

6.7.2 Equivalence with Baseline-Category Logit Models

Writing $\log\{\pi_j(x)/\pi_k(x)\} = \sum_{l=j}^{k-1} \log\{\pi_l(x)/\pi_{l+1}(x)\}$, it is simple to show that both the PO and NPO versions of the ACL model for ordinal responses can be written as BCL models for nominal responses (see [2], Section 4.1) where the k category is used as reference.

In particular, the NPO version of the ACL model in (6.10) is equivalent to a BCL model with

$$\log \frac{\pi_j(x)}{\pi_k(x)} = \gamma_j + \delta_j^\top x \quad (j = 1, \dots, k-1), \quad (6.11)$$

where $\gamma_j = \sum_{l=j}^{k-1} \alpha_l$ and $\delta_j = \sum_{l=j}^{k-1} \beta_l$. The PO version of the ACL model in (6.8) is equivalent to a BCL model with

$$\log \frac{\pi_j(x)}{\pi_k(x)} = \gamma_j + (k-j)\zeta^\top x \quad (j = 1, \dots, k-1), \quad (6.12)$$

where $\gamma_j = \sum_{l=j}^{k-1} \alpha_l$ and $\beta = \zeta$.

6.7.3 Maximum Likelihood Estimation

A consequence of the equivalence between the BCL and ACL models is that we can estimate the latter using the ML estimates for the former. The equivariance of the maximum ML estimator under one-to-one transformations of the model parameters guarantees that after computing the ML estimates for the parameters of BCL model (6.11), the model parameters of the NPO version of the ACL can be estimated as $\hat{\alpha}_j = \hat{\gamma}_j - \hat{\gamma}_{j+1}$ and $\hat{\beta}_j = \hat{\delta}_j - \hat{\delta}_{j+1}$ ($j = 1, \dots, k-1$) with $\hat{\gamma}_k = 0$ and $\hat{\beta}_k = 0_p$, where 0_p is a p -vector of zeros. Correspondingly, once the ML estimates for the parameters of BCL model (6.12) have been obtained, the model parameters of the PO version of the ACL model can be estimated as $\hat{\alpha}_j = \hat{\gamma}_j - \hat{\gamma}_{j+1}$ and $\hat{\beta} = \hat{\zeta}$ ($j = 1, \dots, k-1$).

So, ML estimation of ACL models can be performed using ready ML implementations for fitting the BCL models (6.11) and (6.12), like the `multinom()` function of the `nnet` R package [39] that exploits the equivalence of BCL models with neural networks, and the `brmultinom()` function of the `brglm2` R package [21] that exploits the equivalence of BCL models with Poisson log-linear models.

6.7.4 Exponential Families

The BCL model is a full exponential family distribution with natural parameters γ_j and δ_j for the NPO version (6.11) of the ACL model, and γ_j and ζ for the PO version (6.12) of the ACL model ($j = 1, \dots, k - 1$). Hence, another consequence of the equivalence of ACL models to BCL models is that both the PO and NPO versions of the ACL model are full exponential families. Specifically, the sufficient statistics in the NPO parameterization are $\sum_{l=1}^j \sum_{i=1}^n y_{il}$ for α_j , $\sum_{l=1}^j \sum_{i=1}^n y_{il}x_i$ for β_j , and $\sum_{l=1}^j \sum_{i=1}^n y_{il}$ for α_j and $\sum_{j=1}^{k-1} \sum_{i=1}^n (k - j)y_{ij}x_i$ for β in the PO parameterization ($j = 1, \dots, k - 1$) (see, also, [2], Section 4.1).

6.7.5 Infinite Maximum Likelihood Estimates

As is the case for their equivalent BCL models, depending on the data configuration, the ML estimates of ACL models can have infinite components, resulting in issues for both iterative estimation procedures and for first-order inference about the parameters. In fact, infinite ML estimates for the PO and NPO versions of the ACL model result if, and only if, separation occurs for the equivalent BCL models. Example 6.3 below uses a real data set to illustrate the consequences that separation can have in the estimation of, and inference from, ACL models.

Example 6.3 (Infinite ML Estimates in ACL Models) The data set in Table 6.2 comes from [36] and concerns an experiment for investigating factors that affect the bitterness of white wine. There are two factors in the experiment, namely temperature at the time of crushing the grapes (with two levels, “cold” and “warm”) and contact of the juice with the skin (with two levels “Yes” and “No”). For each combination of factors two bottles were rated on their bitterness by a panel of 9 judges. The responses of the judges on the bitterness of the wine were taken on a continuous scale in the interval from 0 (“None”) to 100 (“Intense”) and then they were grouped correspondingly into 5 ordered categories, labelled as “1”, “2”, “3”, “4”, and “5”.

Figure 6.3 shows the empirical adjacent logits $\log\{(y_{ij} + 1/2)/(y_{ij+1} + 1/2)\}$ ($j = 1, \dots, 4$) for the bitterness rating for all combinations of temperature and contact. Note that 1/2 has been added to all frequencies as a means of getting

Table 6.2 The wine tasting data [36]

Temperature	Contact	Bitterness rating				
		1	2	3	4	5
Cold	No	4	9	5	0	0
Cold	Yes	1	7	8	2	0
Warm	No	0	5	8	3	2
Warm	Yes	0	1	5	7	5

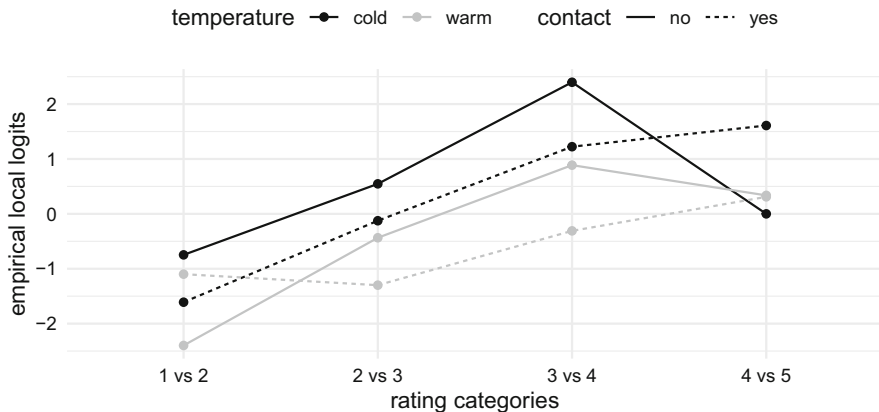


Fig. 6.3 The empirical adjacent logits $\log\{(y_{ij} + 1/2)/(y_{ij+1} + 1/2)\}$ ($j = 1, \dots, 4$) for the bitterness rating for all combinations of levels for temperature and contact

estimates of the adjacent-categories logits with second-order mean bias (see, for example, [13]), avoiding infinite estimates in the process.

There seems to be evidence that the adjacent logits for the combinations of temperature and contact are parallel (see, also, Fig. 6.2), or in other words, the adjacent odds ratios across temperature and/or contact levels do not depend on the rating. The latter hypothesis can be formally tested by estimating the NPO version of the ACL model

$$\log \frac{\pi_j(t, c)}{\pi_{j+1}(t, c)} = \alpha_j + \beta_{1j}t + \beta_{2j}c \quad (j = 1, \dots, 4), \tag{6.13}$$

where t is 1 if temperature is warm and 0 otherwise, c is 1 if contact is yes and 0 otherwise, and $\pi_j(t, c)$ is the probability of a bitterness rating j at t and c . The hypotheses of parallel adjacent logits can then be written in terms of the model parameters as $\beta_{11} = \dots = \beta_{14} = \beta_1$ and $\beta_{21} = \dots = \beta_{24} = \beta_2$, and tested using the value of the Wald statistic

$$W = \hat{\theta}^\top C^\top \left\{ C i(\hat{\theta})^{-1} C^\top \right\}^{-1} C \hat{\theta}, \tag{6.14}$$

where $\hat{\theta}$ is the ML estimate of $\theta = (\alpha_1, \dots, \alpha_4, \beta_{11}, \dots, \beta_{14}, \beta_{21}, \dots, \beta_{24})^\top$ for model (6.13), and $i(\theta)$ is the expected information matrix at θ . The contrast matrix C we use in (6.14) has the form

$$C = \begin{bmatrix} 0_{3 \times 4} & C_1 & 0_{3 \times 4} \\ 0_{3 \times 4} & 0_{3 \times 4} & C_1 \end{bmatrix} \quad \text{with} \quad C_1 = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix},$$

Table 6.3 Top: ML estimates and estimated standard errors (in parenthesis) from fitting the ACL model in (6.13) on the data in Table 6.2. The estimates are obtained using the `vglm()` function of the VGAM R package [42] version 1.1-5 with default converge criteria (`epsilon = 10-7` in `vglm.control()`). Bottom: ML estimates and estimated standard errors using stricter convergence criteria (`epsilon = 10-9` in `vglm.control()`). The estimated standard errors are computed as the square roots of the diagonal of the inverse of the expected information matrix at the ML estimates. The column $\ell(\hat{\theta})$ gives the maximized log-likelihood for each fit

epsilon	$\ell(\hat{\theta})$	Rating (j)	$\hat{\alpha}_j$	$\hat{\beta}_{1j}$	$\hat{\beta}_{2j}$
10^{-7}	-15.29	1	-0.83 (0.59)	-20.26 (10047.96)	-1.10 (1.21)
		2	0.67 (0.52)	-1.21 (0.66)	-0.87 (0.64)
		3	3.08 (1.05)	-1.98 (0.92)	-1.54 (0.83)
		4	20.22 (10732.18)	-19.89 (10732.18)	0.04 (1.08)
10^{-9}	-15.29	1	-0.83 (0.59)	-25.26 (122409.18)	-1.10 (1.21)
		2	0.67 (0.52)	-1.21 (0.66)	-0.87 (0.64)
		3	3.08 (1.05)	-1.98 (0.92)	-1.54 (0.83)
		4	25.22 (130748.2)	-24.89 (130748.24)	0.04 (1.08)

where $0_{a \times b}$ is an $a \times b$ matrix of zeros. General results about the limiting distribution of the ML estimator under mild regularity conditions (see, for example, [30, Section 7.1 and Section 7.2] and [8, Section 9.1]) can be used to show that the Wald statistic has asymptotically a χ^2_6 distribution.

Table 6.3 shows the ML estimates of the ACL model in (6.13), as computed using the `vglm()` function of the VGAM R package [42]. No warnings or errors were returned when fitting the model. As has been the case in the logistic regression model of Example 6.1, the estimates and estimated standard errors for α_4 , β_{11} and β_{14} are atypically large in absolute value. It is also clear that these estimates and estimated standard errors increase in absolute value as the convergence criteria get stricter, while the maximized log-likelihood value remains the same to the displayed accuracy.

These issues are not due to the implementation of the `vglm()` function; instead they are consequences of quasi-complete separation for this particular combination of data and model (6.13). The ML estimates $\hat{\alpha}_4$, $\hat{\beta}_{11}$ and $\hat{\beta}_{14}$ in Table 6.3 are formally ∞ , $-\infty$ and $-\infty$, the corresponding estimated standard errors are all ∞ , and the likelihood surface has an asymptote at -15.29 as α_4 , β_{11} and β_{14} diverge to ∞ , $-\infty$ and $-\infty$, respectively, along a ray in the parameter space.

Note here that the estimated standard errors appear to diverge faster than the ML estimates do as the convergence criteria get stricter. As a result, the typically reported Z -statistics for individual hypothesis tests about the parameters will tend to be spuriously small in absolute value regardless of the strength of the evidence against the hypotheses. Hence, the naive use of the computer output for inference about the parameters of ACL models is likely to lead to invalid conclusions when data separation occurs. More importantly, having estimates on the boundary of the parameter space violates the assumptions required for the asymptotic χ^2 distribution

of (6.14). Consequently, it is hard to justify the performance and validity of the Wald statistic in that case.

6.8 Mean and Median Bias Reduction for ACL Models

A consequence of the ACL models being full exponential family distributions (see Sect. 6.7.4) is that mean BR can be implemented by maximizing the penalized likelihood in (6.2). Nevertheless, as for ML, mean BR estimates for ACL models can be conveniently computed through a ready implementation for mean BR in BCL models coupled with the equivariance of the mean BR estimator under linear transformations (see Sect. 6.6.2).

Kosmidis and Firth [23] prove that the equivalence of BCL models and Poisson log-linear models (see, also, [34] and [5] for authoritative descriptions of that equivalence) extends to the mBR estimates, and describe a simple algorithm for mBR estimation of BCL models, each iteration of which consists of the following steps:

- P1 Rescale the Poisson means to match the observed multinomial totals.
- P2 Add half a leverage based on the rescaled means to the observed multinomial frequencies.
- P3 Estimate, using ML, the equivalent Poisson log-linear model to the adjusted frequencies.

Iteration stops when the differences between successive estimates or, alternatively, the mean BR adjusted scores in (6.1) are smaller than a pre-determined, small positive constant. An alternative criterion can be based on the change of the mean BR penalized likelihood (6.2) between successive iterations. mBR estimates for ACL models can then be computed as follows

- S1 Compute mBR estimates of the parameters γ_j and δ_j of the BCL model in (6.11) for the NPO version (or γ_j and ζ of the BCL model in (6.12) for the PO version) ($j = 1, \dots, q$) by iterating steps P1, P2, and P3.
- S2 Calculate the mBR estimates for the NPO version of the ACL model as $\alpha_j^* = \gamma_j^* - \gamma_{j+1}^*$ and $\beta_j^* = \delta_j^* - \delta_{j+1}^*$ (or $\alpha_j^* = \gamma_j^* - \gamma_{j+1}^*$ and $\beta^* = \zeta^*$ for the PO version) ($j = 1, \dots, q$), with $\gamma_k^* = 0$ and $\beta_k^* = 0_p$.

Implementation of mdBR for ACL models is not as direct as that of mBR. A maximum penalized likelihood interpretation of mdBR does not exist for general ACL models, like it does for mBR. Also, since contrasts of parameters are not component-wise transformations, algorithms for mdBR for BCL models (see [26, Section 6] for extensions of the results in [23]) can only be used to get mdBR estimates β^\dagger of β in the PO version of the ACL model. In other words, the estimates $\gamma_j^\dagger - \gamma_{j+1}^\dagger$ and $\beta_j^\dagger = \delta_j^\dagger - \delta_{j+1}^\dagger$ ($j = 1, \dots, q$) are not mdBR estimates, unless $k = 2$. Hence, for general ACL models, computing the mdBR estimates θ^\dagger must

Table 6.4 Mean BR estimates and estimated standard errors (in parenthesis) from fitting the ACL model in (6.13) on the data in Table 6.2. The estimates are obtained using the `brac1()` function of the `brglm2` R package [21] version 0.7.2 with default convergence criteria. The estimated standard errors are computed as the square roots of the diagonal of $i(\theta^*)^{-1}$

Rating (j)	α_j^*		β_{1j}^*		β_{2j}^*	
1	-0.76	(0.59)	-1.65	(1.60)	-0.82	(1.08)
2	0.62	(0.52)	-1.12	(0.66)	-0.80	(0.64)
3	2.73	(0.99)	-1.75	(0.87)	-1.38	(0.81)
4	1.53	(1.83)	-1.26	(1.68)	0.07	(1.03)

rely on implementing and solving the mBR adjusted score equations (6.3). That can certainly be done (using, for example, the quasi-Fisher scoring iteration (6.4)), with the only effort being in deriving $P_t(\theta)$ using the expressions for mBR in BCL models in [18, Appendix B.5].

Example 6.4 (Infinite ML Estimates in ACL Models (Continued.)) Table 6.4 gives the mBR estimates from fitting the ACL model in (6.13) on the data in Table 6.2. The mBR estimates are computed using the `brac1()` function of the `brglm2` R package, which implements mBR through the corresponding Poisson log-linear model, as detailed earlier. No convergence issues have been reported; the absolute values of the components of the adjusted score functions in (6.1) at the mBR estimates are all less than 10^{-6} , and all estimates and estimated standard errors remain unchanged to the reported accuracy as the convergence criteria get stricter.

The Wald statistic (6.14) when $\hat{\theta}$ is replaced by θ^* has value 1.067, which is small compared to the value of the 95% quantile of a χ_6^2 distribution (12.592), providing no evidence against the simpler PO model with $\beta_{11} = \dots = \beta_{14} = \beta_1$ and $\beta_{21} = \dots = \beta_{24} = \beta_2$.

Comparing the mBR estimates in Table 6.4 to the ML ones in Table 6.3, we notice that the mBR estimates are shrunken towards zero relative to ML ones. As a result, the fitted multinomial probabilities at the mBR estimates are closer to $(1/5, 1/5, 1/5, 1/5, 1/5)^\top$ than ones at the ML estimates. In other words, mBR shrinks the model towards equi-probability across observations. This is a generalization of the shrinkage effect of mBR we observed in Example 6.2 and that [24] study theoretically in the special case of logistic regression ($k = 2$).

It is interesting to note that the shrinkage direction of mBR in cumulative logit models for global cumulative odds [19] is rather different; the fitted multinomial probabilities at the mBR estimates for the PO version of the cumulative logit model would be closer to $(1/2, 0, 0, 0, 1/2)^\top$ than the ones at the ML estimates. In other words, mBR shrinks the cumulative logit model towards a logistic regression model for the end categories.

The ML, mBR, and mBR estimates for β_1 for the PO version of the ACL model are -1.69 , -1.61 , and -1.56 , respectively, with corresponding estimated standard errors 0.41 , 0.39 , and 0.38 . The respective estimates for β_2 are -0.96 , -0.92 , and -0.90 , respectively, with corresponding estimated standard errors 0.32 , 0.31 , and

0.31. The shrinkage towards equi-probability that mBR delivers is also apparent in the estimates for the PO version of the ACL model. As is the case in logistic regression, mdBR also tends to shrink estimates towards zero, but that shrinkage effect is less strong than from mBR.

6.9 Mean Bias Reduction of Ordinal Superiority Summaries

The mean BR estimates for ACL models can be used to get improved estimates of other model summaries by using the bias of transformations of the mean RB estimator in expression (6.5).

A prominent example of such a summary are the ordinal, model-based superiority measures for comparing distributions of two groups, adjusted for covariates that are introduced in [3]. In ordinal-response models with a latent variable interpretation, such as cumulative-link models [29], ordinal superiority measures can be defined directly on the latent scale, which results in exact (for probit, log-log, and complementary log-log link) or approximate expressions (for logit link) that are functions of only the coefficient of the indicator variable characterizing the two groups being compared. This fact has been exploited in [12], who used the equivariance properties of the mdBR estimator (see Sect. 6.6.3) to directly transform the mdBR estimates of the group indicator parameter to deliver mdBR estimates of ordinal superiority measures.

In more general models for ordinal responses that may also lack a latent variable interpretation (like ACL models), ordinal superiority measures are instead defined in terms of category probabilities that necessarily depend on all model parameters. Suppose that the covariate vector is $(w^\top, z)^\top$, where z is a group indicator variable taking value 0 for group 1 and value 1 for group 2, and denote by $\pi_j(w, 1)$ and $\pi_j(w, 0)$ ($j = 1, \dots, k$) the model-based probabilities of category j at covariate values w , for group 1 and group 2, respectively. The dependence of the probabilities on the model parameters has been suppressed here for notational convenience.

Agresti and Kateri [3] propose comparing the distribution of the ordinal response at group 1 to that at group 2, at covariate values w , through the ordinal superiority measure

$$\Delta(w; \theta) = \sum_{r>s} \pi_r(w, 1)\pi_s(w, 0) - \sum_{s>r} \pi_r(w, 1)\pi_s(w, 0). \quad (6.15)$$

If the two distributions are identical then $\Delta(w; \theta) = 0$. Positive values of $\Delta(w; \theta)$ indicate that for covariates w , it is more likely to observe higher response categories in group 1 than in group 2, and vice versa for negative values. A related ordinal superiority measure is

$$\gamma(w; \theta) = 2\Delta(w; \theta) - 1, \quad (6.16)$$

which takes values between 0 and 1, and is interpreted as the probability that the response category in group 1 is higher than the response category in group 2, while adjusting for covariates w (see [16], for details). In practice, the covariate setting w can be taken to be a representative value from a sample of covariate values w_1, \dots, w_n , e.g. $\bar{w} = \sum_{i=1}^n w_i/n$. Alternatively, if the sample of covariate values is representative of the population of interest then summary ordinal superiority measures can be defined as

$$\bar{\Delta}(\theta) = \frac{1}{n} \sum \Delta(w_i; \theta) \quad \text{and} \quad \bar{\gamma}(\theta) = \frac{1}{n} \sum \gamma(w_i; \theta). \quad (6.17)$$

Agresti and Kateri [3] propose estimating the ordinal superiority measures by replacing θ in expressions (6.15), (6.16), and (6.17) by the ML estimator $\hat{\theta}$, and use the delta method to construct inferences about those measures. Note here that because of the specific equivariance properties of the mBR and mdBR estimator (see Sect. 6.6), replacing θ by the mBR estimator θ^* or a mdBR estimator θ^\dagger does not, in general, result in mBR or mdBR estimators of the measures. In fact, despite it being the case that the resulting estimators will be consistent under the same conditions that their ML counterparts are, they may end up having much worse finite-sample mean and/or median bias properties than the ML version does.

mdBR estimators of (6.15), (6.16), and (6.17) are not easy to construct. In contrast, an easy-to-compute mBR estimator of $\Delta(w; \theta)$ and of the other ordinary superiority measures can be derived using expression (6.5). In particular, an mBR estimator of $\Delta(w; \theta)$ is

$$\Delta^*(w; \theta^*) = \Delta(w; \theta^*) - B^*(w; \theta^*).$$

where

$$B^*(w; \theta) = \frac{1}{2} \text{trace} \left\{ i(\theta)^{-1} \nabla \nabla^\top \Delta(w; \theta) \right\},$$

is the first term in the right-hand side of expression (6.5). Computing $\Delta^*(w; \theta^*)$ requires only the mBR estimator θ^* that can be obtained using the procedures in Sect. 6.8, the corresponding estimated category probabilities at $(w^\top, 1)$ and $(w^\top, 0)$, the matrix $i(\theta^*)^{-1}$, and the hessian $\nabla \nabla^\top \Delta(w; \theta^*)$. All these quantities, except $\nabla \nabla^\top \Delta(w; \theta^*)$, are readily available or can be readily computed once the model has been estimated using mBR, as is done, for example, in Sect. 6.3 for ACL models and in [19] for cumulative link models. For specific ordinal-response models, the hessian $\nabla \nabla^\top \Delta(w; \theta)$ can be analytically obtained with some algebraic effort. For example, if $\pi_j(w, z)$ is based on cumulative link models one can work with the expressions for the derivatives of $\Delta(w; \theta)$ in [3, Web appendix A]. Alternatively, a very accurate approximation of $\nabla \nabla^\top \Delta(w; \theta^*)$ can be obtained for general models using a ready implementation of $\Delta(w; \theta)$ and numerical differentiation routines, like the ones provided in the `numDeriv` R package [11]. This is the route that the `ordinal_superiority()` method of the `brglm2` R package takes.

Due to the equivariance properties of mBR estimation in Sect. 6.6.2 under linear transformations, mBR estimators of $\gamma(w; \theta)$, $\Delta^\dagger(\theta)$, and $\gamma^\dagger(\theta)$ are readily obtained by replacing $\Delta(w; \theta)$ by $\Delta^*(w; \theta^*)$ in expressions (6.16) and (6.17). Wald-type inferences about the mBR estimators of the ordinal superiority measures can be constructed as proposed in [3, Section 5], using the mBR estimates of the ordinal superiority measures along with estimated standard errors obtained using the delta method, based on $i(\theta^*)$, and numerical gradients.

Example 6.5 (mBR for Ordinal Superiority Measures from ACL Model) In order to assess the finite sample properties of the mBR estimator of ordinal superiority scores in ACL models, we consider the example in [7, Section 4.3], where the bitterness ratings “2”, “3”, and “4” in Table 6.2 are merged into a single rating “2–4”. Like the PO version of the cumulative logit model [see 7, Section 4.8], the ML estimates for α_{2-4} and β_1 for the PO version of the ACL model (6.13) are $+\infty$ and $-\infty$ respectively. The mBR estimates of $\theta = (\alpha_1, \alpha_{2-4}, \beta_1, \beta_2)^\top$, on the other hand, take the finite values $\theta^* = (-1.247, 5.331, -3.291, -1.181)^\top$. If $\gamma(w, \theta)$ is the ordinal superiority measure for temperature setting w ($w = 0$ for cold and $w = 1$ for warm), and z indicates contact ($z = 1$) or not ($z = 0$) of the juice with the skin, then $\gamma(0, \theta^*) = 0.594$ and $\gamma(1, \theta^*) = 0.575$, indicating that there is almost 60% chance of higher bitterness ratings when there is contact of the juice with the skin.

We simulate 10,000 samples from the PO version of the ACL model at $\bar{\theta}$, and we compute $\gamma(w, \hat{\theta})$ and $\gamma^*(w, \theta^*)$ for each sample. The simulation-based estimates of the finite-sample relative biases of $\gamma(w, \hat{\theta})$ are 0.84% and 1.56% for $w = 0$ and $w = 1$, respectively. As expected, the mBR version $\gamma^*(w, \theta^*)$ is found to have smaller finite-sample relative biases at 0.13% and -0.02% for $w = 0$ and $w = 1$, respectively. The corresponding percentages of underestimation are 48.48% and 44.69% for $\gamma(w, \hat{\theta})$, and 52.12% and 51.14% for $\gamma^*(w, \theta^*)$. Hence, in this case, mBR also results in improvements in median bias. Finally, both estimators appear to perform satisfactorily in terms of Wald-type inferences based on them. The coverage probability of the nominally 95% Wald-type confidence intervals based on $\gamma(w, \hat{\theta})$ are 94.8% ($w = 0$) and 94.6% ($w = 1$), and 94.7% ($w = 0$) and 95.1% ($w = 1$) for those based on $\gamma^*(w, \theta^*)$.

6.10 Supplementary Material

The supplementary material consists of three scripts that replicate all the numerical results and graphics reported in the paper, and is available at https://ikosmidis.com/files/bracl_supplementary_v0.2.zip. The results are exactly reproducible in R version 4.1.2, and with the following packages: VGAM version 1.1-5 [42], tibble 3.1.6 [32], dplyr 1.0.7 [41], ggplot2 3.3.5 [40], colorspace 2.0-2 [43], and ordinal 2019.12-10 [7], brglm2 0.8.2 [21], enrichwith 0.3.1 [20], and detectseparation 0.2 [25].

Acknowledgments The author greatly appreciates the constructive discussions with Alan Agresti and Anestis Touloumis during the Challenges for Categorical Data Analysis 2018 Workshop in Aachen University on the equivalence between ACL and BCL models, which informed this work. Ioannis Kosmidis has been partially supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1.

References

1. Agresti, A.: *Categorical Data Analysis*, 2nd edn. Wiley-Interscience, New York (2002)
2. Agresti, A.: *Analysis of Ordinal Categorical Data*, 2nd edn. Wiley, Hoboken (2010)
3. Agresti, A., Kateri, M.: Ordinal probability effect measures for group comparisons in multinomial cumulative link models. *Biometrics* **73**(1), 214–219 (2017)
4. Albert, A., Andershan, J.: On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **71**(1), 1–10 (1984)
5. Baker, S.G.: The multinomial-Poisson transformation. *J. R. Stat. Soc. Ser. D (Statistician)* **43**(4), 495 (1994)
6. Bull, S.B., Mak, C., Greenwood, C.M.T.: A modified score function estimator for multinomial logistic regression in small samples. *Comput. Stat. Data Anal.* **39**, 57–74 (2002)
7. Christensen, R.H.B.: *Ordinal—regression models for ordinal data*. R package version 2019.12-10 (2019). <https://CRAN.R-project.org/package=ordinal>
8. Cox, D.R., Hinkley, D.V.: *Theoretical Statistics*. Chapman & Hall Ltd., London (1974)
9. Di Caterina, C., Kosmidis, I.: Location-adjusted Wald statistics for scalar parameters. *Comput. Stat. Data Anal.* **138**, 126–142 (2019)
10. Firth, D.: Bias reduction of maximum likelihood estimates. *Biometrika* **80**(1), 27–38 (1993)
11. Gilbert, P., Varadhan, R.: *numDeriv: Accurate Numerical Derivatives*. R package version 2016.8-1.1 (2019)
12. Gioia, V., Kenne Pagui, E.C., Salvan, A.: Median bias reduction in cumulative link models. *Commun. Stat.* **52**, 1–17 (2021)
13. Haldane, J.: The estimation of the logarithm of a ratio of frequencies. *Ann. Hum. Genet.* **20**, 309–311 (1955)
14. Heinze, G., Schemper, M.: A solution to the problem of separation in logistic regression. *Stat. Med.* **21**(16), 2409–2419 (2002)
15. Kenne Pagui, E.C., Salvan, A., Sartori, N.: Median bias reduction of maximum likelihood estimates. *Biometrika* **104**(4), 923–938 (2017)
16. Klotz, J.H.: The Wilcoxon, ties, and the computer. *J. Am. Stat. Assoc.* **61**(315), 772–787 (1966)
17. Konis, K.: *Linear programming algorithms for detecting separated data in binary logistic regression models*. Ph. D. thesis, University of Oxford (2007)
18. Kosmidis, I.: *Bias reduction in exponential family nonlinear models*. Ph. D. thesis, University of Warwick (2007)
19. Kosmidis, I.: Improved estimation in cumulative link models. *J. R. Stat. Soc. Ser. B (Stat Methodol.)* **76**(1), 169–196 (2014)
20. Kosmidis, I.: *enrichwith: Methods to enrich list-like R objects with extra components*. R package version 0.3.1 (2020)
21. Kosmidis, I.: *brglm2: Bias reduction in generalized linear models*. R package version 0.7.2 (2021)
22. Kosmidis, I., Firth, D.: Bias reduction in exponential family nonlinear models. *Biometrika* **96**(4), 793–804 (2009)
23. Kosmidis, I., Firth, D.: Multinomial logit bias reduction via the poisson log-linear model. *Biometrika* **98**(3), 755–759 (2011)

24. Kosmidis, I., Firth, D.: Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. *Biometrika* **108**(1), 71–82 (2021)
25. Kosmidis, I., Schumacher, D.: detectseparation: Detect and Check for Separation and Infinite Maximum Likelihood Estimates. R package version 0.2 (2021)
26. Kosmidis, I., Kenne Pagui, E.C., Sartori, N.: Mean and median bias reduction in generalized linear models. *Stat. Comput.* **30**, 43–59 (2020)
27. Lesaffre, E., Albert, A.: Partial separation in logistic discrimination. *J. R. Stat. Soc. Ser. B (Methodol.)* **51**(1), 109–116 (1989)
28. Lyles, R.H., Guo, Y., Greenland, S.: Reducing bias and mean squared error associated with regression-based odds ratio estimators. *J. Stat. Plan. Infer.* **142**(12), 3235–3241 (2012)
29. McCullagh, P.: Regression models for ordinal data. *J. R. Stat. Soc. Ser. B (Methodol.)* **42**, 109–142 (1980)
30. McCullagh, P.: *Tensor Methods in Statistics*, 2nd edn. Dover Publications, Mineola (2018)
31. Mehrabi, Y., Matthews, J.N.S.: Likelihood-based methods for bias reduction in limiting dilution assays. *Biometrics* **51**, 1543–1549 (1995)
32. Müller, K., Wickham, H.: *tibble: Simple Data Frames*. R package version 3.1.6 (2021)
33. Pace, L., Salvan, A.: *Principles of statistical inference from a Neo-Fisherian perspective*. World Scientific, Singapore (1997)
34. Palmgren, J.: The Fisher information matrix for log linear models arguing conditionally on observed explanatory variables. *Biometrika* **68**(2), 563 (1981)
35. Peterson, B., Harrell, J., Frank, E.: Partial proportional odds models for ordinal response variables. *Appl. Stat.* **39**, 205–217 (1990)
36. Randall, J.H.: The analysis of sensory data by generalised linear model. *Biom. J.* **7**, 781–793 (1989)
37. R Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna (2021)
38. Silvapulle, M.J.: On the existence of maximum likelihood estimators for the binomial response models. *J. R. Stat. Soc. Ser. B (Methodol.)* **43**(3), 310–313 (1981)
39. Venables, W.N., Ripley, B.D.: *Modern Applied Statistics with S*, 4th edn. Springer, New York (2002). ISBN 0-387-95457-0
40. Wickham, H.: *ggplot2: Elegant Graphics for Data Analysis*. Springer, New York (2016)
41. Wickham, H., François, R., Henry, L., Müller, K.: *dplyr: A Grammar of Data Manipulation*. R package version 1.0.7 (2021)
42. Yee, T.W.: *VGAM: Vector Generalized Linear and Additive Models*. R package version 1.1-5 (2021)
43. Zeileis, A., Fisher, J.C., Hornik, K., Ihaka, R., McWhite, C.D., Murrell, P., Stauffer, R., Wilke, C.O.: *colorspace: A toolbox for manipulating and assessing colors and palettes*. *J. Stat. Softw.* **96**(1), 1–49 (2020)

Chapter 7

Regularization and Predictor Selection for Ordinal and Categorical Data



Jan Gertheiss and Gerhard Tutz

7.1 Introduction

In regression modeling, categorical variables can be challenging. Categorical predictors, for instance, are typically included in the model in the form of dummy variables that encode the occurrence of specific categories. That means a categorical predictor with k categories contributes $k - 1$ parameters. If the numbers of categories for single variables are large and/or several categorical predictors are available, the number of parameters becomes very large, and inference can be affected strongly. Also, an uneven distribution of the observations across the categories of the covariate can result in poor estimates. In classical linear regression, for instance, dummy coefficients referring to categories with only a small number of observations typically have large variance. In binary regression, it even happens from time to time that only observations from one response class are found within a specific category of an explanatory variable, which leads to inflated dummy coefficients tending towards $\pm\infty$. Therefore, in typical applications only a few categorical predictors with a manageable number of categories are used, often obtained after fusing categories. If those categorical variables are ordinal, an alternative, and quite popular approach in applied statistics, is to treat those variables as metric and use the classical, or corresponding generalized linear model.

A typical example for ordinal variables as described above is the the ICF (International Classification of Functioning, Disability and Health) [74], which can

J. Gertheiss (✉)
Helmut Schmidt University, Hamburg, Germany
e-mail: jan.gertheiss@hsu-hh.de

G. Tutz
Ludwig Maximilians University, Munich, Germany
e-mail: gerhard.tutz@stat.uni-muenchen.de

be used by health professionals to document the health and functioning of patients in a standardized form. For instance, the patient's ability to walk can be assessed on a five-point (ordinal) scale ranging from 'no problem' to 'complete problem'. Usually, however, not only one aspect of a person's health will be evaluated but one may easily end up with dozens of corresponding variables. In addition, extreme categorizations will (hopefully) occur less often than those for milder problems. Finally, not only five-point but also nine-point scales exist within the ICF; for details, see Sect. 7.4, where the so-called ICF Core Set for chronic widespread pain will be considered. At this point, we can note that, if using ICF assessment as independent variables in a regression model, we are usually faced with a situation as sketched above: a relatively large number of (ordinal) covariates, with at least some having a relatively large number of levels, and potentially uneven distribution of the observations across levels. For the adequate handling of such settings, advanced and flexible methods that provide a sparse representation of categorical variables are needed, including, but not limited to, the selection of relevant variables. One approach to obtain sparse models uses regularization methods based on adequate penalty terms that reduce the number of effective parameters. Alternatives include Bayesian methods and both supervised and unsupervised machine learning tools.

Within a Bayesian framework, model selection may be done by using the spike and slab distribution, which has been propagated as a modeling tool in structured additive regression [53]. Specifically, sparse Bayesian modeling with nominal and ordinal categorical predictors can be done by placing a spike and slab prior on appropriate differences of regression coefficients [49]. As an alternative, (model-based) clustering may be applied on the categories' effects, which can also be done in a Bayesian framework [36].

Penalization methods may become computationally demanding if the number of categories is very large. Then, an alternative way to fuse categories and select variables is to use recursive partitioning methods, also called trees. Tree-type methods for structuring categorical predictors that also work in high dimensions have been considered by Tutz & Berger [64]. Also, boosting methods can be used for both model fitting and selection, in particular with ordinal covariates [5, 23, 30, 39].

In this chapter we will focus on penalty-based regularization for categorical, in particular ordinal, predictors, but will also make some comments on regularization for nominal covariates and categorical response models in Sect. 7.5. The approaches are embedded within the general class of (generalized) additive regression models.

Given a response y with distribution from a simple exponential family, and a set of covariates x_1, \dots, x_p , a generalized additive model has the form [29]:

$$\eta = \alpha + f_1(x_1) + \dots + f_p(x_p), \quad \mu = h(\eta), \quad (7.1)$$

where μ is the (conditional) mean of y given the covariates, h is a (known) response function, and η is the equivalent of the linear predictor in generalized linear models [41, 45]. This differs from a generalized linear model in that *non-linear* functions f_j , $j = 1, \dots, p$, are allowed in η , but still the structure of η

is additive. Of course, if the functions f_j are restricted to be linear, a generalized linear model is obtained as a special case.

Generalized additive models are, in particular, useful to obtain more flexible forms of predictor terms for quantitative, continuous explanatory variables. For those kinds of predictors, it is typically assumed that functions f_j are reasonably smooth; and one way to fit such models, as for instance implemented in the popular R [51] package `mgcv` [78], is to specify a set of basis functions for each predictor. That means, one assumes that

$$f_j(x) = \sum_{r=1}^{q_j} \beta_{jr} B_{jr}(x), \quad (7.2)$$

with $B_{j1}(x), \dots, B_{jq_j}(x)$ being the set of basis functions chosen for function f_j , and $\beta_{j1}, \dots, \beta_{jq_j}$ the corresponding basis coefficients; q_j is the number of basis functions chosen for predictor x_j . By fitting the basis coefficients, the function f_j is fitted implicitly. For instance, a popular choice in case of a continuous x is the so-called B-spline basis, leading to f_j being modeled as a spline function; see, e.g., [11, 13, 14] for details on (B-)splines. The big advantage of the basis functions approach is that after plugging-in the observed data x_{ij} , $i = 1, \dots, n$, $j = 1, \dots, p$, the vector $\vec{f}_j = (f_j(x_{1j}), \dots, f_j(x_{nj}))^\top$ can be written as $\vec{B}_j \boldsymbol{\beta}_j$, with matrix $(\vec{B}_j)_{ir} = B_{jr}(x_{ij})$ and vector $\boldsymbol{\beta}_j = (\beta_{j1}, \dots, \beta_{jq_j})^\top$. So model (7.1) can be written as a (generalized) linear model with coefficients β_{jr} , $j = 1, \dots, p$, $r = 1, \dots, q_j$, and basis coefficients can be fitted accordingly, for instance by maximum likelihood (ML).

However, basis functions do not need to be smooth, and also categorical predictors can be modeled within the framework of additive models. Suppose one has a categorical predictor x_j with levels $1, \dots, k_j$. Then there is a somewhat natural basis: the basis of (dummy) functions often referred to as dummy variables ($l = 1, \dots, k_j$)

$$B_{jl}(x) = \begin{cases} 1 & \text{if } x=l, \\ 0 & \text{otherwise.} \end{cases} \quad (7.3)$$

Since it is known that x_j can only take values $1, \dots, k_j$, we do not need to think about the type and number of basis functions, placing of knots, etc., as one usually has to do with continuous covariates. For means of identifiability, however, some linear restrictions need to be placed on the basis/dummy coefficients β_{jl} . Typically, this is done by specifying a so-called reference category, e.g., category 1 for categorical predictor x_j , and setting the corresponding $\beta_{j1} = 0$. However, one may also set $\sum_l \beta_{jl} = 0$, which is also known as ‘effect coding’. In case of a continuous x_j a popular constraint is

$$\sum_{i=1}^n f_j(x_{ij}) = 0 \text{ for all } j, \quad (7.4)$$

which translates into $\sum_l n_{jl} \beta_{jl} = 0$ for categorical predictor x_j , with n_{jr} being the number of samples with level r being observed for x_j . Since the latter is the typical constraint in generalized additive models, also used in `mgcv` [78], we will use (7.4) here as well. After having fit the functions/basis coefficients, however, one can switch between constraints easily, because switching between the constraints mentioned above for some f_j yields an equivalent model as it simply means a vertical shift of the entire function f_j and a corresponding change in the constant α in (7.1). However, it should be noted that by changing the constraint, the interpretation of β -coefficients changes, too. In the case of standard dummy coding with reference category, elements of β_j refer to differences to the reference category. In the cases of the other two constraints given above, those β s give differences to some, hypothetical, ‘mean category’. This also needs to be taken into account when interpreting measures of uncertainty such as confidence intervals.

The rest of this chapter is organized as follows. In Sect. 7.2, we will discuss regularization for ordinal covariates in the framework of generalized additive models. We will present different types of penalties for smoothing, fusion of categories, and/or variable selection. In addition, we will sketch some tools for statistical inference that are available in generalized additive models employing quadratic (smoothing) penalties and show how those can be used for ordinal predictors with corresponding penalties as well. Based on the latter, we will present stepwise selection, in particular forward selection, as an alternative, and more classical, method for variable/model selection with ordinal predictors. We will compare those more classical procedures to L_1 -type regularization in numerical experiments in Sect. 7.3. As already mentioned, Sect. 7.4 presents a case study on real-world data (ICF Core Set), comparing forward model selection for ordinal predictors in a generalized additive modeling framework to results obtained earlier [23] when using L_1 -regularization. Section 7.5 discusses L_1 -type regularization for nominally scaled predictors and categorical response, and Sect. 7.6 concludes.

7.2 Regularization for Ordinal Covariates

In generalized additive models with continuous covariates, the problem with the basis functions approach is that typically a rather large number of basis functions needs to be chosen to be sufficiently flexible with respect to the type of functions that can be fitted. With a large basis, however, the number of basis coefficients to be fitted becomes large, too. As a consequence, resulting functions tend to be wiggly and thus hard to interpret. Therefore, a penalty term $J_j(\beta_j)$ is typically added for each covariate x_j , penalizing wiggly basis coefficients and thus wiggly functions f_j . Instead of maximizing the usual log-likelihood $l(\beta)$ one maximizes the penalized log-likelihood

$$l_p(\beta) = l(\beta) - \sum_{j=1}^p \lambda_j J_j(\beta_j),$$

where the parameters λ_j , $j = 1, \dots, p$, are non-negative, variable-specific smoothing parameters, $\boldsymbol{\beta}_j$ contains all coefficients that belong to covariate x_j (as defined above), and $\boldsymbol{\beta}$ is a vector comprising all β -coefficients. For B-splines with equally spaced knots, widely used penalties are quadratic difference penalties

$$J_j(\boldsymbol{\beta}_j) = \sum_{s=d+1}^{k_j} (\Delta^d \beta_{js})^2,$$

where Δ is the difference operator, operating on adjacent B-spline coefficients, that is, $\Delta \beta_{js} = \beta_{js} - \beta_{j,s-1}$, $\Delta^2 \beta_{js} = \Delta(\beta_{js} - \beta_{j,s-1}) = \beta_{j,s} - 2\beta_{j,s-1} + \beta_{j,s-2}$, and so on. The method is referred to as *P-splines* (for penalized splines). An alternative form of the penalties uses the representation $\sum_{s=d+1}^{k_j} (\Delta^d \beta_{js})^2 = \boldsymbol{\beta}_j^\top \mathbf{K}_d \boldsymbol{\beta}_j$, where the corresponding matrix \mathbf{K}_d has a banded structure and gives the differences in matrix form; for details see [14, 60]. B/P-splines have the advantage that in the limit, with strong smoothing, a polynomial is fitted. If a penalty of order d is used and the degree of the B-spline is higher than d , for large values of λ_j the fit of the function $f_j(\cdot)$ will approach a polynomial of degree $d - 1$. Other penalties explicitly penalize f_j 's curvature as given by the second derivative. A P-spline with second-order penalty can be considered a discrete approximation.

7.2.1 Quadratic Smoothing Penalties for Ordinal Predictors

Quadratic smoothing penalties as described above are very common when fitting generalized additive models with continuous covariates. For instance, they are implemented in the popular R add-on package `mgcv` [78].

If the predictor is categorical, the number of basis/dummy coefficients to be fitted can easily become large as well (as already described above), and fitted coefficients tend to have high variance and be wiggly across categories. Consequently, one may use penalized fitting analogously to the case of continuous covariates. In particular with ordinal predictors, the approach is straightforward.

7.2.1.1 Basic Ideas

If a (quadratic) difference penalty is put on the (dummy) coefficients with basis (7.3), this gives exactly the smoothing penalty for ordinal predictors proposed earlier [21, 65, 66]. More precisely, with β_{jl} , $l = 1, \dots, k_j$, denoting the dummy coefficient of level l of predictor x_j , the penalty primarily used is the first-order penalty

$$J_j(\boldsymbol{\beta}_j) = \sum_{l=2}^{k_j} (\beta_{jl} - \beta_{j,l-1})^2, \quad (7.5)$$

with $\boldsymbol{\beta}_j = (\beta_{j1}, \dots, \beta_{jk_j})^\top$. Alternatively, however, the second-order penalty

$$J_j(\boldsymbol{\beta}_j) = \sum_{l=2}^{k_j-1} (\beta_{j,l+1} - 2\beta_{jl} + \beta_{j,l-1})^2, \quad (7.6)$$

penalizing deviations from linearity can be used as well [20]. The strength of the penalty is determined by parameter λ_j , which may be different for different predictors x_j . In case of $\lambda_j = 0$, for both penalties (7.5) and (7.6), the usual maximum likelihood estimates are obtained. If $\lambda_j \rightarrow \infty$ in case of penalty (7.5), the fit $f_j(x) = 0$ for all $x \in \{1, \dots, k_j\}$ is obtained because of the constraint (no matter which one is chosen from the options given above). If penalty (7.6) is chosen, large λ_j leads to a function f_j being linear in the class labels $1, \dots, k_j$. One of the benefits of considering ordinal predictors along with penalties (7.5) and (7.6) in the framework of generalized additive models is that after implementing basis (7.3) in the appropriate way, `gam()` from `mgcv` can be used for model fitting [24].

7.2.1.2 Further Statistical Inference in Generalized Additive Models

Besides model fitting, considering ordinal predictors in the framework of generalized additive models has additional advantages, since we can make use of various tools originally developed for continuous covariates. As also described in [24], this particularly refers to:

- estimation of penalty/smoothing parameters,
- further statistical inference, such as confidence intervals, and
- checking significance of smooth terms.

A prerequisite for at least some of those tools is to rewrite the model with the quadratic smoothing penalty as a (generalized) linear mixed model. Starting with penalty (7.5), we may rewrite our model in terms of $\tilde{\boldsymbol{\beta}}_j = \tilde{\mathbf{D}}_1 \boldsymbol{\beta}_j$ with

$$\tilde{\mathbf{D}}_1 = \begin{bmatrix} 1 & 0 & \dots & 0 \\ & \mathbf{D}_1 & & \end{bmatrix}, \text{ and } \mathbf{D}_1 = \begin{pmatrix} -1 & 1 & 0 & 0 & \dots \\ 0 & -1 & 1 & 0 & \dots \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}.$$

Let us denote the subvector consisting of the last $k_j - 1$ elements of $\tilde{\boldsymbol{\beta}}_j$ by $\mathbf{u}_j = (u_{j1}, \dots, u_{j,k_j-1})^\top$, such that $u_{jl} = \beta_{j,l+1} - \beta_{jl}$ and $\mathbf{u}_j = \mathbf{D}_1 \boldsymbol{\beta}_j$, with \mathbf{D}_1 from above (and $\mathbf{D}_1^\top \mathbf{D}_1 = \mathbf{K}_1$ from page 202). The entries of \mathbf{u}_j are now interpreted as iid random effects with $u_{jl} \sim N(0, \tau_j^2)$; compare [18, 20, 57]. Then, for given variance parameters τ_j^2 , $j = 1, \dots, p$ (note, the random effects' variance may vary between covariates), maximizing the log-likelihood over $\tilde{\boldsymbol{\beta}}_j$ yields estimates that are equivalent to the smoothed dummies obtained via penalty (7.5), with a one-to-one

correspondence of penalty parameter λ_j and variance parameter τ_j . Alternatively, smoothed dummy coefficients can be derived in a Bayesian framework (as the mode of the posterior density) by putting a Gaussian random walk prior (with prior variance τ_j^2) on the dummy coefficients [21]. Analogously, penalty (7.6) can be derived/interpreted in a mixed model/Bayesian framework. For that purpose, we replace $\tilde{\mathbf{D}}_1$ above by $\tilde{\mathbf{D}}_2$ with

$$\tilde{\mathbf{D}}_2 = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ & & \mathbf{D}_2 & & \end{bmatrix}, \text{ and } \mathbf{D}_2 = \begin{pmatrix} 1 & -2 & 1 & 0 & 0 & \dots \\ 0 & 1 & -2 & 1 & 0 & \dots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}.$$

Then, the last $k_j - 2$ elements of $\tilde{\boldsymbol{\beta}}_j$ are denoted by $\mathbf{u}_j = (u_{j1}, \dots, u_{j,k_j-2})^\top$, such that $u_{jl} = \beta_{j,l+2} - 2\beta_{j,l+1} + \beta_{jl}$, and as before $u_{jl} \sim N(0, \tau_j^2)$ are interpreted as iid random effects; compare [20, 78].

Estimation of Penalty Parameters Both approaches, the mixed model and Bayesian interpretation, can be used for determining the variance components τ_j^2 , and thereby the penalty parameters λ_j . In theory, we may integrate out the random effects from the joint density of the response and random effects, giving the marginal likelihood of the fixed effects and the variance parameters. Maximizing this likelihood leads to ML estimates of the fixed effects and variance parameters. In generalized linear mixed models, however, calculating the integral analytically is often not possible, and also numerically demanding. The standard approach is the so-called *Laplace approximation* [4], which essentially cycles through the penalized log-likelihood given above and plugging-in the corresponding estimates of regression/basis coefficients to obtain an approximate profile likelihood for the variance parameters that can be maximized. In the Bayesian framework, the smoothed dummies are then interpreted as an *empirical Bayes* estimator, since τ_j^2 are estimated from the data. In a fully Bayesian approach, we could choose a hyperprior, e.g., an Inverse Gamma, for τ_j^2 and apply Markov Chain Monte Carlo (MCMC); but we won't follow this path here.

A problem with ML estimation of variance parameters is that those estimates are typically biased downwards, that is, the true variance tends to be underestimated, in particular if the number of fixed effects is large. As an alternative to ML that reduces this bias, so-called *restricted* maximum likelihood (REML) estimation has been proposed, which can be motivated in different ways [15, 27, 28, 33, 48]. Eventually, in the linear mixed model, the (profile) log-likelihood is (additively) corrected such that the number/structure of fixed effects is taken into account. In generalized mixed models, this can be done analogously within Laplace approximation. It should be noted that REML cannot be used to compare models with different fixed effect structures. Nevertheless, REML is very popular for estimating variance components in mixed models, due to the reduced bias, and hence for determining smoothing/penalty parameters in (generalized) additive models as well [76]. With

ordinal predictors as considered here, it is exactly the latter that REML will be used for. Model comparison/selection will be discussed in more detail in Sect. 7.2.2 below. Besides likelihood-based methods, prediction error methods such as (generalized) cross-validation or the Akaike Information Criterion (AIC) can also be used for smoothness selection [75, 79].

Confidence Intervals In particular, the Bayesian interpretation of quadratic smoothing penalties is useful to derive confidence intervals. In the Gaussian identity link/linear mixed model case, one can derive the covariance matrix of the regression/basis coefficients' posterior distribution. In the generalized case the corresponding matrix from the penalized iteratively weighted least squares algorithm (PIRLS) used to estimate the parameters is taken. Using this matrix, let's say \mathbf{V}_β , one can calculate (point-wise) credible intervals for function $\hat{\mathbf{f}}_j = \mathbf{B}_j \hat{\boldsymbol{\beta}}$, denoting the vector of $\hat{f}_j(x)$ values at evaluation points x_{ij} . For each element \hat{f}_{ji} of $\hat{\mathbf{f}}_j$, we obtain an approximate $(1 - \alpha)100\%$ credible interval via $\hat{f}_{ji} \pm z_{1-\alpha/2} \sqrt{v_{ji}}$, where v_{ji} is the i th diagonal element of $\mathbf{B}_j \mathbf{V}_\beta \mathbf{B}_j^\top$, and z_q is the q -quantile of the standard normal distribution. It turns out that those credible regions also have good frequentist coverage rates [37, 46], and this is also the case when applied to smoothed ordinal effects [24]. As pointed out earlier though, interpretation of regression parameters, and hence the confidence intervals, changes depending on the constraint chosen (compare Sect. 7.1). Therefore, coverage can only be calculated/interpreted with respect to a "true" function where the chosen constraint holds. A problem, however, can occur with the second-order penalty (7.6), where substantial under-coverage is observed if the fitted regression function is close to being linear [24]. This problem is also found for (generalized) additive models with continuous covariates, and the suggested fix is to change the target of inference to the smooth term plus the overall model intercept [37, 78]. If simultaneous confidence intervals/confidence bands are needed instead of point-wise ones, we could also proceed like in the continuous case by posterior sampling [78].

Significance of Ordinal Predictors In general, we would like to test null hypothesis $H_0 : f_j(x) = 0$ for all potential x . With ordinal predictors, this means for all levels of the predictor of interest. Analogously to the confidence intervals above, let $\hat{\mathbf{f}}_j = \mathbf{B}_j \hat{\boldsymbol{\beta}}$ denote the vector of $\hat{f}_j(x)$ evaluated at the predictor levels (i.e., with appropriately chosen dummy-matrix \mathbf{B}_j). Then, following [77, 78], we can define a Wald-type test statistic

$$T = \hat{\mathbf{f}}_j^\top \mathbf{V}_j^- \hat{\mathbf{f}}_j,$$

where \mathbf{V}_j^- is an appropriately chosen pseudo-inverse of $\mathbf{V}_j = \mathbf{B}_j \mathbf{V}_\beta \mathbf{B}_j^\top$. Under H_0 , the distribution of T is obtained via a linear combination of random variables following specific χ^2 -distributions. For details on this linear combination and the pseudo-inverse \mathbf{V}_j^- , see [77, 78].

When taking the mixed models perspective of penalty (7.5), the null hypothesis above can alternatively be written as $H_0 : \tau_j^2 = 0$. In the Gaussian/linear mixed model with only one smooth term/variance component a (restricted) likelihood ratio test can be used [9, 10, 52], which also works well with ordinal data [18, 20, 57], and provides extensions/approximations in case of multiple ordinal predictors/smooth terms [18, 26, 52]. To the best of our knowledge, however, no generalization exists beyond models with Gaussian response. Therefore we will use the Wald-type test here.

7.2.2 Smoothing and Selection

Smoothing of ordinal predictors strongly reduces the effective number of parameters. However, if many explanatory variables are available smoothing alone is not sufficient. Typically, the researcher also wants to select variables, that is, he/she wants to include only variables that contribute to explaining the variation of the response. A modern, and overwhelmingly popular, approach in the “theory and methods”-oriented statistics community, is the so-called *sparsity-inducing* penalties approach, which means that by using appropriately designed penalty terms, the number of covariates in the model is also reduced. In applied statistics, however, more classical approaches, such as stepwise selection, are also still quite common. In what follows, we will hence introduce both penalty-based variable selection for categorical, in particular ordinal, factors and stepwise selection using the inferential tools presented above (Sect. 7.2.1.2).

7.2.2.1 Group Lasso and Similar Approaches

Selection of categorical predictors can be obtained by using the penalty term

$$J_j(\boldsymbol{\beta}_j) = \sqrt{k_j} \|\boldsymbol{\beta}_j\|_2, \quad (7.7)$$

where $\|\boldsymbol{\beta}_j\|_2 = (\beta_{j_1}^2 + \dots + \beta_{j_{k_j}}^2)^{1/2}$ is the L_2 -norm of the parameters of the j th group, which refers to one categorical predictor. The penalty has been called the *group lasso* [80], and is an extension of Tibshirani’s lasso [58]. The latter penalizes the sum of absolute values of regression coefficients, which leads to parameters that can be estimated to be exactly zero. If the regression coefficient of a continuous or binary covariate is exactly zero, the variable is excluded from the model. That means, all the covariates with non-zero coefficients are selected. With multi-categorical predictors, however, exclusion of single (dummy) coefficients does not necessarily lead to variable selection since the entire *group* of dummy coefficients that belong to a factor needs to be set to zero simultaneously to have it excluded from the model. The group lasso penalty (7.7) is able to do exactly that: it select groups

of parameters simultaneously and, given an appropriate penalty parameter, excludes whole predictors from the explanatory term. As an alternative to the group lasso, we may also use other sparsity-inducing penalties, such as SCAD [16] or MCP [81], on the norm of subvectors β_j , leading to group SCAD, group MCP, etc.; see also [31].

Penalty (7.7) can be used for any categorical variable, including ordinal ones. With ordinal covariates, however, penalty (7.7) does not exploit the additional information obtained from the categories' ordering. In order to use this additional information, and to smooth across categories of the selected variables as we did in Sect. 7.2.1, we can switch to a group-wise, sparsity-inducing penalty on the vectors of pairwise differences $\delta_j = (\delta_{j1}, \dots, \delta_{j,k_j-1})^\top$, with $\delta_{jl} = \beta_{j,l+1} - \beta_{jl}$ [23], similarly to Sect. 7.2.1.2. If all components from δ_j are set to zero, we have coefficients $\beta_{j1} = \dots = \beta_{jk_j}$. That means, all levels of predictor x_j have the same 'effect'. If all levels have the same effect/coefficient, it does not make sense to distinguish between them, and x_j can be excluded from the model. In fact, when looking at the potential constraints on the coefficients $\beta_{j1}, \dots, \beta_{jk_j}$ from Sect. 7.1, such as (7.4), it becomes clear that $\beta_{j1} = \dots = \beta_{jk_j}$ also means that $\beta_{j1} = \dots = \beta_{jk_j} = 0$.

It should be noted that $\beta_{j1} = \dots = \beta_{jk_j} = 0$ follows from $\delta_j = 0$ no matter which constraint is chosen: dummy coding with reference category r , i.e., $\beta_{jr} = 0$, where r may even change across variables, effect coding with $\sum_l \beta_{jl} = 0$, or (7.4). More generally speaking, when using a group-wise penalty on δ_j the penalized maximum likelihood estimate is invariant against the constraint chosen, since switching between the constraints mentioned in Sect. 7.1 simply means a vertical shift of the entire β_j vector (as already pointed out), and hence does not affect δ_j at all. Also, the model fit is not affected, because the constant α (which does change if changing the constraint) is not penalized. If using the more general penalty (7.7), by contrast, the result is not invariant under changes of the constraint. That is why dummy variables (7.3) are typically 'standardized' in some sense [12, 54] when penalty (7.7) is used. The same applies to analogous group-wise, sparsity-inducing penalties, such as group SCAD and group MCP.

7.2.2.2 Forward/Backward Selection in Generalized Additive Models

When considering ordinal smoothing penalties in the framework of generalized additive models, methods such as statistical testing developed there can also be used for stepwise variable selection. Generally speaking, in case of so-called *forward selection* we usually start with a model containing only an intercept, i.e., no covariates at all (but we may also start with a small model containing only a set of mandatory covariates, which have to be included for substantial reasons). Next, we add covariates until the model is not improved anymore. Using *backward selection*, by contrast, one starts with the full model containing all the covariates available, and successively excludes covariates until the model does not improve anymore/deteriorates significantly. Of course, we may also combine the two directions by either including or excluding covariates in each step of

the algorithm. If the number of predictors is relatively large compared to the sample size, forward selection in particular appears attractive, since the full model (the starting point for backward selection) is sometimes hard to fit. Furthermore, stepwise selection typically leads to smaller models that are easier to interpret. When implementing the concrete algorithm, however, one needs to choose both the criterion deciding on the covariate to include/exclude, and the stopping criterion.

A popular choice is the AIC, which can be used for both tasks. For instance, the AIC is the default in the standard R `step()` procedure. When choosing the mixed models interpretation of generalized additive models, we may either use the marginal or conditional AIC. The marginal AIC approach, however, favors simpler models excessively [78], while the conditional AIC is biased towards larger/too large models [25]. A correction to fix the latter problem has, for example, been proposed by Wood et al. [79] and implemented in `mgcv`. Consequently, we may easily use this version of the AIC with smoothed ordinal predictors as well.

A common alternative to the AIC, or other information criteria, is stepwise selection via p -values/statistical testing (an approach that is particularly popular among applied researchers) either “automatically”, in terms of an algorithm, or intuitively “by hand”. In case of forward selection, for instance, in each step the variable that maximizes the model’s goodness-of-fit is added, until the newly added variable is not statistically significant, i.e., the model does not improve significantly anymore. With backward selection, we exclude variables until the model becomes significantly worse than the model from the step before or the initial/full model. By using the tests presented in Sect. 7.2.1.2, we can also proceed this way with ordinal predictors. For selecting the variable to add, we can, for instance, use the model’s deviance (as the measure of goodness-of-fit). However, it should be noted that, regardless of whether the AIC, p -values, or some other criteria were used for model selection, standard inference for the model previously selected on the data at hand can be misleading. This issue is typically discussed under the term *post model selection inference*; see, e.g., [34] and references therein. For instance, this means that tests as given above typically do not control the type I error rate, as they would on a fixed model (which had, for example, been chosen for some substantial reason without looking at the data also used for estimating unknown model parameters). Also, the α -level used for stepwise selection as described above should be interpreted as a tuning parameter controlling rather model complexity than the type I error rate with respect to irrelevant predictors; see also the simulation studies in Sect. 7.3.

7.2.3 Level Fusion

An alternative approach to reducing the number of parameters is to identify clusters of categories that are to be distinguished in their effect on the responses. When dealing with ordinal covariates, this can, for instance, be done by a variant of

the fused lasso [59]. Alternatively, more classical approaches, such as stepwise selection, can be employed on appropriately recoded dummy variables [70].

7.2.3.1 Fused Lasso and Similar Approaches

For ordered predictors it is natural to assume that clusters of categories refer to adjacent categories. A penalty that is inspired by the so-called *fused* lasso [59] and enforces the fusion of adjacent categories is [22, 32, 65]

$$J_j(\boldsymbol{\beta}_j) = \sum_{l=1}^{k_j-1} w_l^{(j)} |\beta_{j,l+1} - \beta_{jl}|, \quad (7.8)$$

where the $w_l^{(j)}$ are additional, appropriately chosen weights. Those weights may, for instance, depend on the number of observations in the respective categories [2, 6], compare also Sect. 7.5. Also, they can be chosen proportionally to the absolute difference between unpenalized dummy coefficients, yielding a variant of the adaptive (fused) lasso [82]. Of course, instead of the lasso-type penalty on the differences of adjacent coefficients, we may also use sparsity-inducing penalties such as SCAD [16] or MCP [81]. In general, we recode the model using parameters $\delta_{jl} = \beta_{j,l+1} - \beta_{jl}$ as done in Sect. 7.2.2.1, and may then use any penalty that is able to shrink δ -parameters exactly to zero, but individually (that is, not in a group-wise fashion, as done in Sect. 7.2.2.1).

A problem of the standard fused lasso (7.8) is that in special cases it does not lead to the desired fusion at all. Specifically, if weights are constant across levels, i.e., $w_1^{(j)} = \dots = w_{k_j-1}^{(j)}$ (e.g., in a balanced design), and unpenalized estimates are monotone across levels, the penalty essentially only depends on the distance between the two most extreme categories and no fusion is obtained in between. In this case, switching to the adaptive version or non-convex penalties can help [56]. Also compare SCOPE and the “Range” penalty in Sect. 7.5.1.2.

7.2.3.2 Stepwise Selection After Recoding

Technically, reparametrization by δ -coefficients as described above is done by introducing modified dummy variables

$$z_{jl} = \begin{cases} 1 & \text{if } x_j > l, \\ 0 & \text{otherwise.} \end{cases} \quad (7.9)$$

For instance, if having an ordinal predictor x_j with four levels, we need dummies z_{j1}, z_{j2}, z_{j3} such that [70]:

Level	z_{j1}	z_{j2}	z_{j3}
1	0	0	0
2	1	0	0
3	1	1	0
4	1	1	1

Instead of sparsity-inducing penalties, we could also run classical selection algorithms on the recoded data, that means, on the z -dummies. For instance, AIC-based stepwise selection as implemented in the R function `step()`.

7.3 Numerical Experiments: L_1 -Regularization vs. Forward Selection

In recent decades there has been tremendous methodological research on L_1 -regularization for variable and model selection in a range of models and settings. With more than 35,000 citations on Google Scholar, for instance, the original lasso paper [58] is one of the most popular contributions to statistical learning. Many applied researchers, however, still use more classical approaches for model selection, such as forward/stepwise selection, as pointed out already. In this section, we hence compare the forward selection procedures for ordinal predictors described above to L_1 -regularization in some illustrative simulation studies. We will focus on ‘classical’, medium-sized problems with both a moderate number of (ordinal) predictors and sample size.

7.3.1 *Smoothing and Selection of Covariates*

Assuming the data analyst’s goal is variable selection rather than fusion of categories, we are going to compare forward selection based on the `ordinal` basis within `mgcv` (compare Sect. 7.2.2.2) to our group lasso approach (from Sect. 7.2.2.1). In addition, we will consider the common R `step()` function on a linear (`lm()`) model with either linear effects across levels or classical factor modeling. That means, ordinal predictors are either included in `lm()` as numerical variables or factors.

7.3.1.1 Simulation Setup

In our first simulation study, we assume that there are three influential (ordinal) predictors with effects as shown in Fig. 7.1. The effect of X_1 is almost linear across

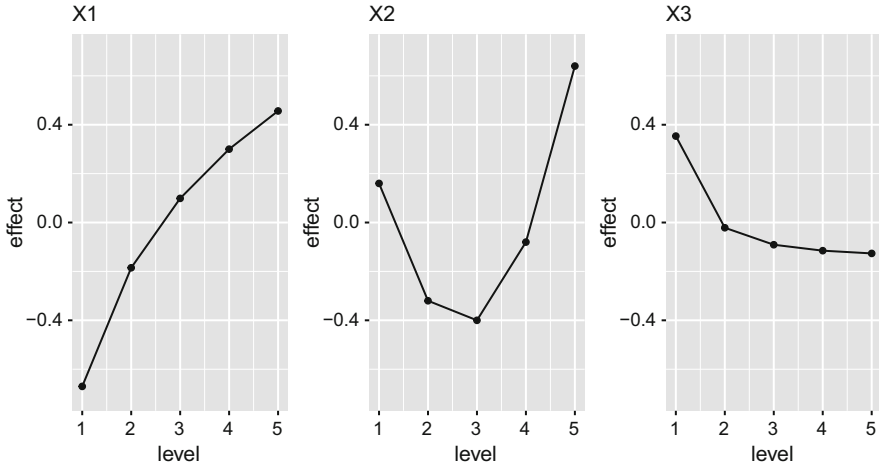


Fig. 7.1 True effects of influential predictors in simulation study 1

levels, X2 is non-monotone, and X3 is monotone but nonlinear, with smaller effect size than X1 and X2. In addition, we generate seven irrelevant predictors X4–X10, i.e., with effects being zero. Factor levels are randomly drawn from $\{1, \dots, 5\}$, which means that the design is approximately balanced with each covariate having the same number of categories. The sample size is $n = 100$, and the error term is standard normal.

Using the data generated, we select the most appropriate model by the methods given above. In case of the ordinal forward selection, we consider $\alpha = 0.05$, $\alpha = 0.1$, and the (conditional) AIC. Smoothing parameters are determined by REML here, whereas the group lasso uses fivefold cross-validation.

7.3.1.2 Results and Discussion

We ran the simulation 1000 times and report the results in terms of selection frequencies and mean squared error (MSE) in Figs. 7.2 and 7.3, respectively. Mean model sizes in terms of the number of predictors/factors being chosen (not the number of estimated parameters) are provided in Table 7.1 (so the ‘truth’ is 3 here). Grey bars in Fig. 7.2 indicate that the covariate is actually relevant (i.e., X1, X2, X3, which is known in simulations). So, heights close to one are desirable here. In case of black bars (X4–X10) on the other hand, the covariate is just noise, which means that selection frequencies, which are averaged across X4–X10 in Fig. 7.2 (bottom right), should be low. In case of an influential covariate (X1, X2, X3), the mean squared error is calculated as the mean across all five levels, and illustrated as boxplots across replicates $1, \dots, 1000$ in Fig. 7.3. In case of X4–X10 (Fig. 7.3, bottom right), it is also averaged across X4–X10 for each replicate. That is why

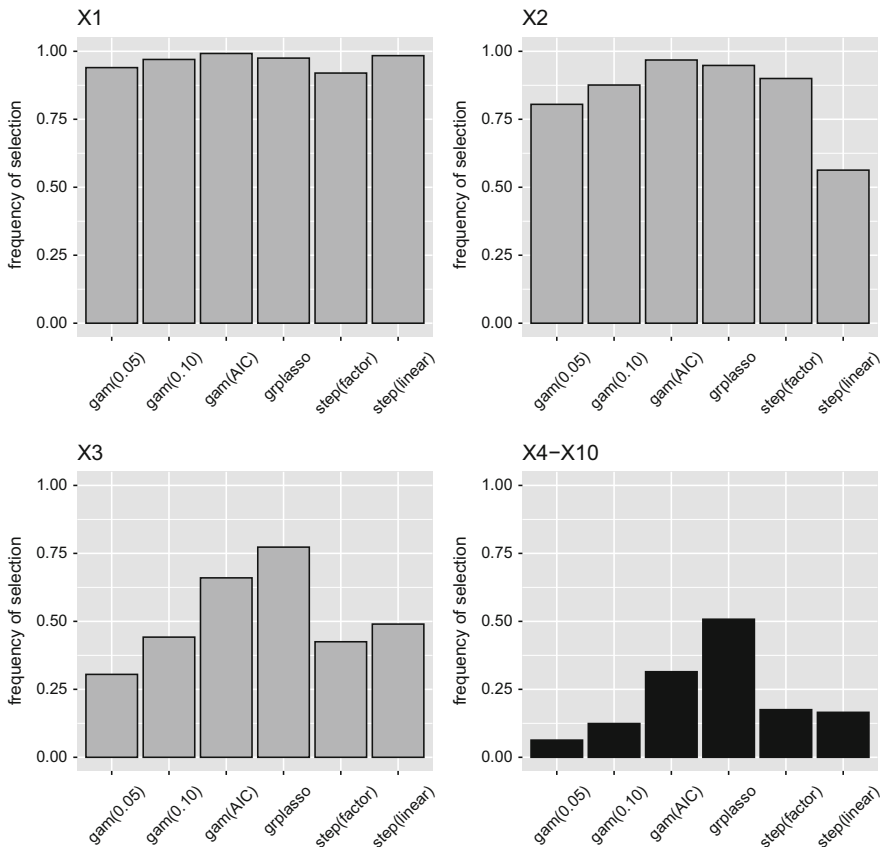


Fig. 7.2 Selection frequencies for predictors in simulation study 1

each boxplot in Fig. 7.3 is made of 1000 data points, and the box is still visible even if selection frequencies for noise variables are below 25% (for a non-zero MSE, just one noise variable has to be selected).

When looking at Fig. 7.2, we see that all the methods considered quite consistently select predictor X1 having approximately linear effects. If the effects are non-monotone (X2), the linear model in particular suffers, which is not surprising, of course. If effect sizes go down (X3) selection frequencies drop as well. Only the group lasso and AIC-based forward selection for ordinal predictors within `gam()` still produce selection frequencies well above 50%. This, however, comes at a price: irrelevant covariates (X4–X10) are selected very frequently as well; see also Table 7.1. This behavior (of selecting too large models) is well known for L_1 -regularization with tuning parameters chosen via (standard) cross-validation (see, e.g., [17] and references therein). That is why remedies such as the relaxed lasso [43] and stability selection [44] have been proposed, and are highly recommended for ordinal/categorical predictors as well. Although correction against over-complex

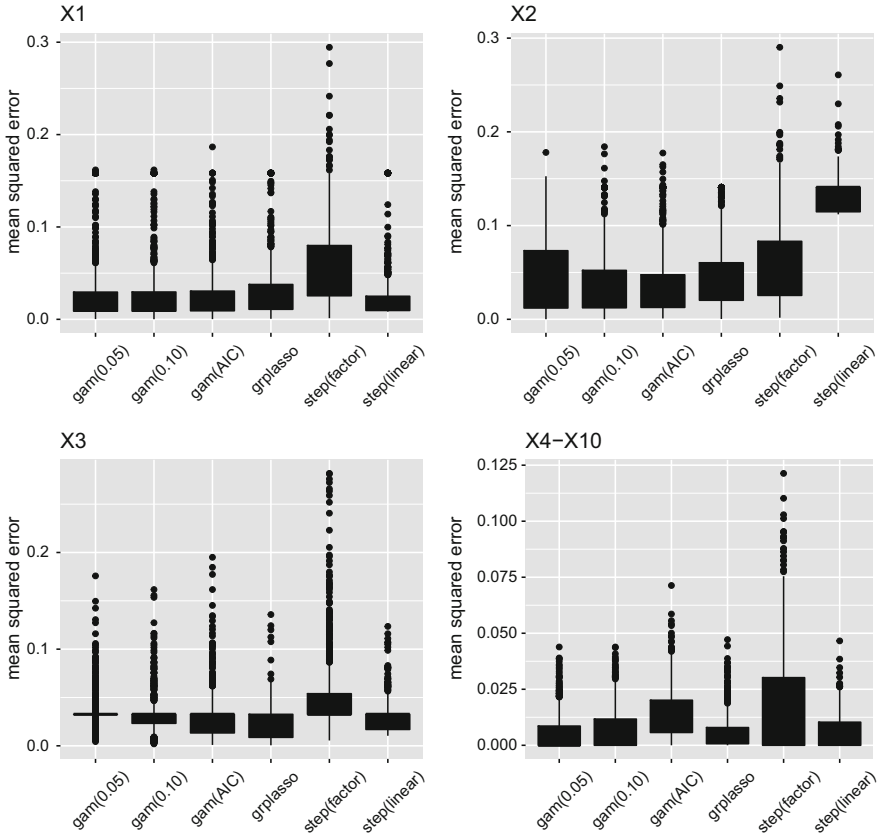


Fig. 7.3 Estimation accuracy (MSE) in simulation study 1

Table 7.1 Mean model sizes (incl. standard deviation) in terms of the number of predictors/factors being chosen

gam(0.05)	gam(0.1)	gam(AIC)	grplasso	step(factor)	step(linear)
2.493 (0.964)	3.158 (1.126)	4.824 (1.446)	6.251 (2.302)	3.472 (1.480)	3.195 (1.274)

models [79] is supposed to be employed within `gam()` when selecting the model via (conditional) AIC, this might not be sufficient here, see Table 7.1. Eventually, however, it is for the data analyst to decide whether he/she is willing to risk that some influential covariates are missed for the sake of a sparse model, as is the case here when using forward selection with a small α . As already pointed out in Sect. 7.2.2 though, α should merely be interpreted as a tuning parameter for stepwise selection, which does not necessarily control the type I error rate with respect to irrelevant predictors. Figure 7.2 indeed shows that the relative frequency of truly irrelevant covariates X4–X10 is (slightly) larger than α ; with $\alpha = 0.1$, for example, it is about 13% in the setting considered here.

In terms of the MSE (Fig. 7.3), it becomes clear that smoothing, either within `gam()` or the group lasso, works very well for ordinal predictors. It adapts well to different situations, whereas purely linear modeling obviously suffers in case of very non-linear, or even non-monotone, effects. The classical factor model does not seem to be a preferable choice for ordinal data either.

7.3.2 *Level Fusion and Selection*

L_1 -regularization, in particular variants of the fused lasso [59], have become very popular for fusion of regression parameters, and might even be called the current state of the art. When it comes to ordinal predictors, those methods can also be used for level fusion (compare Sect. 7.2.3.1). Rather classical approaches, such as forward selection, applied on split-coded variables as described in Sect. 7.2.3.2, however, have been largely neglected. In what follows, we are going to compare those two approaches in a simulation study, that is, the fused lasso with penalty being chosen via fivefold cross-validation and the standard R (AIC-based) `step()` procedure on split-coded variables. As before, we will focus on medium-sized problems.

7.3.2.1 **Simulation Setup**

The setting from above is slightly altered such that predictors X1 and X2 now have partly constant effects over categories (see Fig. 7.4). The rest of the setting (sample size, error term, etc.), however, remains unchanged. In total, we still have 10 ordinal covariates (with 5 levels each), seven of which (X4–X10) have zero effect on the response.

7.3.2.2 **Results**

Figure 7.5 shows the overall selection frequencies of the (ordinal) factors, with a covariate being counted as selected if at least two categories were not fused. We see that both methods are very successful in selecting X1 and X2 (having the largest effects). In case of X3, selection frequencies are a bit higher for the fused lasso, but this is also the case when looking at noise variables X4–X10. When looking at fusion frequencies between levels (Fig. 7.6), we obtain the same (and expected) result: L_1 -regularization with tuning parameters chosen via cross-validation tends to select slightly more complex models. The main advantage of penalization is found with respect to the MSE (Fig. 7.7), in particular for covariates with small (X3) or zero (X4–X10) effect.

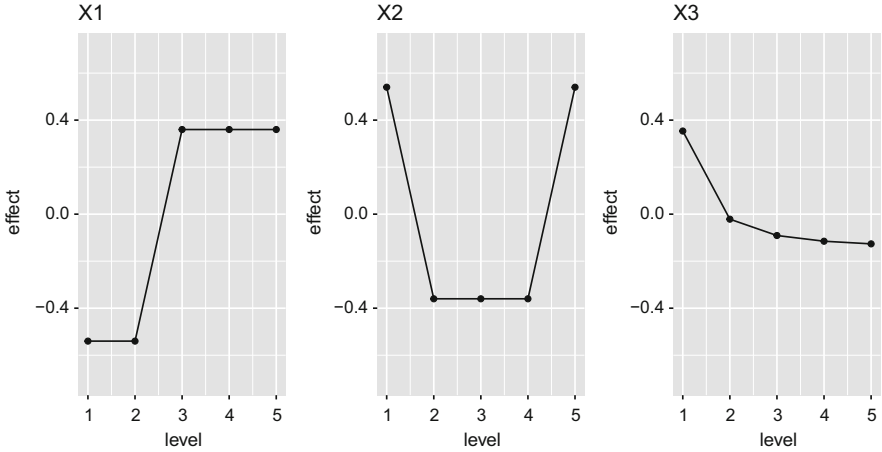


Fig. 7.4 True effects of influential predictors in simulation study 2

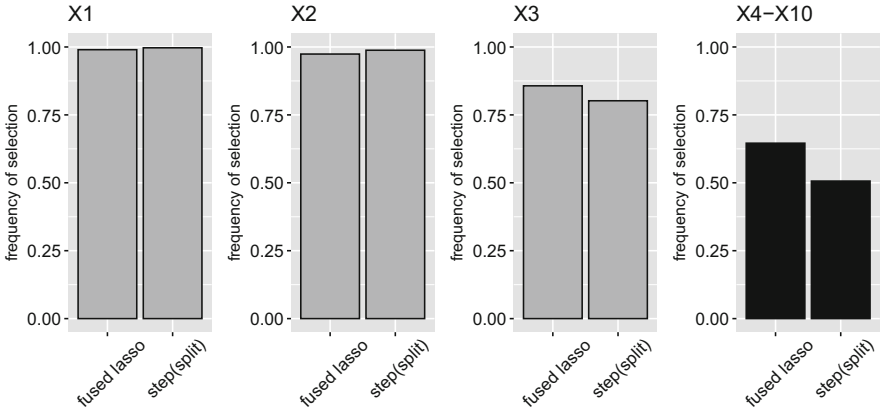


Fig. 7.5 Selection frequencies for predictors in simulation study 2

7.4 Case Study: The ICF

The ordinal group lasso [23] as described in Sect. 7.2.2.1 has originally been developed for analyzing the ICF Core Set for chronic widespread pain (CWP). ICF Core Sets constitute a first attempt to make the ICF [74] applicable in clinical practice. In total, the ICF consists of about 1400 so-called ICF *categories*, each of which refers to a health or a health-related domain. (To clarify, ICF *categories* should not be confused with the *categories* of a categorical variable. In World Health Organization (WHO) terminology *category* denotes the whole factor.) Each ICF category is attributed to one of four ICF components:

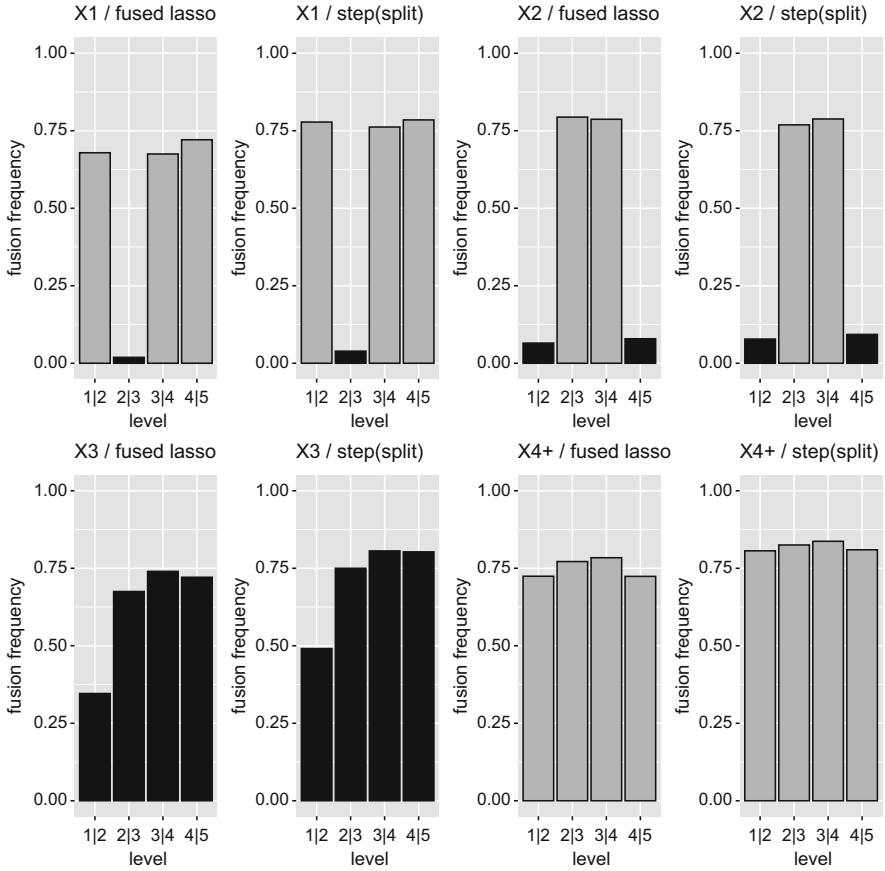


Fig. 7.6 Fusion frequencies in simulation study 2 when using the fused lasso or stepwise/forward selection on split-coded predictors

- (b) body functions (e.g. b140 ‘attention function’)
- (d) activities and participation (e.g. d450 ‘walking’)
- (e) environmental factors (e.g. e1101 ‘drugs’)
- (s) body structures (e.g. s770 ‘additional musculoskeletal structures related to movement’)

The coding scheme for (b), (d), and (s) is 0 (no problem), 1 (mild problem), 2 (moderate problem), 3 (severe problem) and 4 (complete problem). For (e), a differentiation is made between barriers and facilitators resulting in the coding scheme -4 (complete barrier), \dots , -1 (mild barrier), 0 (no barrier/facilitator), $+1$ (mild facilitator), \dots , $+4$ (complete facilitator). In other words, ICF categories are ordinally scaled variables with five or nine levels. Those variables may be used by

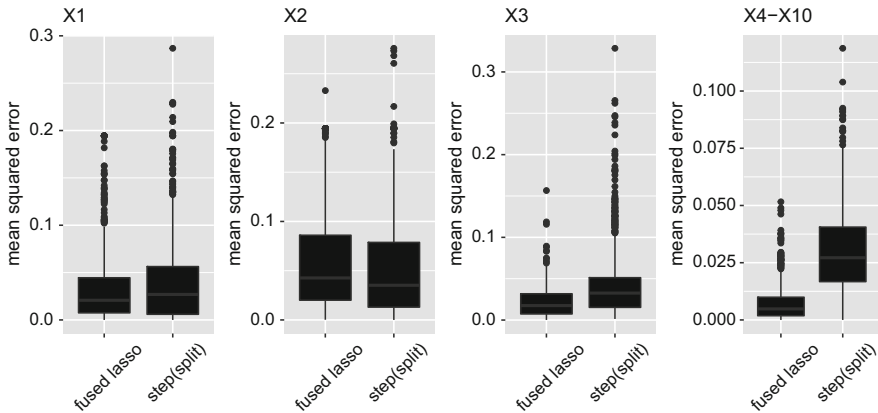


Fig. 7.7 Estimation accuracy (MSE) for predictors in simulation study 2

health professionals to document the health and functioning of patients by rating these on the levels described above [23].

In a first step, condition-specific ICF Core Sets, such as the Core Set for CWP consisting of 67 categories, which represent a large selection from the 1400 original ICF categories, were constructed based on mostly qualitative methods [7, 8]. On the one hand, the goal of the data analysis revisited here was to validate the ICF Core Set for CWP by comparing it to the physical health component summary measure (PCS) calculated from the well-established questionnaire SF-36 [42, 72], which is filled out by the patients. On the other hand, ICF categories should be identified that could also be used in more general health surveys. That is why the SF-36 outcome was regressed on the 67 (ordinal) factors from the ICF Core Set for CWP while using the (ordinal) group lasso for variable selection in a Gaussian model with identity link [23]. The data set considered is publicly available as part of the R package *ordPens* [19].

The group lasso selected 33 predictors, that means, about half of the predictors available. As we learned from Sect. 7.3, however, this model might be too large. So we ran the forward selection as proposed here as well, with $\alpha = 0.05$ and $\alpha = 0.1$. After convergence, we manually removed predictor e580, which clearly showed insignificant effect in the final model for both α values. The output of the resulting models is given in Figs. 7.8 and 7.9, and the fitted smooth effects are provided in Figs. 7.10 and 7.11, respectively. Interestingly, for instance, the model from Fig. 7.9 only has 11 predictors explaining more than 44% of the deviance (note, in the Gaussian model this value is simply calculated as $1 - \text{residual sum of squares} / \text{total sum of squares}$). The group lasso, for comparison, selects 33 predictors explaining about 46% [23]. Also for comparison, when running the GAM for ordinal predictors on the experts-selected, so-called *Brief Core Set* consisting of 26 variables, the deviance explained is about 41% (not shown in detail here), so even worse than the model with only 10 predictors in Figs. 7.8 and 7.10. In

```

Family: gaussian
Link function: identity

Formula:
phcs ~ s(d450, bs = "ordinal", m = 2) +
s(e1101, bs = "ordinal", m = 2) +
s(d455, bs = "ordinal", m = 2) +
s(e450, bs = "ordinal", m = 2) +
s(d910, bs = "ordinal", m = 2) +
s(b140, bs = "ordinal", m = 2) +
s(d410, bs = "ordinal", m = 2) +
s(d720, bs = "ordinal", m = 2) +
s(b455, bs = "ordinal", m = 2) +
s(b730, bs = "ordinal", m = 2)

Parametric coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  32.4075     0.3089   104.9  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
              edf Ref.df      F  p-value
s(d450)      1.000  1.001 30.171 < 2e-16 ***
s(e1101)     5.026  6.019  4.502 0.000207 ***
s(d455)      3.111  3.637  4.782 0.003969 **
s(e450)      3.481  4.453  2.340 0.047827 *
s(d910)      1.000  1.001 21.434 5.4e-06 ***
s(b140)      2.456  2.962  3.049 0.033594 *
s(d410)      1.000  1.000  4.979 0.026199 *
s(d720)      1.000  1.001  6.790 0.009504 **
s(b455)      1.000  1.001  6.493 0.011198 *
s(b730)      1.000  1.001  4.406 0.036435 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq. (adj) = 0.399  Deviance explained = 42.8%
-REML = 1364.6  Scale est. = 40.073  n = 420

```

Fig. 7.8 R output for the final ICF model after using forward selection with $\alpha = 0.05$ and `mgcv` with second-order smoothing penalty on dummy coefficients

summary, forward selection appears to work well here for choosing a sparse model with good fit to the data.

When looking at the fitted effects (Figs. 7.10 and 7.11), one can see that smooth terms with effective degrees of freedom (edf) being (close to) one in Figs. 7.8 and 7.9, respectively, are fitted as (virtually) linear. For most of the variables with ‘no problem’, ‘mild problem’, ... scale, such as d450 (‘walking’), d455 (‘moving around’), and d910 (‘community life’), fitted functions are decreasing, indicating negative effects on overall health (as given by PCS). Interestingly, however, CWP

```

Family: gaussian
Link function: identity

Formula:
phcs ~ s(d450, bs = "ordinal", m = 2) +
s(e1101, bs = "ordinal", m = 2) +
s(d455, bs = "ordinal", m = 2) +
s(e450, bs = "ordinal", m = 2) +
s(d910, bs = "ordinal", m = 2) +
s(b140, bs = "ordinal", m = 2) +
s(b640, bs = "ordinal", m = 2) +
s(d540, bs = "ordinal", m = 2) +
s(d720, bs = "ordinal", m = 2) +
s(b455, bs = "ordinal", m = 2) +
s(b740, bs = "ordinal", m = 2)

Parametric coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  32.407      0.306   105.9 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
              edf Ref.df      F p-value
s(d450)      1.000  1.000 31.556 < 2e-16 ***
s(e1101)     5.002  5.995  4.531 0.000178 ***
s(d455)      3.006  3.555  5.086 0.002244 **
s(e450)      3.252  4.167  2.211 0.063234 .
s(d910)      1.000  1.000 16.358 6.33e-05 ***
s(b140)      2.405  2.911  2.851 0.044812 *
s(b640)      2.353  2.906  3.019 0.051379 .
s(d540)      1.000  1.000  2.680 0.102397
s(d720)      2.117  2.543  3.244 0.025819 *
s(b455)      1.001  1.002  8.020 0.004856 **
s(b740)      1.000  1.000  5.242 0.022556 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) = 0.41   Deviance explained = 44.3%
-REML = 1360.5   Scale est. = 39.338   n = 420

```

Fig. 7.9 R output for the final ICF model after using forward selection with $\alpha = 0.1$ and `mgcv` with second-order smoothing penalty on dummy coefficients

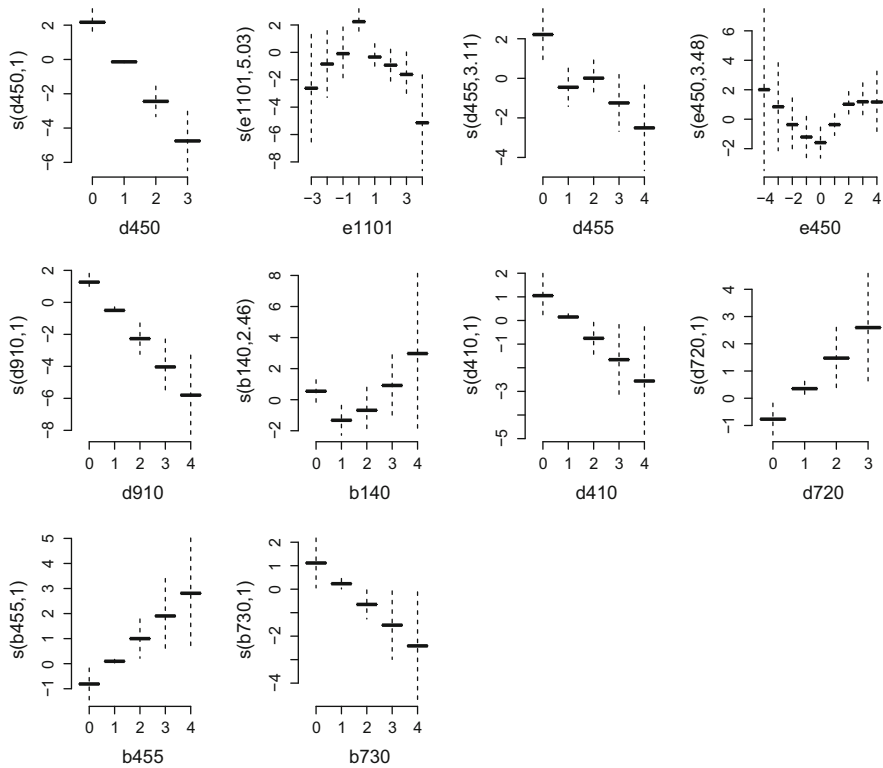


Fig. 7.10 R/mgcv plots for the model from Fig. 7.8 chosen by forward selection with $\alpha = 0.05$

patients with poorer overall health seem to have less problems with exercise tolerance functions (b455) and complex interpersonal interactions (d720). For b140 (‘attention functions’), estimated effects in the upper categories appear less reliable as indicated by the wide confidence intervals (dashed lines). The same is true for b640 (‘sexual functions’) in Fig. 7.11. The two environmental factors e1101 (‘drugs’) and e450 (‘individual attitudes of health professionals’) show reverse effects. With respect to e1101, patients for whom drugs are neither a facilitator nor barrier tend to have the best overall health. On the other hand, patients with poor health do not, or cannot afford to, care about individual attitudes of health professionals (e450).

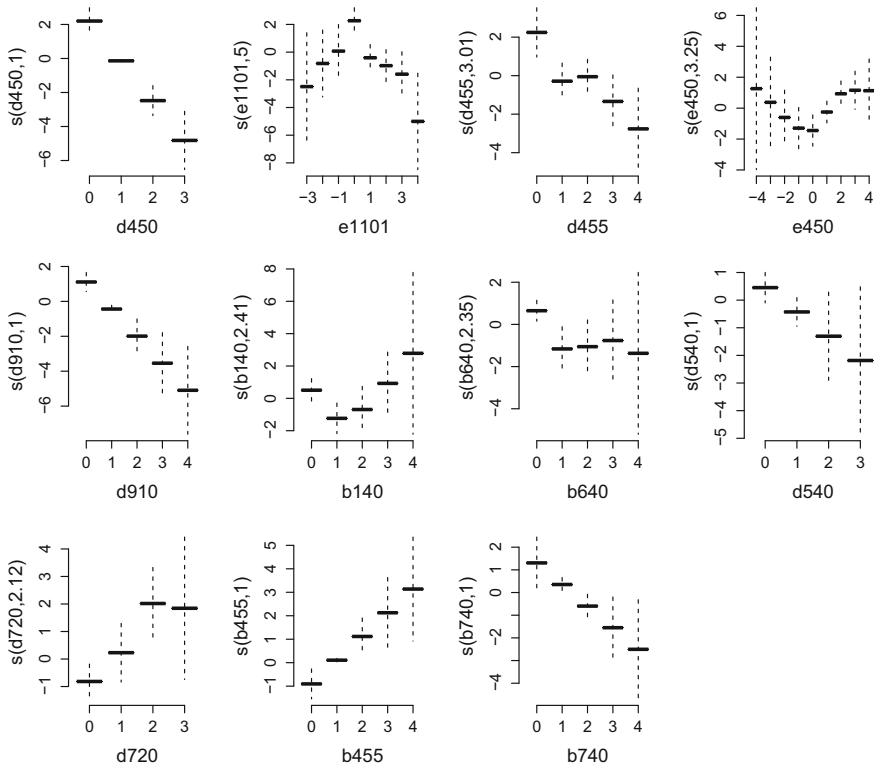


Fig. 7.11 R/mgcv plots for the model from Fig. 7.9 chosen by forward selection with $\alpha = 0.1$

7.5 Nominal Predictors and Categorical Response

7.5.1 Fusion Penalties for Nominal Predictors

Now let the categorical predictor x_j have unordered levels $1, \dots, k_j$. That means the response categories have no inherent ordering and the numbers $1, \dots, k_j$ have to be considered as mere labels for response categories. Then, it is natural to use the dummy variables (7.3) as basis functions since they do not use any order.

7.5.1.1 All-Pairs Penalties

A penalty that enforces the building of clusters of categories that share the same effect is

$$J_j(\beta_j) = \sum_{r < s} w_{rs}^{(j)} |\beta_{jr} - \beta_{js}|, \tag{7.10}$$

where the sum (within each categorical predictor) is over all pairs of categories r, s . Penalty (7.10) was originally introduced in the ANOVA framework under the name CAS-ANOVA (for collapsing and shrinkage in ANOVA) [2], and later used for regression as well [22]. Since the penalty only considers differences of coefficients, it is invariant against (group-wise) vertical shifts and hence invariant against the concrete constraint that is chosen (see Sect. 7.1). For the weights $w_{rs}^{(j)}$ a useful choice is $w_{rs}^{(j)} = (k_j + 1)^{-1} \sqrt{(n_r^{(j)} + n_s^{(j)})/n}$, where $n_r^{(j)}$ and $n_s^{(j)}$ are the number of observations in category r and s of predictor c_j , respectively [2, 22]. But other types of weights have been proposed as well; for instance, weights that make sure that coefficient paths have a ‘tree-like’ structure, which means that categories that have been fused for some λ remain fused for increased penalty as well, in other words, they “cannot ‘split’ anymore in the future” [6]. As before with ordinal predictors, weights can be chosen to be data-adaptive giving a version of the adaptive lasso [82], also with oracle properties [2, 22]. Alternatively, concave and non-decreasing penalties can be used [35, 47].

A problem with “all-pairs” penalties as given above is that they favor clusters of unequal size [56]. To make things clearer, consider the following very simple but illustrative example (which is similar to [22, 56]). There is only one factor with nine levels and ten observations on each level. The true level-specific means of y show a three-cluster structure as illustrated by the dashed line in Fig. 7.12 (top left); simulated data (assuming standard normal errors/noise) is given as boxplots. Figure 7.12 (top right) also shows the coefficient paths when applying CAS-ANOVA 7.10 to the data at hand (top left), with the y-axis corresponding to the strength of the penalty, and colors giving the “true” grouping. We see that the green coefficients on the right remain to form their own groups, even for penalization where all the other (red and black) coefficients are fused already. This means that for no value of the penalty parameter λ , is the correct grouping obtained in this example (with the data simulated). Under some circumstances, results are better when switching to the adaptive version, but a more promising approach appears to be the SCOPE penalty proposed very recently [56], and described below.

7.5.1.2 The SCOPE and Range Penalty and Tree-Structured Approaches

In order to circumvent the preference of “all-pairs” penalties for clusters of unequal size, the basic idea of SCOPE [56] is to switch to a first-order difference penalty on sorted regression coefficients. Let $\beta_{j(l)}$ be the l th smallest entry of β_j , then SCOPE is defined as

$$J_j(\beta_j) = \sum_{l=1}^{k_j-1} \rho_j(\beta_{j(l+1)} - \beta_{j(l)}), \quad (7.11)$$

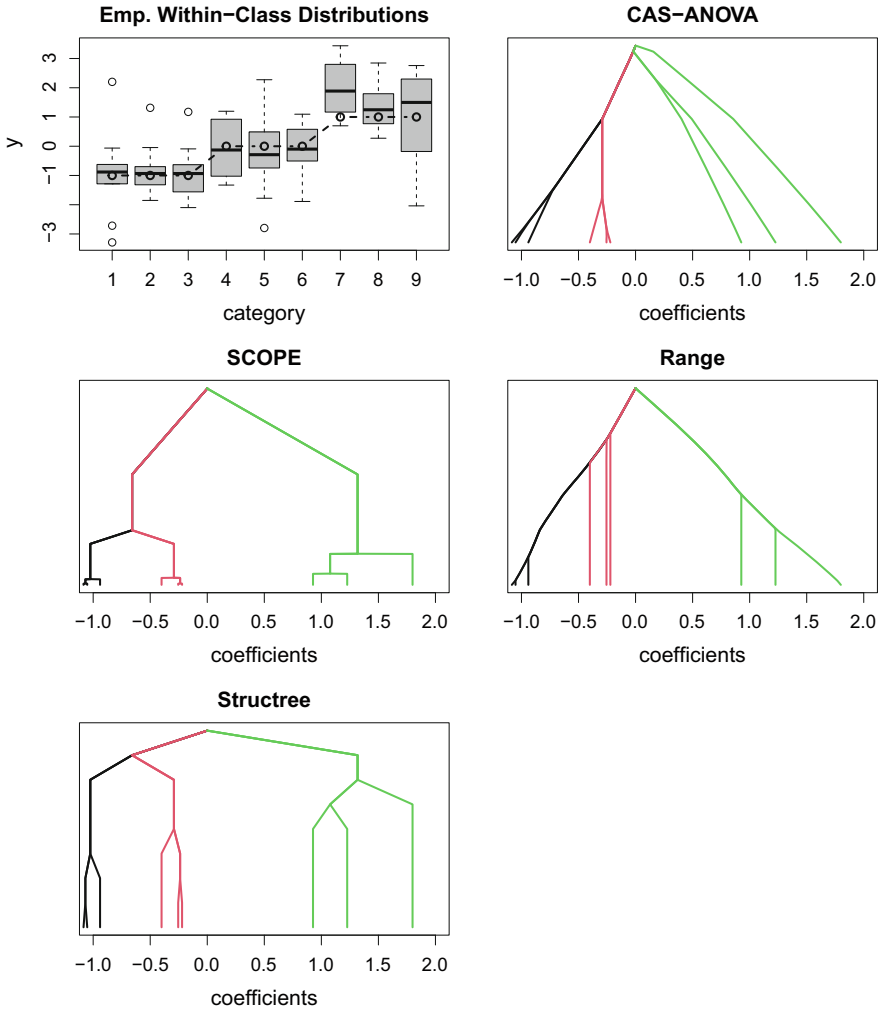


Fig. 7.12 Toy example with simulated data (top left) and coefficient paths for CAS-ANOVA (top right), SCOPE (center left) and Range penalty (center right), and Structree (bottom left); y-axis corresponds to the amount of penalty/fusion step if not given otherwise

where ρ_j is a concave (and non-convex) non-decreasing penalty function [56], specifically, the MCP [81]. Although the penalty looks similar to the ordinal case from Sect. 7.2.3.1, there is an important difference: now the ordering is not defined by the levels' ordering but by the size of the coefficients. An algorithm for fitting linear and logistic models with the SCOPE penalty is implemented in R package CatReg [55]. When applying this penalty to the simulated data from Fig. 7.12 (top left), we indeed obtain very nice coefficient paths as seen from Fig. 7.12 (center left).

Although not observed in the example given here, SCOPE-paths do not necessarily have a tree-like structure, which may be seen as a drawback [6].

One important feature of the SCOPE-algorithm [55] applied above is that it starts by sorting the categories/coefficients according to the unpenalized estimates. So we may think that we could do the same but then proceed with the fused lasso, like in the ordinal case. This, however, leads to the so-called ‘Range’ penalty [56]

$$J_j(\boldsymbol{\beta}_j) = \sum_{l=1}^{k_j-1} |\beta_{j(l+1)} - \beta_{j(l)}| = \max_l \beta_{jl} - \min_l \beta_{jl}, \quad (7.12)$$

which only shrinks the largest and the smallest of the coefficients together, but does not encourage fusion in between. This is also nicely seen in Fig. 7.12 (center right): even categories with ordinary least squares estimates very close to each other are not fused until they are collapsed with one of the most extreme levels. This also makes clear that the same happens with the ordinal fusion penalty (7.8), if weights are constant across levels and unpenalized estimates are monotonically increasing or decreasing.

An alternative approach that is not based on penalties but designed such that a clear hierarchy of clusters is always obtained is tree-structured modeling [64]. The method explicitly fuses categories of nominal or ordinal predictors successively by recursive partitioning with stopping based on an information criterion (AIC/BIC), statistical testing, or cross-validation. It also allows parametric and smooth components in the predictor and can be seen as a combination of parameter estimates of a generalized additive model and a hierarchical clustering process for categories. It can be applied by using the R package `structree` [1], which yields the coefficient paths shown in Fig. 7.12 (bottom left) when applied to the toy data from Fig. 7.12 (top left). It is seen that the coefficients distinctly reflect the true underlying structure of the clusters.

7.5.2 Regularization for Multi-categorical Response Models

Now let the response variable be multi-categorical with $y \in \{1, \dots, k\}$ and $\pi_r(\mathbf{x}) = P(y = r|\mathbf{x})$ denoting the probability of a response in category r given a vector of explanatory variables \mathbf{x} . We first consider ordinal responses, then the case of nominal responses.

7.5.2.1 Ordinal Responses

Classical ordinal response models have the form

$$g_r(\boldsymbol{\pi}(\mathbf{x})) = F(\beta_{r0} + \mathbf{x}^\top \boldsymbol{\beta}_r), \quad r = 1, \dots, k - 1, \quad (7.13)$$

where $\boldsymbol{\pi}(\mathbf{x})^\top = (\pi_1(\mathbf{x}), \dots, \pi_k(\mathbf{x}))$, $g_r(\cdot)$ are transformation functions, and $F(\cdot)$ a strictly monotonic distribution function. The most widely used models are:

- the cumulative model, which uses $g_r(\boldsymbol{\pi}(\mathbf{x})) = \pi_1(\mathbf{x}) + \dots + \pi_r(\mathbf{x})$,
- the adjacent categories model, with $g_r(\boldsymbol{\pi}(\mathbf{x})) = \pi_r(\mathbf{x})/(\pi_{r+1}(\mathbf{x}) + \pi_r(\mathbf{x}))$,
- the sequential model, which uses $g_r(\boldsymbol{\pi}(\mathbf{x})) = \pi_r(\mathbf{x})/(\pi_r(\mathbf{x}) + \dots + \pi_k(\mathbf{x}))$.

If $\boldsymbol{\beta}_1 = \dots = \boldsymbol{\beta}_{k-1}$, and $F(\cdot)$ is the logistic distribution function one obtains the proportional odds model [40]. If parameters may vary across categories the logistic version is known as the non-proportional odds model or partial proportional odds model [3]. However, the general form generates a variety of models; see, for example, [60].

In the general model (7.13) the parameters are category-specific, which makes it a model with many parameters, typically leading to estimates with large variance and interpretation being difficult. The most often used restriction (as, for instance, in the proportional odds model) is $\boldsymbol{\beta}_1 = \dots = \boldsymbol{\beta}_{k-1} = \boldsymbol{\beta}$, which means one has global effects, that is, effects that do not vary across categories. However, the restriction is frequently too restrictive and yields poor data fit.

A compromise between data fit and models with a simple structure (a specific type of ‘bias-variance trade-off’) is obtained by assuming that parameters vary not too strongly across categories. With $\boldsymbol{\beta}_r^\top = (\beta_{r1}, \dots, \beta_{rp})$ a penalty that enforces the smoothness of effects across categories is [68]

$$J_j(\boldsymbol{\beta}_{\cdot j}) = \sum_{l=2}^{k-1} (\beta_{lj} - \beta_{l-1,j})^2,$$

where $\boldsymbol{\beta}_{\cdot j}^\top = (\beta_{1j}, \dots, \beta_{k-1,j})$ collects the parameters linked to variable x_j . The penalty is similar to (7.5); however, smoothing is not across categories of the explanatory variable but across categories of the response. For large smoothing parameters the parameter becomes global. A disadvantage, at least in some situations, is that this type of difference penalty does not enforce variable selection. An alternative and sparsity-inducing version is

$$J_j(\boldsymbol{\beta}_{\cdot j}) = \left(\sum_{l=1}^{k-1} \beta_{lj}^2 \right)^{1/2} + \zeta \left(\sum_{l=2}^{k-1} (\beta_{lj} - \beta_{l-1,j})^2 \right)^{1/2} = \|\boldsymbol{\beta}_{\cdot j}\|_2 + \zeta \|\mathbf{D}\boldsymbol{\beta}_{\cdot j}\|_2,$$

where \mathbf{D} is a matrix that generates differences and ζ is an additional smoothing parameter. The first term enforces selection of variables while the second term enforces the choice between global and category-specific effects. For large, but not too large, ζ some but not all of the variables will have global effects. For very large ζ all effects are global. Details on how to obtain solutions are given in [50].

An alternative way to simplify the complex parameter structure that is found in the general models (7.13) is to separate the location structure from the dispersion structure. Models that are able to separate these effects are the location-scale model

and the location-shift model. The *location-scale model*, which was proposed by [40] for cumulative models, has the form

$$g_r(\boldsymbol{\pi}(\mathbf{x}, \mathbf{z})) = F\left(\frac{\beta_{r0} + \mathbf{x}^\top \boldsymbol{\beta}}{\exp(\mathbf{z}^\top \boldsymbol{\gamma})}\right), \quad r = 1, \dots, k-1,$$

where \mathbf{z} is an additional vector of covariates, which can be distinct, can overlap with \mathbf{x} or can be identical to \mathbf{x} , and $\boldsymbol{\gamma}$ are the corresponding weights. While $\beta_{r0} + \mathbf{x}^\top \boldsymbol{\beta}$ represents the location, the term $\exp(\mathbf{z}^\top \boldsymbol{\gamma})$ represents dispersion. If $\mathbf{z}^\top \boldsymbol{\gamma} \rightarrow \infty$ in the cumulative model, the response probabilities are concentrated in the extreme categories $\pi_1(\mathbf{x}, \mathbf{z})$ and $\pi_k(\mathbf{x}, \mathbf{z})$ indicating strong dispersion; if $\mathbf{z}^\top \boldsymbol{\gamma} \rightarrow -\infty$ the probability becomes one for one of the response categories indicating low dispersion.

The location-shift model avoids the multiplicative structure by postulating

$$g_r(\boldsymbol{\pi}(\mathbf{x}, \mathbf{z})) = F\left(\beta_{r0} + \mathbf{x}^\top \boldsymbol{\beta} + (k/2 - r)\mathbf{z}^\top \boldsymbol{\gamma}\right), \quad r = 1, \dots, k-1,$$

where the weight $(k/2 - r)$ scales the linear term $\mathbf{z}^\top \boldsymbol{\gamma}$ such that the difference between adjacent linear predictors $\eta_r = \beta_{r0} + \mathbf{x}^\top \boldsymbol{\beta} + (k/2 - r)\mathbf{z}^\top \boldsymbol{\gamma}$ becomes $\eta_r - \eta_{r-1} = \beta_{r0} - \beta_{r-1,0} - \mathbf{z}^\top \boldsymbol{\gamma}$. Thus the difference between adjacent predictors is widened or shrunk by $\mathbf{z}^\top \boldsymbol{\gamma}$ depending on the sign. Its value determines whether there is more concentration in middle or extreme categories, which indicates dispersion effects; details are given in [62, 63], and an overview on ordinal models was given by [61]. Both the location-scale and the location-shift models simplify to the ordinal models with global parameters if $\boldsymbol{\gamma} = \mathbf{0}$.

Variable selection now refers to two sources of variation, namely location and dispersion effects, which suggests a penalty of the form

$$J(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \lambda \sum_{j=1}^p |\beta_j| + \zeta \sum_{j=1}^m |\gamma_j|,$$

where λ and ζ are smoothing parameters, and \mathbf{z} is a vector of dimension m . If $\zeta \rightarrow \infty$, models with global parameters are fitted. Thus, the penalty also chooses between models with category-specific and those with global effects.

7.5.2.2 Nominal Response Models

A general model for nominal response categories with global and category-specific explanatory variables is the multinomial logit model

$$\pi_r = P(y = r|\mathbf{x}) = \frac{\exp(\beta_{r0} + \mathbf{x}^\top \boldsymbol{\beta}_r + (\mathbf{w}_r - \mathbf{w}_k)^\top \boldsymbol{\alpha})}{\sum_{s=1}^k \exp(\beta_{s0} + \mathbf{x}^\top \boldsymbol{\beta}_s + (\mathbf{w}_s - \mathbf{w}_k)^\top \boldsymbol{\alpha})}, \quad (7.14)$$

which can also be given in the form

$$\log\left(\frac{\pi_r}{\pi_s}\right) = \beta_{r0} - \beta_{s0} + \mathbf{x}^\top(\boldsymbol{\beta}_r - \boldsymbol{\beta}_s) + (\mathbf{w}_r - \mathbf{w}_s)^\top \boldsymbol{\alpha},$$

The vector \mathbf{x} contains all the variables that do not vary across categories; for example, in choice experiments variables like gender and age. The vectors $\mathbf{w}_1, \dots, \mathbf{w}_k$ represent vectors that vary across categories, for example the price in the choice of travel mode. In the model, the difference between prices, with category k as the reference category, determine the response probabilities. If \mathbf{w}_r are constant across all categories one obtains the classical multinomial logit model.

With $\boldsymbol{\beta}_r^\top = (\beta_{r1}, \dots, \beta_{rp})$ and the parameters referring to the variable x_j collected in $\boldsymbol{\beta}_{\cdot j}^\top = (\beta_{1j}, \dots, \beta_{k-1,j})$ a penalty that enforces selection of variables is

$$J(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \lambda \sum_{j=1}^p \|\boldsymbol{\beta}_{\cdot j}\|_2 + \zeta \sum_{j=1}^m |\alpha_j|,$$

where $\boldsymbol{\alpha}^\top = (\alpha_1, \dots, \alpha_m)$. Details are given in [67], and an extended version is given in [38].

7.6 Concluding Remarks

Structuring and selection of explanatory variables in regression has been considered within the framework of generalized additive models. In particular, penalty-based methods for ordinal predictors have been compared to stepwise procedures to investigate the merits of the different approaches. In general, the penalty terms that are to be used depend on the scaling of the explanatory variables, nominal variables call for other penalty terms than ordinal variables. Although we restricted ourselves to the framework of generalized additive models here, at least some of the penalties discussed could also be used in an extended framework such as *quasi-likelihood* methods. This is particularly true for the quadratic smoothing penalties implemented for use within `mgcv`, since the latter also supports quasi-likelihood (and further extensions); see [78] for details.

Stepwise selection, potentially combined with quadratic smoothing penalties for ordinal predictors as considered in detail here, works well in moderate settings with not too many categories and variables. For high-dimensional problems with a very large number of categories and/or covariates, appropriately designed, sparsity-inducing penalties and tree-based approaches should be more efficient.

While the focus of this paper was on categorical, in particular ordinal, predictors, we also briefly considered multi-categorical response models, where some ideas from categorical predictors can be borrowed. A topic of future research could be the

combination of both. This appears particularly relevant for, e.g., questionnaires with many ordinally scaled items where some items may serve as independent and some as dependent variables. Taking one step further, we may also try to describe and analyze dependence within a large set of ordinal variables by appropriately designed graphical models or structural equations. Furthermore, ordinal data are found in settings such as time series and functional data. Generally speaking, a very popular approach for handling ordinal data is to assume a latent continuous variable, with ordinal data being obtained via (latent) thresholds. This has, for instance, already been done for ordinal functional data [71] and structural equation modeling [69]. Assuming a latent continuous variable, however, is not always reasonable. For example, some measures of location, dispersion, symmetry, and (serial) dependence for the case of ordinal (time series) data which do not require a latent continuous variable are found in [73]. Although some of the ordinal regression models given in Sect. 7.5, such as the proportional odds model, can also be motivated as a latent variable approach, many others cannot, and we have not made this assumption at any time in this chapter. For many other settings with ordinal data, however, methods that do not rely on a latent variable assumption are still needed. Eventually, of course, it must be decided by the data analyst whether the latent variable approach makes sense or not; and the greater the number of available options that are tailored to specific situations the better.

Acknowledgments We would like to thank two anonymous reviewers for their encouraging and very constructive comments which helped us to improve the initial version of this chapter. Furthermore, support from Deutsche Forschungsgemeinschaft (DFG) through grant GE2353/2-1 is gratefully acknowledged.

References

1. Berger, M.: *structree: Tree-Structured Clustering*. R package version 1.1.7 (2020)
2. Bondell, H., Reich, B.: Simultaneous factor selection and collapsing levels in ANOVA. *Biometrics* **65**, 169–177 (2009)
3. Brant, R.: Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics* **46**, 1171–1178 (1990)
4. Breslow, N.E., Clayton, D.G.: Approximate inference in generalized linear mixed models. *J. Am. Stat. Assoc.* **88**, 9–25 (1993)
5. Bühlmann, P., Gertheiss, J., Hieke, S., Kneib, T., Ma, S., Schumacher, M., Tutz, G., Wang, C.-Y., Wang, Z., Ziegler, A.: Discussion of “the evolution of boosting algorithms” and “extending statistical boosting”. *Methods Inf. Med.* **53**, 436–445 (2014)
6. Chiquet, J., Gutierrez, P., Rigai, G.: Fast tree inference with weighted fusion penalties. *J. Comput. Graph. Stat.* **26**, 205–216 (2017)
7. Cieza, A., Ewert, T., Berdirhan Üstün, T., Chatterji, S., Kostanjsek, N., Stucki, G.: Development of ICF Core Sets for patients with chronic conditions. *J. Rehabil. Med. Suppl.* **44**, 9–11 (2004)
8. Cieza, A., Stucki, G., Weigl, M., Kullmann, L., Stoll, T., Kamen, L., Kostanjsek, N., Walsh, N.: ICF Core Sets for chronic widespread pain. *J. Rehabil. Med. Suppl.* **44**, 63–68 (2004)

9. Crainiceanu, C.M., Ruppert, D.: Likelihood ratio tests in linear mixed models with one variance component. *J. R. Stat. Soc. B* **66**, 165–185 (2004)
10. Crainiceanu, C.M., Ruppert, D., Claeskens, G., Wand, M.P.: Exact likelihood ratio tests for penalised splines. *Biometrika* **92**, 91–103 (2005)
11. de Boor, C.: *A Practical Guide to Splines*. Springer, New York (1978)
12. Detmer, F.J., Cebal, J., Slawski, M.: A note on coding and standardization of categorical variables in (sparse) group lasso regression. *J. Stat. Plan. Infer.* **206**, 1–11 (2020)
13. Dierckx, P.: *Curve and Surface Fitting with Splines*. Clarendon Press, Oxford (1993)
14. Eilers, P.H.C., Marx, B.D.: Flexible smoothing with B-splines and penalties. *Stat. Sci.* **11**, 89–121 (1996)
15. Fahrmeir, L., Kneib, T., Lang, S., Marx, B.: *Regression—Models, Methods and Applications*. Springer, Berlin (2013)
16. Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **96**, 1348–1360 (2001)
17. Feng, Y., Yu, Y.: The restricted consistency property of leave- n_v -out cross-validation for high-dimensional variable selection. *Stat. Sin.* **29**, 1607–1630 (2019)
18. Gertheiss, J.: ANOVA for factors with ordered levels. *J. Agric. Biol. Environ. Stat.* **19**, 258–277 (2014)
19. Gertheiss, J., Hoshiyar, A.: *ordPens: Selection, Fusion, Smoothing and Principal Components Analysis for Ordinal Variables*. R package version 1.0.0 (2021)
20. Gertheiss, J., Oehrlin, F.: Testing relevance and linearity of ordinal predictors. *Electron. J. Stat.* **5**, 1935–1959 (2011)
21. Gertheiss, J., Tutz, G.: Penalized regression with ordinal predictors. *Int. Stat. Rev.* **77**, 345–365 (2009)
22. Gertheiss, J., Tutz, G.: Sparse modeling of categorical explanatory variables. *Ann. Appl. Stat.* **4**, 2150–2180 (2010)
23. Gertheiss, J., Hogger, S., Oberhauser, C., Tutz, G.: Selection of ordinally scaled independent variables with applications to International Classification of Functioning core sets. *J. R. Stat. Soc. C* **60**, 377–395 (2011)
24. Gertheiss, J., Scheipl, F., Lauer, T., Ehrhardt, H.: Statistical inference for ordinal predictors in generalized linear and additive models with application to bronchopulmonary dysplasia. Preprint (2021). Available at <https://arxiv.org/abs/2102.01946>
25. Greven, S., Kneib, T.: On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika* **97**, 773–789 (2010)
26. Greven, S., Crainiceanu, C., Küchenhoff, H., Peters, A.: Restricted likelihood ratio testing for zero variance components in linear mixed models. *J. Comput. Graph. Stat.* **17**, 870–891 (2008)
27. Harville, D.A.: Bayesian inference for variance components using only error contrasts. *Biometrika* **61**, 383–385 (1974)
28. Harville, D.A.: Maximum likelihood approaches to variance component estimation and to related problems. *J. Am. Stat. Assoc.* **72**, 320–338 (1977)
29. Hastie, T., Tibshirani, R.: *Generalized Additive Models*. Chapman & Hall, London (1990)
30. Hofner, B., Hothorn, T., Kneib, T., Schmid, M.: A framework for unbiased model selection based on boosting. *J. Comput. Graph. Stat.* **20**, 956–971 (2011)
31. Huang, J., Breheny, P., Ma, S.: A selective review of group selection in high-dimensional models. *Stat. Sci.* **27**, 481–499 (2012)
32. Huang, L., Hang, W., Chao, Y.: High-dimensional regression with ordered multiple categorical predictors. *Stat. Med.* **39**, 294–309 (2020)
33. Laird, N.M., Ware, J.H.: Random-effects models for longitudinal data. *Biometrics* **38**, 963–974 (1982)
34. Leeb, H., Pötscher, B.M.: Model selection and inference: facts and fiction. *Economet. Theor.* **21**, 21–59 (2005)
35. Ma, S., Huang, J.: A concave pairwise fusion approach to subgroup analysis. *J. Am. Stat. Assoc.* **112**, 410–423 (2017)

36. Malsiner-Walli, G., Pauer, D., Wagner, H.: Effect fusion using model-based clustering. *Stat. Model.* **18**, 175–196 (2018)
37. Marra, G., Wood, S.N.: Coverage properties of confidence intervals for generalized additive model components. *Scand. J. Stat.* **39**, 53–74 (2012)
38. Mauerer, I., Pössnecker, W., Thurner, P., Tutz, G.: Modeling electoral choices in multiparty systems with high-dimensional data: a regularized selection of parameters using the Lasso approach. *J. Choice Model.* **16**, 23–42 (2015)
39. Mayr, A., Binder, H., Gefeller, O., Schmid, M.: Extending statistical boosting—an overview of recent methodological developments. *Methods Inf. Med.* **53**, 428–435 (2014)
40. McCullagh, P.: Regression model for ordinal data (with discussion). *J. R. Stat. Soc. B* **42**, 109–127 (1980)
41. McCullagh, P., Nelder, J.A.: *Generalized Linear Models*, 2nd edn. Chapman & Hall, New York (1989)
42. McHorney, C.A., Ware, J.E., Raczek, A.E.: The MOS 36-item short-form health survey (SF-36): II. psychometric and clinical tests of validity in measuring physical and mental health constructs. *Med. Care* **31**, 247–263 (1993)
43. Meinshausen, N.: Relaxed lasso. *Comput. Stat. Data Anal.* **52**, 374–393 (2007)
44. Meinshausen, N., Bühlmann, P.: Stability selection. *J. R. Stat. Soc. B* **72**, 417–473 (2010)
45. Nelder, J.A., Wedderburn, R.W.M.: *Generalized linear models*. *J. R. Stat. Soc. A* **135**, 370–384 (1972)
46. Nychka, D.: Bayesian confidence intervals of smoothing splines. *J. Am. Stat. Assoc.* **83**, 1134–1143 (1988)
47. Oelker, M.-R., Pössnecker, W., Tutz, G.: Selection and fusion of categorical predictors with L0-type penalties. *Stat. Model.* **15**, 389–410 (2015)
48. Patterson, H.D., Thompson, R.: Recovery of interblock information when block sizes are unequal. *Biometrika* **58**, 545–554 (1971)
49. Pauer, D., Wagner, H.: Bayesian effect fusion for categorical predictors. *Bayesian Anal.* **14**, 341–369 (2019)
50. Pössnecker, W., Tutz, G.: A general framework for the selection of effect type in ordinal regression. Technical Report 186, Department of Statistics LMU (2016)
51. R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna (2020). <https://www.R-project.org/>
52. Scheipl, F., Greven, S., Küchenhoff, H.: Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Comput. Stat. Data Anal.* **52**, 3283–3299 (2008)
53. Scheipl, F., Fahrmeir, L., Kneib, T.: Spike-and-slab priors for function selection in structured additive regression models. *J. Am. Stat. Assoc.* **500**, 1518–1532 (2012)
54. Simon, N., Tibshirani, R.: Standardization and the group lasso penalty. *Stat. Sin.* **22**, 983–1001 (2012)
55. Stokell, B.: *CatReg: Solution Paths for Linear and Logistic Regression Models with SCOPE Penalty*. R package version 2.0.1. (2020)
56. Stokell, B.G., Shah, R.D., Tibshirani, R.J.: Modelling high-dimensional categorical data using nonconvex fusion penalties. *J. R. Stat. Soc. B* **83**, 579–611 (2021)
57. Sweeney, E., Crainiceanu, C., Gertheiss, J.: Testing differentially expressed genes in dose-response studies and with ordinal phenotypes. *Stat. Appl. Genet. Mol. Biol.* **15**, 213–235 (2016)
58. Tibshirani, R.: Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* **58**, 267–288 (1996)
59. Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K.: Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. B* **67**, 91–108 (2005)
60. Tutz, G.: *Regression for Categorical Data*. Cambridge University Press, Cambridge (2012)
61. Tutz, G.: Ordinal regression: a review and a taxonomy of models. *WIREs Comput. Stat.* **14**, e1545 (2022)

62. Tutz, G., Berger, M.: Response styles in rating scales – simultaneous modelling of content-related effects and the tendency to middle or extreme categories. *J. Educ. Behav. Stat.* **41**, 239–268 (2016)
63. Tutz, G., Berger, M.: Separating location and dispersion in ordinal regression models. *Eco. Stat.* **2**, 131–148 (2017)
64. Tutz G., Berger, M.: Tree-structured modelling of categorical predictors in generalized additive regression. *Adv. Data Anal. Classif.* **12**, 737–758 (2018)
65. Tutz, G., Gertheiss, J.: Rating scales as predictors – the old question of scale level and some answers. *Psychometrika* **79**, 357–736 (2014)
66. Tutz, G., Gertheiss, J.: Regularized regression for categorical data (with discussion and rejoinder). *Stat. Model.* **16**, 161–260 (2016)
67. Tutz, G., Pössnecker, W., Uhlmann, L.: Variable selection in general multinomial logit models. *Comput. Stat. Data Anal.* **82**, 207–222 (2015)
68. Ugba, E.R., Mörlin, D., Gertheiss, J.: Smoothing in ordinal regression: an application to sensory data. *Stats* **4**, 616–633 (2021)
69. Vegelius, J. Jin, S.: A semiparametric approach for structural equation modeling with ordinal data. *Struct. Equ. Model. Multidiscip. J.* **28**, 497–505 (2021)
70. Walter, S.D., Feinstein, A.R., Wells, C.K.: Coding ordinal independent variables in multiple regression analyses. *Am. J. Epidemiol.* **125**, 319–323 (1987)
71. Wang, B., Shi, J.Q.: Generalized gaussian process regression model for non-gaussian functional data. *J. Am. Stat. Assoc.* **109**, 1123–1133 (2014)
72. Ware, J.E., Sherbourne, C.: The MOS 36-item short-form health survey (SF-36): I. conceptual framework and item selection. *Med. Care* **30**, 473–483 (1992)
73. Weiß, C.H.: Distance-based analysis of ordinal data and ordinal time series. *J. Am. Stat. Assoc.* **115**, 1189–1200 (2020)
74. WHO: International Classification of Functioning, Disability and Health: ICF. World Health Organization, Geneva (2001)
75. Wood, S.N.: Fast stable direct fitting and smoothness selection for generalized additive models. *J. R. Stat. Soc. B* **70**, 495–518 (2008)
76. Wood, S.N.: Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc. B* **73**, 3–36 (2011)
77. Wood, S.N.: On p-values for smooth components of an extended generalized additive model. *Biometrika* **100**, 221–228 (2013)
78. Wood, S.N.: *Generalized Additive Models: An Introduction with R*, 2nd edn. CRC Press, Boca Raton (2017)
79. Wood, S.N., Pya, N., Saefken, B.: Smoothing parameter and model selection for general smooth models (with discussion). *J. Am. Stat. Assoc.* **111**, 1548–1575 (2016)
80. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. B* **68**, 49–67 (2006)
81. Zhang, C.-H.: Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.* **38**, 894–942 (2010)
82. Zou, H.: The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* **101**, 1418–1429 (2006)

Chapter 8

An Overview of ARMA-Like Models for Count and Binary Data



Mirko Armillotta, Alessandra Luati, and Monia Lupparelli

8.1 Introduction

Traditionally, time series modelling has been mostly applied to data that are continuously valued. From the early specifications of [59] and [56], to the formalization by Box and Jenkins [8, 9], autoregressive (AR) and moving average (MA) models have been regularly applied in many fields, from finance to energy and neural networks, see for example [39, 57] and [49]. Non-linear models, such as the generalized autoregressive conditional heteroscedastic models [7, 26] or the threshold and smooth transition models [53, 54], up to the class of score driven models [16, 36], are essentially grounded on autoregressive dynamics. Though often employed regardless of the discrete nature of the data generating process, continuous models cannot adequately describe the dynamic trend of count or binary data. Notable examples where ad hoc models for discrete data are required include the number of clicks on a website and the daily counts of people infected with a rare disease or, as far as binary data are concerned, the presence or absence of an edge in a random network system and the success or failure of an industrial process.

Despite some relevant instances that we aim to discuss in this chapter, ARMA models for discrete valued time series have not enjoyed the same popularity of linear models for continuous time series. One of the reasons certainly lies in the fact that linear processes are related to second order stationarity, which fully characterizes Gaussian time series, while for discrete or count data the concept

M. Armillotta · A. Luati

Department of Statistical Sciences, University of Bologna, Bologna, Italy
e-mail: mirko.armillotta2@unibo.it; alessandra.luati@unibo.it

M. Lupparelli (✉)

Department of Statistics, Computer Science, Applications, University of Florence, Florence, Italy
e-mail: monia.lupparelli@unifi.it

© Springer Nature Switzerland AG 2023

M. Kateri, I. Moustaki (eds.), *Trends and Challenges in Categorical Data Analysis*,
Statistics for Social and Behavioral Sciences,
https://doi.org/10.1007/978-3-031-31186-4_8

233

of autocovariance needs to be adapted [50]. Moreover, the Wold representation, which allows every covariance-stationary process to be written as the sum of two processes, one deterministic and one stochastic, has no direct interpretation in the integer-valued case, see [21]. As a matter of fact, modelling discrete valued time series entails challenging aspects which are directly related to the nature of the random generating process.

In recent years, the interest in the analysis of discrete dynamic data has been considerably increasing. A useful classification of time series models in two main families is due to [15], who distinguished between observation driven models [61] and parameter driven ones [60]. In parameter driven models, two different time series processes are the object of inference: the process generating the observed data, say $\{Y_t\}_{t \in \mathbb{Z}}$, and an unobservable, latent, process $\{\xi_t\}_{t \in \mathbb{Z}}$ which presents a dynamic formulation and carries a stochastic error term $\{e_t\}_{t \in \mathbb{Z}}$. Observation driven models, on the other hand, are fully described by the observed time series coming from the process $\{Y_t\}_{t \in \mathbb{Z}}$, since the latent process $\{\xi_t\}_{t \in \mathbb{Z}}$ is simply defined as a deterministic function of the past history of Y_t . For example, the well known generalized autoregressive conditional heteroskedastic (GARCH) model introduced by Bollerslev [7] generalizing [26], is an observation driven model. Indeed, in a GARCH model, the unobservable latent process is the conditional variance σ_t^2 and is just defined as a non linear function of past values of the observation process $\{Y_t\}_{t \in \mathbb{Z}}$. Specifically, to complete the example, in a GARCH(1,1) model one has that $\sigma_t^2 = \omega + \alpha y_{t-1}^2 + \beta \sigma_{t-1}^2$ with $\omega > 0, \alpha > 0, \beta > 0$. Conversely, in one of the most popular parameter driven models for the latent variance, the so called Stochastic Volatility model, [52], the latent process is modelled as $a_t = \alpha + \beta a_{t-1} + e_t$, where $a_t = 2 \log \sigma_t$ and e_t is an independent and identically distributed sequence of Gaussian random variables, in short $e_t \sim IIDN(0, \nu)$. Since for parameter driven models estimation can be difficult as it is often not possible to compute the associated likelihood in closed form, observation driven models are sometimes preferred. For this reason, in the present contribution, the focus will be on observation-driven models.

Contributions related with observation driven models for discrete-valued time series include the works by Davis et al. [19], Benjamin et al. [5] and Ferland et al. [27], among others. With the focus on the dynamic trend of count data, recent contributions can be envisaged in the works of [1, 18, 40, 47] and [14] and [34]. A different stream of the literature concerned with the development of models for time dependent count data is constituted by Integer Autoregressive models (INAR) [2, 3]. The latter are also categorized as observation-driven models and an ARMA version (INARMA) has been recently discussed in [58]. However, such models rely on the so-called generalized Steutel and van Harn thinning operator [51] and have a completely different methodological treatment with respect to the models that we shall discuss in the present chapter. For this reason, we shall not cover the theory of integer autoregressive models. A comprehensive review of these models can be found in [48].

The aim of this chapter is to provide a comprehensive overview of the literature on observation driven models for discrete valued time series, with a special focus

on count and binary data. In particular, stochastic properties and estimation are discussed for notable ARMA-like models, such as BARMA [41], GARMA [5], GLARMA [19], M-GARMA [62], and log-linear Poisson [30] models. These models are generally referred as ARMA-like models as they are designed to account for the direction and the magnitude of three relevant effects in the analysis of temporal data. More precisely, ARMA-like models may include an autoregressive-like effect, a moving average type effect and the dependence with respect to the past predictions of the random process. The specification for these effects eventually depends on a suitable link function which is selected according to the probabilistic assumption underlying the data generating process.

The stochastic properties of discrete ARMA models can be derived by following two different methods, one based on the theory of Markov chains and the other on the perturbation approach. The latter developed by Fokianos et al. [30] is based on the analysis of a modified version of the discrete process, which allows one to derive properties of the original processes. An alternative method, based on Markov chain theory without irreducibility assumptions, has been considered by Matteson et al. [42] and Douc et al. [23]. This approach leads to obtaining probabilistic properties of the discrete variable by defining the latent process as a Markov chain of order one. To illustrate these methods, an example for the GARMA model is given, taken from [42]. An application to log-linear Poisson autoregression provided by Douc et al. [23] is reported, as well.

As far as inference is concerned, the properties of the maximum likelihood estimator (MLE) and Quasi MLE (QMLE) have been widely studied for discrete-valued models; see [18, 23], and [1], among others. Specifically, the use of the generalized linear model (GLM) of [43] for dynamic discrete data provides a natural extension of continuous-valued time series to integer-valued processes. Then, inference based on likelihood theory can be acquired directly from the GLM framework, as well as principles for hypothesis testing and model diagnostics. For the case of misspecified models, results related to quasi-likelihood inference are also illustrated, together with the conditions required for strong, consistent, and asymptotic normality of QMLE, based on the work of [23] and [24]. Clearly, the exact likelihood inference and the asymptotic properties of the MLE are obtained as a special case.

To conclude the review, two applications of ARMA-like models are illustrated. The first illustration is concerned with the analysis of a time series related to the daily number of deaths from COVID-19 in Italy, from March to December 2020. The analysis is performed under the assumption of a Poisson and a negative binomial distribution for the data generating process. Model comparison is carried out by using penalized likelihood criteria. The second empirical analysis regards the binary series of signs of log-returns for the weekly closing prices of Johnson & Johnson, by using BARMA and Bernoulli GARMA and GLARMA models.

8.2 General Overview

Let us consider a stochastic process $\{Y_t\}_{t \in \mathbb{Z}}$, the information set of past observations of the process $\mathcal{F}_{t-1} = \sigma\{(\mathbf{X}_{s+1}, Y_s), s \leq t-1\}$ up to the time $t-1$ and a vector of covariates \mathbf{X}_t up to time t , where $\sigma\{X\}$ refers to the σ -field generated by the random variable X , defined as the smallest sigma-field with respect to which the variable is measurable. For the definition of the sigma-field see Billingsley [6, p. 19-20]. The corresponding realizations are denoted with the lower-case counterparts, y_t and \mathbf{x}_t , respectively. The focus, throughout the chapter, is on the case when $\{Y_t\}_{t \in \mathbb{N}}$ is discrete-valued. Suppose that the distribution of the process lies in the general class of the one-parameter exponential family,

$$q(Y_t | \mathcal{F}_{t-1}) = \exp\{Y_t f(\eta_t) - A(\eta_t) + d(Y_t)\}, \quad (8.1)$$

where the conditional expected value is defined as

$$\mu_t = E(Y_t | \mathcal{F}_{t-1}) = A'(\eta_t)$$

and $\eta_t = g(\mu_t)$ with $g(\cdot)$ a twice-differentiable, one-to-one monotonic function, which is called the link function, see [43].

In Eq. (8.1) it is assumed that the dynamics of the density (or mass) function $q(Y_t | \mathcal{F}_{t-1})$ are captured by the parameter μ_t , or equivalently η_t , called the linear predictor. The function $A(\cdot)$ (log-partition) and $d(\cdot)$ are specific functions which define the particular distribution of interest. In the framework of the exponential family of [43], $f(\eta_t)$ is the canonical parameter. The mapping $f(\cdot)$ is a twice-differentiable bijective function, chosen in accordance with to the model of interest. The conditional variance is

$$\sigma_t^2 = V(Y_t | \mathcal{F}_{t-1}) = A''(\eta_t) = v(\mu_t).$$

Example 8.1 In Eq. (8.1), the Poisson distribution is obtained by setting $f(\eta_t) = \eta_t$, $\eta_t = g(\mu_t) = \log(\mu_t)$, $A(\eta_t) = \exp(\eta_t) = \mu_t$ and $d(Y_t) = \log(1/Y_t!)$. The conditional expectation is then $E(Y_t | \mathcal{F}_{t-1}) = V(Y_t | \mathcal{F}_{t-1}) = \exp(\eta) = \mu_t$.

Clearly, since for the Poisson distribution the canonical parameter is $\eta_t = \log(\mu)$, see [43], one has $f(\eta_t) = \eta_t$.

Example 8.2 The Gaussian distribution (with known variance) is obtained by setting $f(\eta_t) = \eta_t$, $g(\mu_t) = \frac{\mu_t}{\sigma_t^2}$, $A[g(\mu_t)] = \frac{\mu_t^2}{2\sigma_t^2}$ and $d(Y_t) = \log\left[-\frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left(-\frac{Y_t^2}{2\sigma_t^2}\right)\right]$. One can verify that $\mu_t = \sigma_t^2 \eta_t$, so $A(\eta_t) = \sigma_t^2 \eta_t^2 / 2$, whose first and second derivatives are respectively μ_t and σ_t^2 .

It can be convenient to consider the following dynamic representation for the time varying conditional mean,

$$g(\mu_t) = \eta_t = \mathbf{x}_t^T \beta + z_t, \quad (8.2)$$

$$z_t = \sum_{j=1}^p \phi_j \left[h(Y_{t-j}) - \mathbf{x}_{t-j}^T \beta \right] + \sum_{j=1}^k \gamma_j (z_{t-j} + \epsilon_{t-j}) + \sum_{j=1}^q \theta_j \epsilon_{t-j}, \quad (8.3)$$

where p, k , and q are integers representing the maximum lag order of their respective additive terms, and ϵ_t , generally called *prediction error*, is defined in the following way:

$$\epsilon_t = \frac{h(Y_t) - \bar{g}(\mu_t)}{v_t} \quad (8.4)$$

and v_t is some scaling sequence, for example:

- $v_t = \sigma_t$, Pearson residuals
- $v_t = \sigma_t^2$, Score-type residuals
- $v_t = 1$, No scaling
- $v_t = \text{V}[h(y_t) | \mathcal{F}_{t-1}]$

where $\text{V}[h(y_t) | \mathcal{F}_{t-1}]$ is the variance of the function $h(Y_t)$, conditional to the past information \mathcal{F}_{t-1} .

The function $h(Y_t)$ is called the “data-link function” because it is applied to the observation process Y_t whereas $\bar{g}(\mu_t)$ is said to be “mean-link function” because it is applied only to the conditional mean, unlike the link function $g(\cdot)$ which, in principle, can be applied to any parameter or moment of the probability distribution. Both the functions $h(Y_t)$ and $\bar{g}(\mu_t)$ are twice-differentiable, one-to-one monotonic; their shape depends on the specific model (8.2)–(8.3) and the distribution of interest in Eq. (8.1). Note that the terminology “link function” generally refers to the specification of a function $g(\cdot)$ modelling the dependence between a transformation η_t of the conditional expected value μ_t and a linear predictor including information related to past values z_t or to a covariate set \mathbf{x}_t . The same terminology is adopted for the specification of functions $h(\cdot)$ and $\bar{g}(\cdot)$ since, in some instances belonging to the exponential family distribution, convenient choices for these functions correspond to the canonical link function. Nevertheless, $h(\cdot)$ and $\bar{g}(\cdot)$ might be different from $g(\cdot)$, so that the model (8.2)–(8.3) is able to encompass a wide range of existing models developed in the literature, as special cases. Some examples are presented in the next section.

Despite the fact that it is not constrained to assume a specific formulation, in general, it is useful to choose the mean-link function as follows:

$$\bar{g}(\mu_t) = \text{E}[h(Y_t) | \mathcal{F}_{t-1}], \quad (8.5)$$

in order to obtain $\epsilon_t \sim MDS$ (Martingale Difference Sequence), i.e. the difference $E[h(Y_t) - \bar{g}(\mu_t)|\mathcal{F}_{t-1}] = 0$. In general, a MDS has conditional expectation $E[\epsilon_t|\mathcal{F}_{t-1}] = 0$ and, as a consequence, unconditional expectation $E(\epsilon_t) = 0$. Moreover it is uncorrelated, i.e. $E(\epsilon_t\epsilon_{t-s}) = 0$, with $s \neq 0$. This is a really useful construct in probability theory because it does not require the usual assumption of independence of the errors. Furthermore, most limit theorems that hold for an independent sequence will also hold for an MDS.

Moreover, if $v_t = \sqrt{V[h(Y_t)|\mathcal{F}_{t-1}]}$, then the residuals in Eq. (8.4) form a white noise (WN) sequence, with unit variance. In practical situations, an explicit formula for the conditional moments $E[h(Y_t)|\mathcal{F}_{t-1}]$ and $V[h(Y_t)|\mathcal{F}_{t-1}]$ is not always available. In these cases, it seems reasonable to use an approximation constructed from their Taylor expansions; for example, the second order expansions are: $\bar{g}(\mu_t) = E[h(Y_t)|\mathcal{F}_{t-1}] \approx h(\mu_t) + \frac{1}{2}h''(\mu_t)\sigma_t^2$, $V[h(Y_t)|\mathcal{F}_{t-1}] = E[h(Y_t)^2|\mathcal{F}_{t-1}] - E[h(Y_t)|\mathcal{F}_{t-1}]^2 \approx m(\mu_t) + \frac{1}{2}m''(\mu_t)\sigma_t^2 - \bar{g}(\mu_t)^2$, where $m(\cdot) = h(\cdot)^2$.

We remind the reader that the process $\{Y_t\}_{t \in \mathbb{Z}}$ is observed whereas $\{\mu_t\}_{t \in \mathbb{Z}}$ is not. However, it can be shown by backward substitutions in (8.2)–(8.3), that the process $\{\mu_t\}_{t \in \mathbb{Z}}$ is a deterministic function of the past \mathcal{F}_{t-1} . This is the reason why Eqs. (8.2)–(8.3) belong to the class of “observation driven models”, see [15], where error terms are typically defined as MDS.

The parameters ϕ , θ , and γ in Eq. (8.3) model the direction and the magnitude of three relevant effects in the analysis of temporal data. First, the autoregressive-like effect, which represents the dependence on the past observations; then, the effect of the moving average part is considered for modelling the dependence between prediction error terms over time; finally, the effect of the past memory accounts for the dependence with respect to the past predictions rather than to the past observations. In some sense, the latter term can account for the dependence of the process from its whole past, since μ_t depends on all the past observations Y_{t-1}, Y_{t-2}, \dots . In principle, any effect can be specified in the model through different link functions. In practice, however, these functions are typically tailored to the nature of the data generating process.

8.3 Some Relevant Models

This section describes the most relevant models developed in the literature of ARMA-like time series for binary and count observations generated from probability distributions mainly belonging to the exponential family.

8.3.1 GARMA

A well-known specification for discrete-valued time series is the generalized autoregressive moving average model, GARMA, [5]. The distribution of the process is defined to be the one-parameter exponential family (8.1). From Eqs. (8.2)–(8.3) the GARMA model is obtained when $k = 0$, by setting $g \equiv \bar{g} \equiv h$, and $v_t = 1$, so that, the three link functions are equivalent and no scaling is applied:

$$\eta_t = \mathbf{x}_t^T \beta + \sum_{j=1}^p \phi_j \left[g(Y_{t-j}) - \mathbf{x}_{t-j}^T \beta \right] + \sum_{j=1}^q \theta_j \left[g(Y_{t-j}) - \eta_{t-j} \right]. \quad (8.6)$$

The model includes the autoregressive and the moving average effects by using the same link function g . The dependence on the past memory is not considered directly by a specific factor. This means that model (8.6) would be employed when the immediate past values of the observed process Y_{t-j} , $j = 1, \dots, \max(p, q)$ may be considered influential. In general, ϵ_t is not a Martingale difference sequence and then the mean-link function \bar{g} does not follows (8.5); instead, it is just set to be equivalent to g . However, there still is a special case in which $\epsilon_t \sim MDS$, such as $g \equiv h$: *identity* (see the M-GARMA model below).

Although this model is suitably applicable in practice to every distribution encompassed in (8.1), in the context of count data, it has been mainly used with a Poisson or a Negative Binomial (NB) distribution, see [5].

The estimation of the model (8.6) is usually performed by maximizing the log likelihood $L(\rho) = \sum_{t=1}^n \log q(Y_t | \mathcal{F}_{t-1})$, with respect to the associated vector of parameters $\rho = (\beta, \phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q)'$. This is done in practice through an iterated re-weighted least square (IRLS) procedure and the asymptotic normality of the estimator is established as $\sqrt{n}(\hat{\rho} - \rho) \sim N(0, I(\rho)^{-1})$ where $\hat{\rho}$ is the maximum likelihood estimator and $I(\rho)$ is the information matrix evaluated at the true parameter value, see [5, Sec. 3.1]. Under the Poisson distribution, strong consistency of the MLE is available [28, Theorem 3.1]. Section 8.3.4 provides further details about such results. More general results concerning the asymptotic properties of the Quasi MLE are introduced in Sect. 8.6.

8.3.2 M-GARMA

A suitable extension of the GARMA model in Eq. (8.6) has recently been introduced by Zheng et al. [62]; it allows the residuals ϵ_t to be a Martingale difference sequence and for this reason it has been called Martingalized GARMA (M-GARMA). It is obtained from (8.2)–(8.3) for $k = 0$, $g(\mu_t) = E[h(Y_t) | \mathcal{F}_{t-1}] = \bar{g}(\mu_t)$ and $v_t = 1$:

$$\bar{g}(\mu_t) = \mathbf{x}_t^T \beta + \sum_{j=1}^p \phi_j \left[h(Y_{t-j}) - \mathbf{x}_{t-j}^T \beta \right] + \sum_{j=1}^q \theta_j \left[h(Y_{t-j}) - \bar{g}(\mu_{t-j}) \right]. \quad (8.7)$$

For its particular construction, in this model, a crucial role is played by the data-link function h which would entirely determine the mean-link function. The usefulness of M-GARMA lies in the possibility of writing $h(Y_t)$ as a standard ARMA model simply by adding $h(Y_t) - \bar{g}(\mu_t)$ to both sides of (8.7) and rearranging the covariates:

$$h(Y_t) = \mathbf{x}_t^T \alpha + \sum_{j=1}^p \phi_j h(Y_{t-j}) + \epsilon_t + \sum_{j=1}^q \theta_j \epsilon_{t-j},$$

where $\alpha = \left(1 - \sum_{j=1}^p \phi B^j\right) \beta$ and B is the lag operator, such as $B^j \mathbf{x}_t = \mathbf{x}_{t-j}$. Note that when $\bar{g}(\mu_t) = E[h(y_t) | \mathcal{F}_{t-1}] = h(\mu_t)$, a GARMA model with the linear predictor $\eta_t = E[h(y_t) | \mathcal{F}_{t-1}]$ is obtained. Also, the use of the first-order Taylor approximation for $\bar{g}(\cdot)$ around μ_t provides

$$\bar{g}(\mu_t) = E[h(Y_t) | \mathcal{F}_{t-1}] \approx h(\mu_t).$$

Thus, the standard GARMA model has been derived as a particular case of the M-GARMA model when a linear approximation of \bar{g} is used. This leads to consideration of the application of the representation in Eq. (8.7), instead of the usual GARMA model (8.6), in all the cases when the expression $\bar{g}(\mu_t) = E[h(Y_t) | \mathcal{F}_{t-1}]$ has a closed form. This happens only for certain distributions, (such as Log-normal, Gamma, and Beta, among others) and suitable choices of the data-link function $h(\cdot)$. The interested reader can find an exhaustive treatment of these cases under Table 1 of [62].

The estimation of (8.7) has been performed in [62] with the Gaussian MLE, that is a maximum likelihood estimation performed by maximizing the Gaussian likelihood instead of the true likelihood of the model. Only the consistency of such an estimator is available. Asymptotic normality has been developed only under the special case $q = 0$.

8.3.3 GLARMA

A promising class has been developed by Rydberg and Shephard [47] and Davis et al. [19] under the name of generalized Linear Autoregressive Moving Average (GLARMA) models; Again, the distribution belongs to the exponential family (8.1). GLARMA models can be written based on Eqs. (8.2)–(8.3) by setting $p = 0$ and h : *identity*:

$$\begin{aligned} \eta_t &= \mathbf{x}_t^T \beta + z_t, \\ z_t &= \sum_{j=1}^k \gamma_j (z_{t-j} + \epsilon_{t-j}) + \sum_{j=1}^q \theta_j \epsilon_{t-j}, \end{aligned} \tag{8.8}$$

$$\epsilon_t = \frac{Y_t - \mu_t}{\nu_t}.$$

In these models, the error component and the past lag of the latent process are considered. However, the effect of past lags of the discrete process Y_t are not directly specified in the model. This means that model (8.8) would be suitable when the whole past memory of the observed process Y_{t-j} , $j = 1, 2, \dots$ may be influential. The usefulness of model (8.8) with a scaling sequence ν_t is considered when the discrete data are not constrained in a closed interval, that is when counts are modelled [25]. Notice that this model is equivalent to an ARMA model on the linear predictor (minus the constants and covariates):

$$\eta_t - \mathbf{x}_t^T \beta = z_t = \sum_{j=1}^k \gamma_j z_{t-j} + \sum_{j=1}^{\tilde{q}} \tau_j \epsilon_{t-j},$$

where $\tilde{q} = \max(k, q)$ and $\tau_j = \gamma_j + \theta_j$. Or alternatively, in terms of η_t , we have

$$\eta_t = \mathbf{x}_t^T \alpha + \sum_{j=1}^k \gamma_j \eta_{t-j} + \sum_{j=1}^{\tilde{q}} \tau_j \epsilon_{t-j},$$

where $\alpha = \left(1 - \sum_{j=1}^k \gamma_j B^j\right) \beta$.

For the Poisson distribution, the asymptotic normality of the MLE for the parameters $\rho = (\beta, \gamma_1, \dots, \gamma_k, \theta_1, \dots, \theta_q)'$ has been discussed in [20], namely $\sqrt{n}(\hat{\rho} - \rho) \xrightarrow{d} N(0, I(\rho)^{-1})$ where $\hat{\rho}$ is the maximum likelihood estimator and $I(\rho)$ is the information matrix evaluated at the true value of the parameters. For more general results see Sect. 8.6.

8.3.4 Poisson Autoregression

Poisson autoregression, henceforth Pois AR, introduced by Fokianos et al. [30], is obtained when (8.1) is $Pois(\mu_t)$, with $f(\eta_t) = \log(\eta_t)$, and in Eq. (8.2)–(8.3), one has $q = 0$ and $g \equiv h : \textit{identity}$:

$$\mu_t = \mathbf{x}_t^T \alpha + \sum_{j=1}^k \gamma_j \mu_{t-j} + \sum_{j=1}^p \phi_j Y_{t-j}. \quad (8.9)$$

Obviously, the parameters in Eq. (8.9) are constrained to the positive real line. A variant of (8.9) is the log-linear Poisson autoregression, henceforth Pois log-AR, [28] which is obtained when $q = 0$, $f(\eta_t) = \eta_t$, $g(\mu_t) = \log(\mu_t)$, and

$h(Y_t) = \log(Y_t + 1)$:

$$\log(\mu_t) = \mathbf{x}_t^T \boldsymbol{\alpha} + \sum_{j=1}^k \gamma_j \log(\mu_{t-j}) + \sum_{j=1}^p \phi_j \log(Y_{t-j} + 1). \tag{8.10}$$

The models (8.9) and (8.10) consider lagged effects for the discrete variable and the mean process explicitly and do not include an error component. However, note that, for Poisson data, the GARMA model (8.6) with identity or log links can be considered as a constrained Poisson autoregression where $\gamma_j = -\theta_j$ and ϕ_j is replaced by $\phi_j + \theta_j$, in Eqs. (8.9) or (8.10), so that the Poisson autoregression model can be rewritten in ARMA form.

The model in (8.10) could also be used for Negative Binomial data, by rewriting the distribution in terms of the expected value parameter μ_t , see [12]:

$$q(Y_t | \mathcal{F}_{t-1}) = \frac{\Gamma(v + Y_t)}{\Gamma(Y_t + 1)\Gamma(v)} \left(\frac{v}{v + \mu_t}\right)^v \left(\frac{\mu_t}{v + \mu_t}\right)^{Y_t} \tag{8.11}$$

where v is the dispersion parameter (if an integer, it is also known as the number of failures) and the usual probability parameter would be $p_t = \frac{v}{v + \mu_t}$. The distribution (8.11) with model (8.10) is obtained from the distribution (8.1), by setting the non-canonical link $g(\mu_t) = \log(\mu_t)$ and $f(\eta_t) = \eta_t - \log(v + e^{\eta_t})$, with $A(\eta_t) = -v \log\left(\frac{v}{v + e^{\eta_t}}\right)$ and $d(Y_t) = \log\frac{\Gamma(v + Y_t)}{\Gamma(Y_t + 1)\Gamma(v)}$.

Consistency and asymptotic normality of the MLE for the parameters $\rho = (\alpha, \gamma_1, \dots, \gamma_k, \phi_1, \dots, \phi_p)'$ has been established in [30] for the linear model (8.9) and in [28] for the log-linear model (8.10), that is $\sqrt{n}(\hat{\rho} - \rho) \xrightarrow{d} N(0, I(\rho)^{-1})$. The same properties have been discussed in [12], for the Poisson Quasi MLE applied under the Negative binomial distribution (8.11), with limiting covariance matrix $J(\rho)^{-1}I(\rho)J(\rho)^{-1}$, where $J(\rho)$ is the associated Hessian matrix of the Poisson quasi log-likelihood. See also Theorem 8.6 for details about limiting covariance matrices in a quasi-likelihood framework.

8.3.5 BARMA

In case of dynamic binary data, a relevant model is the Binomial ARMA (BARMA) model [41, 50] which is obtained when (8.1) is $Bin(a, \mu_t)$, where the number of trials a is known, and the probability parameter is $p_t = \mu_t/a$. By setting $k = 0$, h : identity and $v_t = 1$ in (8.2)–(8.3), we have

$$\eta_t = \mathbf{x}_t^T \boldsymbol{\beta} + \sum_{j=1}^p \phi_j \left[Y_{t-j} - \mathbf{x}_{t-j}^T \boldsymbol{\beta} \right] + \sum_{j=1}^q \theta_j \left[Y_{t-j} - \mu_{t-j} \right].$$

Note that, when $h : \textit{identity}$, the mean-link function in (8.5) automatically reduces to $E(Y_t | \mathcal{F}_{t-1}) = \mu_t$. Instead, the link function g can be any suitable function, typically logit or probit. This model is designed for Binomial distribution in (8.1). The BARMA model includes the autoregressive effect and the moving average part. The model could be also generalized to consider the dependence with respect to the long memory term with a suitable link function.

Consistency and asymptotic normality of the MLE for the parameters of the BARMA model, $\rho = (\beta, \phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q)'$, are available as special case of Moysiadis and Fokianos [45, Theorem 2], with limiting distribution $N(0, I(\rho)^{-1})$.

8.3.6 Discussion

Models for binary time series have not enjoyed the same developments as models for count data. However, enhancements in this direction could provide useful insights in several fields. The generalization to the non-binary case could be also interesting for the analysis of temporal categorical data. However, the distributions of these kinds of data require the estimation of multiple latent models, e.g. the categorical distribution with K levels needs latent processes $(p_{1,t}, \dots, p_{K,t})$ to be modelled simultaneously, leading to a multivariate time series model not encompassed in the distribution (8.1) and the model (8.2), (8.3). For this reason, the non-binary categorical data are not treated in this contribution. To the best of our knowledge this part of the literature seems to be barely explored; the interested reader can see [29] and [45] on suitable time series models for temporal categorical data.

8.4 Weak Stationarity

We pass on now to examine stationarity and ergodicity for some of the models highlighted in the previous section. A stochastic process $\{X_t\}_{t \in \mathbb{Z}}$, is strictly stationary if, for any integer k and for any ordered set of subscripts, t_1, t_2, \dots, t_k , the joint distribution of $(X_t, X_{t_1}, X_{t_2}, \dots, X_{t_k})$ depends only on $t_1 - t, t_2 - t, \dots, t_k - t$ but not on t . In particular, the distribution of X_t does not depend on the absolute position t . So the mean, variance, and other higher moments, if they exist, remain the same across t . Ergodicity is a property which ensures the sample mean (or moment) of the time series, or any measurable functions of it, converges asymptotically to the associated expectation. See Hayashi [37, p. 101] for details. However, the way to prove such properties strongly depends on the specific nature of the stochastic process itself. See Sect. 8.5 for a brief discussion.

A stochastic process $\{X_t\}_{t \in \mathbb{Z}}$, is said to be weakly stationary (or covariance stationary) if the first two moments are finite and are the same across t . It can be directly verified by checking that the functional form of the first two moments is not time dependent. If the first two moments are finite, then strict stationarity

implies weak stationarity. In this section, we consider weak stationarity conditions for GARMA, M-GARMA and GLARMA models. For the BARMA model, no direct results on weak stationarity are available in the literature so far. However, strong stationarity is proved for BARMA, see [45], which we shall consider in Sect. 8.5 along with the Poisson autoregression, derived by Fokianos et al. [30] and Fokianos and Tjøstheim [28].

8.4.1 GARMA

For the GARMA model in (8.6) for $g \equiv h : \text{identity}$, one has $\epsilon_t = Y_t - \mu_t$, with zero conditional and unconditional mean value. Moreover the process ϵ_t is uncorrelated. The observation process can be expressed in the form

$$Y_t = \mu_t + \epsilon_t. \quad (8.12)$$

By setting $w_t = Y_t - \mathbf{x}_t^T \beta$ and by replacing the expression of (8.6) in (8.12), a standard ARMA model is obtained:

$$w_t = \sum_{j=1}^p \phi_j w_{t-j} + \sum_{j=1}^q \theta_j \epsilon_{t-j} + \epsilon_t. \quad (8.13)$$

Of course (8.13) can be easily rearranged via polynomial notation in:

$$w_t = \Psi(B) \epsilon_t$$

where $\Psi(B) = 1 + \psi_1 B + \psi_2 B^2 + \dots = \Phi(B)^{-1} \Theta(B)$, $\Phi(B) = 1 - \phi_1 B - \dots - \phi_p B^p$, $\Theta(B) = 1 + \theta_1 B - \dots - \theta_q B^q$, and B is the lag operator; provided that $\Phi(B)$ is invertible, i.e. that $\Phi(z) \neq 0$, $\forall z \in \mathbb{C}$ such that $|z| < 1$, see [11]. Indeed, $E(w_t) = \Psi(B)$, $E(\epsilon_t) = 0$ and then $E(Y_t) = \beta$ in the case where $\mathbf{x}_t^T \beta = \beta$. The autocovariance does not depend on time t because of the uncorrelated ϵ_t . Concerning the variance, the situation is slightly more involved:

$$\begin{aligned} V(Y_t) &= V(\mathbf{x}_t^T \beta + w_t) \\ &= V(w_t) = E(\epsilon_t^2) \\ &= E[\Psi(B) \epsilon_t \Psi(B) \epsilon_t] \\ &= \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \psi_i \psi_j E(\epsilon_{t-i} \epsilon_{t-j}) \\ &= \sum_{i=0}^{\infty} \psi_i^2 E(\epsilon_{t-i}^2) \end{aligned}$$

$$= \varphi E \left[\Psi^{(2)}(B) v(\mu_t) \right], \quad (8.14)$$

where $\Psi^{(2)}(B) = 1 + \psi_1^2 B + \psi_2^2 B^2 + \dots$. Expression (8.14) is obtained remembering that $E(\epsilon_t^2) = V(\epsilon_t) = E[E(\epsilon_t^2 | \mathcal{F}_{t-1})] = E[v(\mu_t)]$. The expression of the unconditional variance for the mean can be found as follows: $V(Y_t) = V(\mu_t) + V(\epsilon_t)$ since ϵ_t and μ_t are uncorrelated. So,

$$V(\mu_t) = E \left\{ \left[\Psi^{(2)}(B) - 1 \right] v(\mu_t) \right\}.$$

The particular expression for $v(\mu_t)$ in (8.14) depends on the distribution under investigation from (8.1). For example, in case of Poisson distribution, $v(\mu_t) = \mu_t$ so that

$$V(Y_t) = \Psi^{(2)}(B) E(\mu_t) = \Psi^{(2)}(B) \beta = \Psi^{(2)}(1) \beta,$$

where $\Psi^{(2)}(1) = 1 + \psi_1^2 + \psi_2^2 + \dots = \sum_{j=1}^{\infty} \psi_j^2$; it can be seen that the variance is constant over t and no additional conditions are required for weak stationarity apart from the usual invertibility of $\Phi(B)$. For other distributions, further invertibility conditions could be required; for example, in the Bernoulli case, even $\Psi^{(2)}(B)$ needs to be invertible to assure stationarity. This proof is due to [5].

We remark that these conditions do not work for other link functions that are different from the identity; the reason is that, in general, the prediction error in (8.6) $\epsilon_t = h(Y_t) - \eta_t$ is not an MDS (apart from the special case $g \equiv h$: identity).

In order to develop an asymptotic theory for the maximum likelihood estimator much more attention has been paid to assessing strict stationarity and ergodicity for the GARMA model than to proving weak stationarity. For this reason, we will deal with these results in the following section.

8.4.2 M-GARMA

The M-GARMA model (8.7) allows the prediction error to be an MDS. However, the distribution of ϵ_t does depend on \mathcal{F}_{t-1} ; for this reason, [62] pointed out that, in general, the classical condition of invertibility for $\Phi(B)$ is not sufficient for the existence of a stationary distribution of the process $\{g(Y_t)\}_Z$. By using the theory of Markov chains, the authors showed that the standard invertibility condition holds only for the special cases in which $\bar{g}(\mu_t) = g(\mu_t) + c$, where c is some function which is constant with respect to μ_t ; the authors call these special cases the *canonical link functions* (a survey of these link functions is presented in [62]); for the other cases they provided only strict stationarity conditions. However, the authors required $q(y | \mathcal{F}_{t-1})$ to be positive everywhere (\mathbb{R}^+); this condition is not satisfied for discrete-valued observation process y_t . Thus, the latter results are valid

only for continuous distributions; indeed, in the paper, the attention of the authors is focused on Beta and Gamma distributions.

8.4.3 GLARMA

For the GLARMA models, according to [47] and Dunsmuir and Scott [25, eq. 11-12], an autoregressive representation is available

$$z_t = \sum_{j=1}^k \gamma_j z_{t-j} + \sum_{j=1}^q \theta_j \epsilon_{t-1-j} + \epsilon_{t-1},$$

which can be made equivalent to (8.8) by suitable redefinition of the degrees k and q and the autoregressive and moving average coefficients as shown in Dunsmuir and Scott [25, Sec. 3.4]. Then, similarly to the GARMA model, weak stationarity conditions follow immediately by rewriting the model as an $MA(\infty)$:

$$z_t = \Psi^*(B) \epsilon_t = \sum_{j=1}^{\infty} \psi_j^* \epsilon_{t-j},$$

where $\Psi^*(B) = 1 + \psi_1^* B + \psi_2^* B^2 + \dots = \Gamma(B)^{-1} \Theta(B) - 1$, $\Gamma(B) = 1 - \gamma_1 B - \dots - \gamma_k B^k$, and $\Theta(B) = 1 + \theta_1 B - \dots - \theta_q B^q$. The model is initialized at $z_t = 0$ and $\epsilon_t = 0$ for $t \leq 0$. In general, the process $\{\epsilon_t\}$ is an MDS and, in the special case in which Pearson residuals are chosen, it is a stationary $WN(0,1)$ and, automatically, z_t will be (weakly) stationary (and Y_t as well) under the usual stationarity and invertibility conditions: roots of $\Gamma(B)$ and $\Theta(B)$ lie all outside the unit circle on the complex plan). See [25] for details. Nevertheless, no results are available for strict stationarity apart from the simplest case when $k = 0$, $q = 1$; see [19, 25], and [18].

8.5 Strong Stationarity

Strong stationarity and ergodicity for the models discussed so far can be derived based on several different approaches, see [31] for a comprehensive introduction. We mainly consider two of them. One is the perturbation approach introduced by Fokianos et al. [30] and Fokianos and Tjøstheim [28], for the linear and log-linear Poisson autoregression models, respectively. The other is the Markov chain theory without irreducibility developed by Matteson et al. [42], by extending the perturbation argument with Feller properties. These authors showed an application of their approach to the GARMA model as well, see Sect. 8.5.1. An alternative

approach to Markov chain theory without irreducibility assumption is presented by Douc et al. [23]. In this latter paper, an application to the log-linear Poisson autoregression is available, see Sect. 8.5.2. Similar results are established on the BARMA model, see [45]. For the M-GARMA model, only results for continuous variables are available by Zheng et al. [62]. For the GLARMA model, no direct strict-stationarity results have been developed in the literature.

The perturbation approach is an indirect way to establish the stability properties of the discrete process $\{Y_t\}$ and consists of defining a real-valued version of the process, by adding a small real perturbation m to the original process and then showing stochastic properties on the new perturbed process $\{Y_t^{(m)}\}$. Moreover, it can be proved that, as $m \rightarrow 0$, the two processes are arbitrarily close, Appendix Section “Perturbation Approach” provides details. The Markov chain theory without irreducibility allows one to extend results of the perturbation approach to the original process, by exploiting the fact that $\{\mu_t\}$ can be interpreted as a Markov chain. Showing stationarity and ergodicity for such a chain allows one to draw conclusions on the strict stationarity of the integer-valued process $\{Y_t\}$. The difference in this approach between [42] and [23] lies only in the additional assumptions required.

We first report an application of the perturbation approach and its extension with Feller properties to the GARMA model in Sect. 8.5.1. Then, an example of the approach of [23] to the log-linear Poisson autoregression is presented in Sect. 8.5.2. We postpone all the theoretical tools required for the application of the two methods in Appendix Section “Technical Details”.

8.5.1 Strict Stationarity and Ergodicity for the GARMA Model

In this section, the conditions under which there exists a strict-sense stationary and ergodic version of the observation process $\{Y_t\}_{t \in \mathbb{N}}$ for the GARMA(1,1) model are given. Define

$$Y_t | Y_{0:t-1} \sim q(\mu_t), \quad (8.15)$$

$$g(\mu_t) = \beta + \phi [g(Y_{t-1}^*) - \beta] + \theta [g(Y_{t-1}^*) - g(\mu_{t-1})] \quad (8.16)$$

where Y_t^* is a function which maps the value of Y_t to the domain of g . The process $Y_{0:t-1}$ is the set of past values of Y_t from the time 0 until $t - 1$; $q(\mu_t)$ is a synthetic notation for (8.1). Three separate cases are considered:

1. $q(\mu)$ is defined for any $\mu \in \mathbb{R}$. In this case, the domain of g is \mathbb{R} and $Y_t^* = Y_t$ is taken.
2. $q(\mu)$ is defined for only $\mu \in \mathbb{R}^+$ (or μ on any one-sided open interval by analogy). In this case, the domain of g is \mathbb{R}^+ and $Y_t^* = \max\{Y_t, c\}$ for some $c > 0$ is taken.

3. $q(\mu)$ is defined for only $\mu \in (0, a)$ where $a > 0$ (or any bounded open interval by analogy). In this case, the domain of g is $(0, a)$ and $Y_t^* = \min \{ \max(Y_t, c), (a - c) \}$ for some $c \in (0, a/2)$ is taken.

Valid link functions g are bijective and monotonic. Choices for Case 2 include the log link, which is the most commonly used, and the link, parametrized by $\alpha > 0$,

$$g(\mu) = \log(e^{\alpha \mu} - 1)/\alpha$$

which has the property that $g(\mu) \approx \mu$ for large μ . Examples of valid link functions for Cases 1 and 3 are the identity and logit functions, respectively. Note that model (8.15) is more general than the class of models developed in (8.1) in the sense that it is not necessarily assumed that $q(\cdot)$ belongs to the exponential family.

8.5.1.1 Perturbed Model

The perturbation approach consists of adding a small real-valued perturbation to the discrete-valued time series model in order to obtain a φ -irreducible process (see Definition 8.1 in Appendix Section “Technical Details”); then the standard tools for Markov chains (Appendix Section “Markov Chain Specification”) could be used to assess stationarity and ergodicity for the perturbed version of the GARMA model. First, ergodicity and stationarity results for the following perturbed model are obtained:

$$Y_t^{(m)} \mid Y_{0:t-1}^{(m)} \sim q(\mu_t^{(m)})$$

$$g(\mu_t^{(m)}) = \beta + \phi \left[g(Y_{t-1}^{(m)*}) - \beta \right] + \theta \left[g(Y_{t-1}^{(m)*}) - g(\mu_{t-1}^{(m)}) \right] + mZ_{t-1}, \quad (8.17)$$

where $Z_t \sim N(0, 1)$ are independent, identically distributed random perturbations, for any $m > 0$, which is a scale factor associated with the perturbation. The value $\mu_0^{(m)}$ is a fixed constant that is taken to be independent of m , so that $\mu_0^{(m)} = \mu_0$.

Theorem 8.1 *The process $\{\mu_t^{(m)}\}_{t \in \mathbb{N}}$ specified by the perturbed process (8.17) is an ergodic Markov chain and thus is stationary for an appropriate initial distribution for $\mu_0^{(m)}$, under the conditions below. This implies that the perturbed process $\{Y_t^{(m)}\}_{t \in \mathbb{N}}$ is stationary and ergodic when $\mu_0^{(m)}$ is initialized appropriately. The conditions are:*

1. $E(Y_t^{(m)} \mid \mu_t^{(m)}) = \mu_t^{(m)}$.
2. ($2 + \delta$ moment condition): There exist $\delta > 0$, $r \in [0, 1 + \delta)$ and non-negative constants d_1, d_2 such that

$$E(|Y_t^{(m)} - \mu_t^{(m)}|^{2+\delta} \mid \mu_t^{(m)}) \leq d_1 |\mu_t^{(m)}|^r + d_2.$$

3. g is bijective, increasing, and

- a. $g : \mathbb{R} \mapsto \mathbb{R}$ is concave on \mathbb{R}^+ and convex on \mathbb{R}^- , and $|\phi| < 1$
- b. $g : \mathbb{R}^+ \mapsto \mathbb{R}$ is concave on \mathbb{R}^+ , and $|\phi|, |\theta| < 1$
- c. $|\theta| < 1$; no additional conditions on $g : (0, a) \mapsto \mathbb{R}$.

The proof can be found in the appendix of [42]. This approach yields stationarity and ergodicity of the perturbed model. In order to extend these conclusions to the original unperturbed model the results of the following section are required.

8.5.1.2 Unperturbed Model

In this section, the existence of a stationary distribution for the observation process $\{Y_t\}_{t \in \mathbb{N}}$ of the original (unperturbed) class of GARMA models is proved. Since $\{Y_t\}_{t \in \mathbb{N}}$ is not itself a Markov chain, by using the results of Appendix Section “Feller Conditions”, the existence of a strict-sense stationary ergodic process $\{Y_t\}_{t \in \mathbb{N}}$ is proved by showing that the Markov chain $\{\mu_t\}_{t \in \mathbb{N}}$ has a unique stationary distribution. First, the existence of a stationary distribution for the Markov chain is shown by using the weak Feller property. Let $Y_0(x)$ denote the random variable Y_0 conditioned on $\mu_0 = x$. The results of this section are due to [42].

Theorem 8.2 *The process $\{\mu_t\}_{t \in \mathbb{N}}$ specified by the GARMA model (8.16) has a stationary distribution, and thus is stationary for an appropriate initial distribution for μ_0 , under the following conditions:*

1. $Y_0(x) \Rightarrow Y_0(x')$ as $x \rightarrow x'$.
2. $E(Y_t | \mu_t) = \mu_t$.
3. ($2 + \delta$ moment condition): There exist $\delta > 0$, $r \in [0, 1 + \delta)$, and non-negative constants d_1, d_2 such that

$$E(|Y_t - \mu_t|^{2+\delta} | \mu_t) \leq d_1 |\mu_t|^r + d_2.$$

4. g is bijective, increasing, and

- a. $g : \mathbb{R} \mapsto \mathbb{R}$ is concave on \mathbb{R}^+ and convex on \mathbb{R}^- , and $|\phi| < 1$
- b. $g : \mathbb{R}^+ \mapsto \mathbb{R}$ is concave on \mathbb{R}^+ , and $|\phi|, |\theta| < 1$
- c. $|\theta| < 1$; no additional conditions on $g : (0, a) \mapsto \mathbb{R}$.

The proof is postponed to Appendix Section “Main Proofs”.

Then, uniqueness of the stationary distribution for μ_t is shown. It is further assumed that the distribution $\pi_z(\cdot)$ of $g(Y_t)$ conditional on $g(\mu_t) = z$ varies smoothly and not too quickly as a function of z . This mean that $\pi_z(\cdot)$ has the Lipschitz property

$$\sup_{w, z \in \mathbb{R}: w \neq z} \frac{\|\pi_w(\cdot) - \pi_z(\cdot)\|_{TV}}{|w, z|} < B < \infty \quad (8.18)$$

where $\|\cdot\|_{TV}$ is the total variation norm [44, p. 315].

Theorem 8.3 *Suppose that the conditions of Theorem 8.2 and the Lipschitz condition (8.18) hold, and that there is some $x \in \mathbb{R}$ that is in the support of Y_0 for all values of μ_0 . Then there is a unique stationary distribution for $\{\mu_t\}_{t \in \mathbb{N}}$. This implies that $\{Y_t\}_{t \in \mathbb{N}}$ is strictly stationary when μ_0 is initialized appropriately.*

The proof of the theorem is based on the asymptotic strong Feller property (see Definition 8.7 in the Appendix) and it can be found in [42] and Proposition 8 in [23].

A similar procedure can be followed to prove strict stationarity and ergodicity for the GARMA model with more than one lag. See [42] for further discussion.

8.5.2 Strict Stationarity and Ergodicity for Log-Linear Poisson Autoregression

The work of [23] is intended to provide an alternative proof of stationarity and ergodicity for the discrete process Y_t , by weakening the Lipschitz assumption (8.18), which is not satisfied in several widely applied observation-driven models. The authors specify a broad class of observation-driven models, such as the log-linear Poisson autoregression, as follows. Let (X, d) be a locally compact, complete and separable metric space and denote by \mathcal{X} the associated Borel sigma-field. Let (Y, \mathcal{Y}) be a measurable space, H a Markov kernel from (X, \mathcal{X}) to (Y, \mathcal{Y}) and $(x, y) \mapsto f_y(x)$ a measurable function from $(X \times Y, \mathcal{X} \otimes \mathcal{Y})$ to (X, \mathcal{X}) .

An observation-driven model on \mathbb{N} is a stochastic process $\{(X_t, Y_t)\}_{t \in \mathbb{N}}$ on its space $X \times Y$ satisfying the following recursions: for all $t \in \mathbb{N}$,

$$Y_{t+1} | \mathcal{F}_t \sim H(X_t; \cdot), \quad X_{t+1} = f_{Y_{t+1}}(X_t) \tag{8.19}$$

where $\mathcal{F}_t = \sigma(X_l, Y_l; l \leq t, l \in \mathbb{N})$ and $f_{Y_{t+1}}$ is a generic function depending on the observation process $\{Y_l, \}_{l \leq t+1}$. Similarly $\{(X_t, Y_t)\}_{t \in \mathbb{N}}$ is an observation driven time series model on \mathbb{N} if the previous recursion holds for all $t \in \mathbb{N}$ with $\mathcal{F}_k = \sigma(X_l, Y_l; l \leq t, l \in \mathbb{N})$.

Denote by Q the transition probability associated with $\{X_t\}_{t \in \mathbb{Z}}$, defined implicitly by the recursions (8.19); see Appendix Section “Technical Details” for details. General conditions expressed in terms of H and f are derived by Douc et al. [23] so that the processes $\{X_t\}_{t \in \mathbb{Z}}$ and $\{(X_t, Y_t)\}_{t \in \mathbb{Z}}$ admits a unique invariant probability distribution. In Appendix Section “Technical Details”, we highlight the proof for strict-stationarity and ergodicity for the discrete process in Eq. (8.19). We give only an example of this general approach on the so-called log-linear Poisson autoregression, [28].

Only the aspects of the proof which significantly differ from those in Sect. 8.5.1 are shown. We redirect the reader to Appendix Sections “Coupling Construction”–

“Assumptions and Results of the Alternative Markov Chain Approach Without Irreducibility” for the details.

Let us consider a Markov chain $\{X_t\}_{t \in \mathbb{Z}}$ with a transition kernel Q given implicitly by the following recursive equations:

$$\begin{aligned} Y_{t+1} | X_{0:t}, Y_{0:t} &\sim \mathcal{P}(e^{X_t}) \\ X_{t+1} &= d + a X_t + b \ln(Y_{t+1} + 1) \end{aligned}$$

where $\mathcal{P}(\lambda)$ is the Poisson distribution with parameter λ . Let $\mathbf{X} = \mathbb{R}$ so $d(x, x') = |x - x'|$ and the function $f_y(x) = d + a x + b \ln(1 + y)$.

Theorem 8.4 *If $|a+b| \vee |a| \vee |b| < 1$ and there is some $x \in \mathbb{R}$ that is in the support of Y_0 for all values of X_0 . Then there is a unique stationary distribution for $\{X_t\}_{t \in \mathbb{N}}$. This implies that $\{Y_t\}_{t \in \mathbb{N}}$ is strictly stationary when X_0 is initialized appropriately.*

The proof is in Appendix Section “Main Proofs”. We remark that, for this method, the attention is put on showing stability conditions for the model with only one lag. The extension to orders greater than the first could be challenging; see [23].

8.6 Inference

The inferential procedures related to observation-driven models for discrete processes usually rely on maximum likelihood estimation. However, a misspecified version is available, namely Quasi MLE (QMLE), where the likelihood function considered for the estimation is not necessarily paired with the conditional distribution assumed as a data generating process, see [4, 61], and [38].

For linear and log-linear Poisson autoregressive time series models, [30] and [28] developed maximum likelihood estimation. Quasi-likelihood inference for negative binomial processes has been introduced in [12]. Ahmad and Francq [1] established consistency and asymptotic normality of the QMLE for the specific case of the Poisson distribution. For the general framework (8.19), [23] proved the consistency of MLE and QMLE. Asymptotic normality, in the same setting, is later discussed by Douc et al. [24]. Comparable results have been derived by Davis and Liu [18], based on the approach developed by Neumann [46]. The aim of this section is to give a brief introduction to QMLE for the framework summarized by Eq. (8.19).

Let (Θ, d) be a compact metric subspace of \mathbb{R}^p . Define the parameter vector $\theta \in \Theta$ and the QMLE

$$\hat{\theta}_{n,x} = \arg \max_{\theta \in \Theta} L_{n,x}^\theta \langle Y_{1:n} \rangle, \quad (8.20)$$

with corresponding conditional (quasi) log-likelihood function

$$L_{n,x}^\theta \langle Y_{1:n} \rangle = n^{-1} \log \left(\prod_{t=1}^n h(f^\theta \langle y_{1:t-1} \rangle(x); y_t) \right),$$

where $h(f^\theta \langle y_{1:t-1} \rangle(x); y_t)$ is the density function coming from the kernel H in (8.19) and the notation $f^\theta \langle y_{s:t} \rangle(x) = f_{y_t}^\theta \circ f_{y_{t-1}}^\theta \circ \dots \circ f_{y_s}^\theta(x)$, $s \leq t$ refers to the so-called Iterated Random Function (IRF), see [22], with the convention $f^\theta \langle y_{1:0} \rangle(x) = x$. Moreover, let $X_0 = x$ be the starting value of the chain X_t in (8.19); then the likelihood is conditional to the starting point x . The dependence on the parameter vector θ is emphasized by the notation $f_{y_s}^\theta(\cdot) = f_{y_s}(\cdot)$.

The following results are due to [23] and [24]. We make the following assumptions.

- (B1) $\{Y_t\}_{t \in \mathbb{Z}}$ is a strict-sense stationary and ergodic stochastic process.
- (B2) $\forall(x, y) \in \mathbf{X} \times \mathbf{Y}$, the functions $\theta \mapsto f^\theta y(x)$ and $v \mapsto h(v, y)$ are continuous.
- (B3) There exists a family of finite random variables $\{f^\theta \langle Y_{-\infty:t} \rangle : (\theta, t) \in \Theta \times \mathbb{Z}\}$ such that for all $x \in X$,
 - (i) $\lim_{m \rightarrow \infty} \sup_{\theta \in \Theta} d[f^\theta \langle Y_{-m:0} \rangle(x), f^\theta \langle Y_{-\infty:0} \rangle] = 0$, a.s.
 - (ii) $\lim_{t \rightarrow \infty} \sup_{\theta \in \Theta} |\log h(f^\theta \langle Y_{1:t-1} \rangle(x); Y_t) - \log h(f^\theta \langle Y_{-\infty:t-1} \rangle; Y_t)| = 0$, a.s.
 - (iii) $E \left[\sup_{\theta \in \Theta} (\log h(f^\theta \langle Y_{-\infty:t-1} \rangle; Y_t))_+ \right] < \infty$, where the notation $(\cdot)_+$ is the positive part.
- (B4) The true parameter vector θ^* is assumed to be in Θ° , the interior of Θ .
- (B5) The function $\int H(x^*, dy) \log h(x, y)$ has a unique maximum $\{x^*\}$.

Conditions (B1)–(B2) are required for the estimator $\theta_{n,x}$ to be well-defined. Assumption (B3)-(i) assures that, regardless of the initial value of $X_{-m} = x$, the chain X_0 (and thus X_t) can be approximated by a quantity involving the infinite past of the observations. Intuitively, (B3)-(ii) allows the conditional log-likelihood function to be approximated by a stationary sequence involving the infinite past of Y_t . (B3)-(iii) is required in order to obtain a solvable maximization problem and holds for the discrete Y_t [23, Rem. 18]. Assumption (B5) corresponds to an identification condition.

Theorem 8.5 *Assume that (B1)–(B5) hold and $f^{\theta^*} \langle Y_{-\infty:0} \rangle = f^\theta \langle Y_{-\infty:0} \rangle$ implies that $\theta = \theta^*$. Then, for all $x \in \mathbf{X}$,*

$$\lim_{n \rightarrow \infty} \hat{\theta}_{n,x} = \theta^*, \quad \text{a.s.}$$

These results establish strong consistency of the QMLE. For the proof and other details see [24]. An example of the derivation of Theorem 8.5 for the one lag log-linear Poisson AR can be found in [23]. See also [1], for a similar result.

Finally, the condition under which the QMLE (8.20) is asymptotically normally distributed are investigated. Define the score function

$$\chi^\theta(x_t(\theta), y_t) = \nabla_{\theta} x_t(\theta) \frac{\partial \log h(x_t, y_t)}{\partial x_t},$$

and the Hessian matrix

$$K^\theta(x_t(\theta), y_t) = \nabla_{\theta}^2 x_t(\theta) \frac{\partial \log h(x_t, y_t)}{\partial x_t} + \nabla_{\theta} x_t(\theta) \nabla_{\theta} x_t(\theta)' \frac{\partial^2 \log h(x_t, y_t)}{\partial x_t^2}.$$

Then, define the following functions $f^\bullet \langle Y_{-\infty:t-1} \rangle : \theta \mapsto f^\theta \langle Y_{-\infty:t-1} \rangle$ and $f^\bullet \langle Y_{1:t-1} \rangle(x) : \theta \mapsto f^\theta \langle Y_{1:t-1}(x) \rangle$. A further assumption is required.

(B6): For all $y \in \mathbf{Y}$, the function $v \mapsto h(v, y)$ is twice continuously differentiable. Moreover, there exist $\epsilon > 0$ and a family of a.s. finite random variables

$$\{f^\theta \langle Y_{-\infty:t} \rangle : (\theta, t) \in \theta \times \mathbb{Z}\}$$

such that $f^{\theta^*} \langle Y_{-\infty:0} \rangle$ is in the interior of \mathbf{X} , the function $\theta \mapsto f^\theta \langle Y_{-\infty:0} \rangle$ is twice continuously differentiable on some ball $B(\theta^*, \epsilon)$ and for all $x \in \mathbf{X}$,

(i) a.s.,

$$\lim_{t \rightarrow \infty} \left\| \chi^{\theta^*} (f^\bullet \langle Y_{1:t-1} \rangle(x), Y_t) - \chi^{\theta^*} (f^\bullet \langle Y_{-\infty:t-1} \rangle, Y_t) \right\| = 0$$

where $\|\cdot\|$ is any norm on \mathbb{R}^p .

(ii) a.s.,

$$\lim_{t \rightarrow \infty} \sup_{\theta \in B(\theta^*, \epsilon)} \left\| K^\theta (f^\bullet \langle Y_{1:t-1} \rangle(x), Y_t) - K^\theta (f^\bullet \langle Y_{-\infty:t-1} \rangle, Y_t) \right\| = 0$$

where $\|\cdot\|$ denote any norm on $p \times p$ -matrices with real entries.

(iii)

$$\begin{aligned} & \mathbb{E} \left[\left\| \chi^{\theta^*} (f^\bullet \langle Y_{-\infty:0} \rangle, Y_1) \right\|^2 \right] < \infty, \\ & \mathbb{E} \left[\sup_{\theta \in B(\theta^*, \epsilon)} \left\| K^\theta (f^\bullet \langle Y_{-\infty:0} \rangle, Y_1) \right\| \right] < \infty \end{aligned}$$

Moreover, the matrix

$$\mathcal{J}(\theta_\star) = \mathbb{E} \left[\left(\nabla_\theta g^{\theta_\star} \langle Y_{-\infty:0} \rangle \right) \left(\nabla_\theta f^{\theta_\star} \langle Y_{-\infty:0} \rangle \right)' \frac{\partial^2}{\partial x^2} \right. \\ \left. \times \log h \left(f^{\theta_\star} \langle Y_{-\infty:0} \rangle, Y_1 \right) \right]$$

is non-singular.

Intuitively, (B6) assumes that the score function and the information matrix of the data can be approximated by the their counterpart with the infinite past of the process. In addition, all of these quantities are assumed to exist.

Theorem 8.6 *Assume (B1)–(B6) hold and $\hat{\theta}_{n,x} \xrightarrow{p} \theta^\star$. Then,*

$$\sqrt{n}(\hat{\theta}_{n,x} - \theta^\star) \xrightarrow{d} \mathcal{N}(0, \mathcal{J}(\theta^\star)^{-1} \mathcal{I}(\theta^\star) \mathcal{J}(\theta^\star)^{-1}),$$

where

$$\mathcal{I}(\theta^\star) = \mathbb{E} \left[\left(\nabla_\theta f^{\theta^\star} \langle Y_{-\infty:0} \rangle \right) \left(\nabla_\theta f^{\theta^\star} \langle Y_{-\infty:0} \rangle \right)' \right. \\ \left. \times \left(\frac{\partial}{\partial x} \log h \left(f^{\theta^\star} \langle Y_{-\infty:0} \rangle, Y_1 \right) \right)^2 \right].$$

The proof relies on the argument of [24].

Note that, for a correctly specified MLE, Eq. (8.20) is the exact MLE and $\mathcal{J}(\theta^\star) = \mathcal{I}(\theta^\star)$ in Theorem 8.6, providing the standard ML inference. For further details see [24]. When the quasi-likelihood comes from the Poisson distribution, [1] proved a similar result for Theorem 8.6. An analogous conclusion can be found in [12] for the Negative Binomial distribution.

The results of Theorems 8.5–8.6 apply to all the observation-driven models presented so far, like those introduced in Sect. 8.3, since they can be written as special cases of the framework (8.19).

8.7 Applications

8.7.1 Number of Deaths from COVID-19

The recent outbreak of the new coronavirus called COVID-19 in late 2019 lends itself to a current illustration of the model specified in Eqs. (8.1)–(8.3). The time series we consider is related to the daily number of deaths from COVID-19 in Italy from 21st February 2020 to 20th December 2020. The data can be downloaded

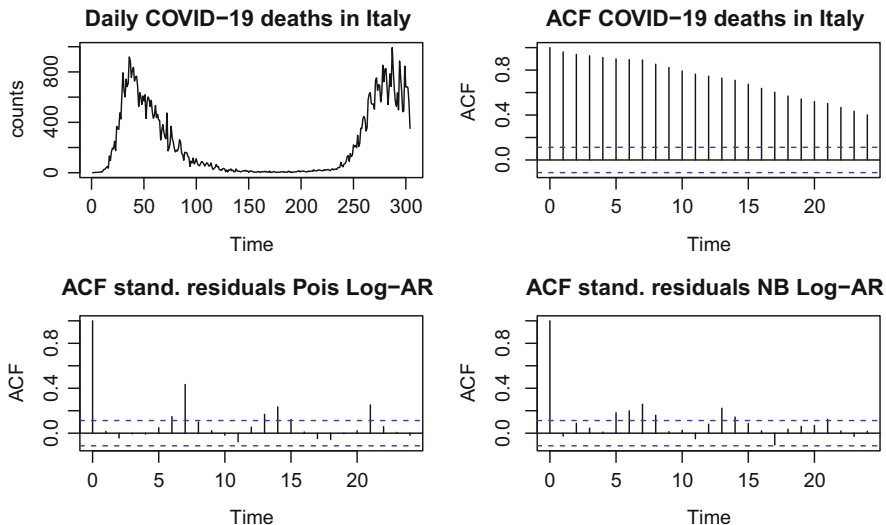


Fig. 8.1 Top: daily count from COVID-19 deaths in Italy (left) and corresponding ACF (right). Bottom-left: ACF standardized residuals for log-AR Poisson model. Bottom-right: ACF standardized residuals for log-AR NB model

from the GitHub repository of the 2019 Novel Coronavirus Visual Dashboard operated by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU), <https://github.com/CSSEGISandData/COVID-19>. The time series has a sample size equal to $n = 304$ and is plotted in Fig. 8.1, along with its autocorrelation function (ACF). The latter shows a temporal correlation spread over several lags in the past. We argue that observation driven models for discrete time series data may be effective in this case. The long-time dependence suggests the use of a feedback mechanism, captured by the latent process.

We fit models coming from two different distributions; the Poisson distribution:

$$P(Y_t = y | \mathcal{F}_{t-1}) = \frac{\exp(-\mu_t) \mu_t^y}{y!}, \quad y = 0, 1, 2, \dots$$

and the NB:

$$P(Y_t = y | \mathcal{F}_{t-1}) = \frac{\Gamma(v + y)}{\Gamma(y + 1)\Gamma(v)} \left(\frac{v}{v + \mu_t}\right)^v \left(\frac{\mu_t}{v + \mu_t}\right)^y, \quad y = 0, 1, 2, \dots \tag{8.21}$$

where $v > 0$ is the dispersion parameter and μ_t is the conditional expectation; the latter is the same for both distributions. Indeed, Eq. (8.21) is defined in terms of mean rather than of the probability parameter $p_t = \frac{v}{v + \mu_t}$ and accounts for overdispersion in the data as, in (8.21), $V(Y_t | \mathcal{F}_{t-1}) = \mu_t (1 + \mu_t / v) \geq \mu_t$. In the Poisson distribution, the mean and variance are equal.

In order to set a model selection procedure we have estimated the following one-lag models: the log-linear Poisson autoregression (8.10)

$$\log(\mu_t) = \alpha + \phi \log(y_{t-1} + 1) + \gamma \log(\mu_{t-1}),$$

the GARMA model (8.6)

$$\log(\mu_t) = \alpha + \phi \log(y_{t-1}^*) + \theta [\log(y_{t-1}^*) - \log(\mu_{t-1})],$$

where $y_{t-1}^* = \max\{y_t, c\}$ with $c = 0.1$ and $\alpha = (1 - \phi)\beta$, and the GLARMA model in Eq. (8.7)

$$\log(\mu_t) = \alpha + \gamma \log(\mu_{t-1}) + \theta \left(\frac{y_{t-1} - \mu_{t-1}}{s_{t-1}} \right),$$

where $s_t = \sqrt{\mu_t}$ for the Poisson distribution and $s_t = \sqrt{\mu_t(1 + \mu_t/\nu)}$ for the NB.

QMLE has been carried out. The log-likelihood function of the Poisson and NB distributions is maximized by using a standard optimizer of R based on the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm. The score functions, written in terms of predictor $x_t = \log \mu_t$, are:

$$\begin{aligned} \chi_n(\theta) &= \frac{1}{n} \sum_{t=1}^n \left(y_t - \exp x_t(\theta) \right) \frac{\partial x_t(\theta)}{\partial \theta}, \\ \chi_n(\theta) &= \frac{1}{n} \sum_{t=1}^n \left(y_t - \frac{(y_t + \nu) \exp x_t(\theta)}{\exp x_t(\theta) + \nu} \right) \frac{\partial x_t(\theta)}{\partial \theta}. \end{aligned}$$

The solution of the system of non-linear equations $\chi_n(\theta) = 0$, if it exists, provides the QMLE of θ (denoted by $\hat{\theta}$). See Sect. 8.6 for details on inference. In NB models, the estimation of ν is required. We used the moment estimator, as in [13]:

$$\hat{\nu} = \left\{ 1/n \sum_{t=1}^n \left[(y_t - \hat{\mu}_t)^2 - \hat{\mu}_t \right] / \hat{\mu}_t^2 \right\}^{-1},$$

where $\hat{\mu}_t = \mu_t(\hat{\theta})$ from the Poisson model. Clearly, we replace each quantity with the sample counterparts computed at $\hat{\theta}$. The results of the analysis are summarized in Table 8.1. In the likelihood-based framework, model selection is based on information criteria, such as the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). All the coefficients of the estimation are significant at the usual 5% level. Both AIC and BIC select the NB log-AR model as the best, in the goodness-of-fit sense.

Table 8.1 MLE results from COVID-19 death counts (standard errors in brackets). Lowest values of AIC and BIC are given in bold

Models	$\hat{\alpha}$	$\hat{\phi}$	$\hat{\gamma}$	$\hat{\theta}$	$\hat{\nu}$	AIC	BIC
Pois log-AR	0.154	0.619	0.357	–	–	24.204	35.355
	(0.035)	(0.060)	(0.062)	–			
Pois GARMA	0.211	0.976	–	-0.360	–	24.163	35.314
	(0.036)	(0.006)	–	(0.061)			
Pois GLARMA	0.187	–	0.961	0.038	–	28.047	39.198
	(0.031)	–	(0.008)	(0.003)			
NB log-AR	0.061	0.569	0.424	–	10.733	15.227	26.378
	(0.023)	(0.036)	(0.035)	–			
NB GARMA	0.157	0.976	–	-0.441	9.123	15.262	26.413
	(0.022)	(0.004)	–	(0.034)			
NB GLARMA	0.712	–	0.822	0.177	4.756	16.636	27.787
	(0.072)	–	(0.016)	(0.011)			

We then assess the adequacy of fit. We check the behaviour of the standardized Pearson residuals $e_t = [Y_t - E(Y_t|\mathcal{F}_{t-1})] / \sqrt{V(Y_t|\mathcal{F}_{t-1})}$, which is done by taking the empirical version \hat{e}_t from the estimated quantities. If the model is correctly specified, the residuals form a white noise sequence with constant variance. The autocorrelation function (ACF) in our case appears uncorrelated for the NB case (see Fig. 8.1, for log-AR models), except for mild residual autocorrelation at weekly lags, which may depend on how the recorded data are reported, rather than on any actual infection cycles or occurrence of events on certain specific days of the week.

Another check comes from the probability calibrations, as defined in [33]. In particular, [17] introduced a non-randomized version of Probability Integral Transform (PIT) for discrete data. It can be built by defining the following conditional distribution function

$$F(u|y_t) = \begin{cases} 0, & u \leq P_t(y_t - 1) \\ \frac{u - P_t(y_t - 1)}{P_t(y_t) - P_t(y_t - 1)}, & P_t(y_t) \leq u \leq P_t(y_t) \\ 1, & u \geq P_t(y_t) \end{cases} \quad (8.22)$$

where $P_t(\cdot)$ is the cumulative distribution function at time t (in our case Poisson or NB). If the model is correct, $u \sim Uniform(0, 1)$ and the PIT (8.22) will appear to be the cumulative distribution function of a $Uniform(0, 1)$. The PIT (8.22) is computed for each realization of the time series $y_t, t = 1 \dots, n$ and for values $u = j/J, j = 1, \dots, J$, where J is the number of bins (usually equal to 10 or 20); then its mean $\bar{F}(j/J) = 1/n \sum_{t=1}^n F(j/J|y_t)$ is taken. The outcomes are probability mass functions, which are obtained in terms of differences $\bar{F}(j/J) - \bar{F}(j-1/J)$ plotted in Fig. 8.2. The NB PIT's appear to be closer to $Uniform(0, 1)$ for log-linear autoregression and GARMA models than for the remaining models.

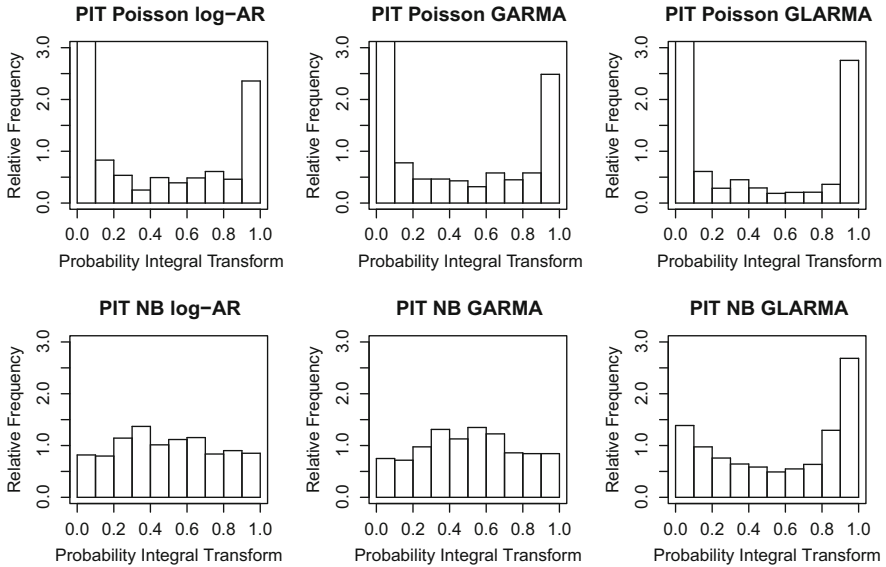


Fig. 8.2 Top: PIT’s for the Poisson models. Bottom: PIT’s for the NB models

Table 8.2 Predictive performance from COVID-19 death counts (smallest values in bold)

Models	Distribution	Logs	qs	sphs	rps	dss
log-AR	Poisson	9.1054	-0.0205	-0.1260	32.6055	21.1890
	NB	4.6168	-0.0324	-0.1458	29.3324	14.0354
GARMA	Poisson	9.0849	-0.0212	-0.1274	32.5241	21.1019
	NB	4.6345	-0.0320	-0.1448	29.7812	14.1704
GLARMA	Poisson	11.0270	0.0009	-0.0822	36.5751	26.0447
	NB	5.3215	-0.0176	-0.1033	74.0710	16.1614

In order to assess the predictive power of each model, we refer to the concept of sharpness of the predictive distribution defined in [33]. It can be measured by some average quantities related to the predictive distribution, which take the form $1/n \sum_{t=1}^n d[P_t(y_t)]$, where $d(\cdot)$ is a scoring rule, see [32]. We used some of the usual scoring rules employed in the literature: the logarithmic score (logs) $-\log p_t(y_t)$, where $p_t(\cdot)$ is the probability mass at the time t ; the quadratic score (qs) $-2p_t(y_t) + \|p\|^2$, where $\|p\|^2 = \sum_{k=0}^{\infty} p_t^2(k)$; the spherical score (sphs) $-p_t(y_t)/\|p\|$; the ranked probability score (rps) $\sum_{k=0}^{\infty} [P_t(k) - \mathbf{1}(y_t \leq k)]$, and the Dawid-Sebastiani score (dss) $(\frac{y_t - \mu_t}{\sigma_t})^2 + 2 \log \sigma_t$, where μ_t and σ_t are the mean and variance of $P_t(y_t)$. These scores are applied to different models and distributions. The results are summarized in Table 8.2. The NB log-AR model is chosen as the best model, as it has the best predictive performance for all the scoring rules, which confirms the result of the goodness-of-fit analysis.

8.7.2 Returns Sign for J&J Stock

In financial time series analysis, it is well known that the expected value of the stock returns is unpredictable, as, unconditionally, returns behave like white noise sequences. However, the sign of the returns can be predicted, as well as its volatility, see for instance [10]. Recently, a literature on sign prediction for stock returns through binary times series has flourished, see Moysiadis and Fokianos [45, Sec. 6.1] for a survey.

We apply the binary time series approach for sign prediction to BARMA, GARMA, and GLARMA models. The time series of logarithmic returns for the weekly closing prices of the Johnson & Johnson (J&J) stock from 2/1/1970 to 17/5/2021 is considered. These data can be found at <http://finance.yahoo.com>. The length of the series is $n = 2682$. More precisely, the log-return series is derived as $r_t = \log(P_t^c) - \log(P_t^o)$, where P_t^c is the closing price of the stock at time t and P_t^o is its associated opening price. The binary time series of signs is generated as follows: $Y_t = 1$ when the logarithmic return at time t is positive, such as when the price change is positive, and $Y_t = 0$ otherwise (negative price change). The time series of stock log-returns is plotted in Fig. 8.3. The ACF of the returns is reported, as well. As expected, no temporal correlation is detected in the log-returns time series, which is in line with the aforementioned unpredictability of return means.

The result of the ML estimation is presented in Table 8.3. All the coefficients are significant at the usual nominal levels. The information criteria select the GLARMA model. The employed models seem to be adequate for the data analysed, as the ACF in Fig. 8.4 show uncorrelated residual errors. The Bernoulli distribution is well adapted to the empirical distribution of the model (Fig. 8.5). The predictive performance is measured by using the same scoring rule of the previous application. In Table 8.4, the GLARMA model is selected again by the majority of the scores.

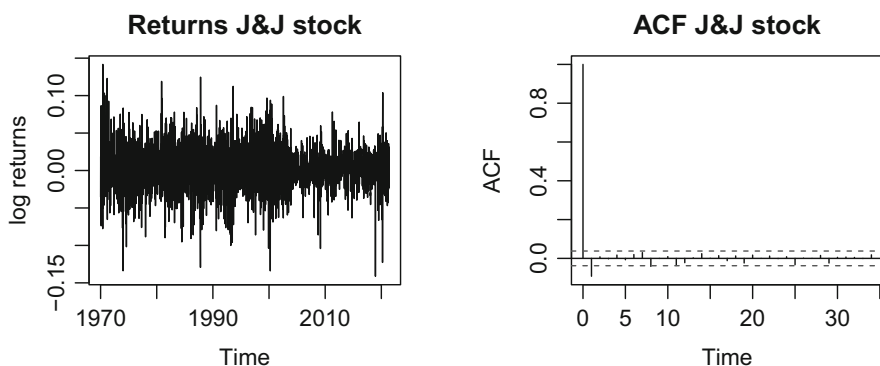


Fig. 8.3 Weekly log-returns for the J&J stock (left) and corresponding ACF (right)

Table 8.3 MLE results for J&J return sign (standard errors in brackets $\times 10^3$). Lowest values of AIC and BIC are given in bold

Models	$\hat{\alpha}$	$\hat{\phi}$	$\hat{\gamma}$	$\hat{\theta}$	$AIC \times 10^{-1}$	$BIC 10^{-1}$
BARMA	0.308	-0.444	-	0.449	371.868	373.637
	(0.4612)	(0.8877)		(0.8891)		
GARMA	0.170	-0.992	-	0.988	371.437	373.205
	(0.0568)	(0.0009)	-	(0.0013)		
GLARMA	0.007	-	0.910	0.035	371.364	373.132
	(0.0009)	-	(0.0112)	(0.0048)		

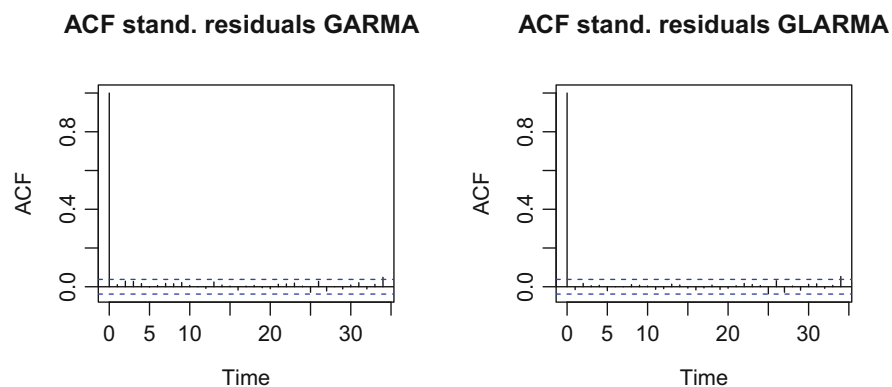


Fig. 8.4 Left: ACF standardized residuals for the Bernoulli GARMA model. Right: ACF of standardized residuals for the Bernoulli GLARMA model

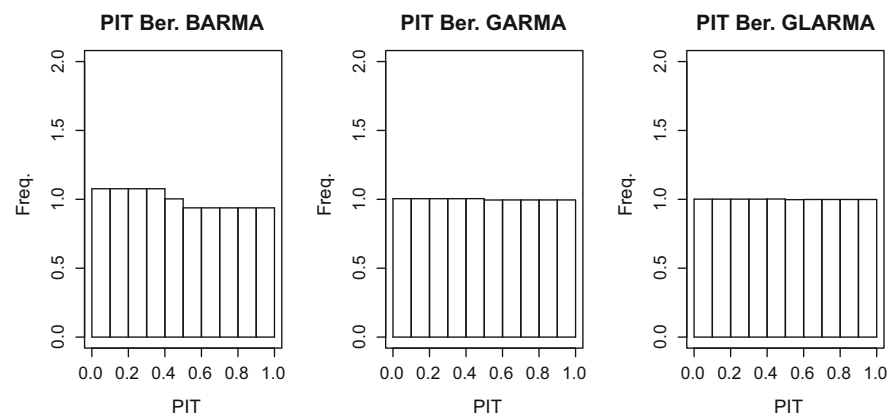


Fig. 8.5 PIT's for the Bernoulli models

Table 8.4 Predictive performance for J&J return sign (smallest values in bold)

Models	Logs	qs	sphs	rps	dss
BARMA	0.6947	-0.4984	-0.7060	0.2508	-1.7808
GARMA	0.6916	-0.5015	-0.7082	0.2492	-1.7788
GLARMA	0.6915	-0.5017	-0.7083	0.2492	-1.7793

8.8 Concluding Remarks

The most notable observation-driven models for discrete data have been reviewed. The basic stochastic properties required to guarantee their correct use have been presented, as well as the technical tools for their practical application. Increased availability and interest in discrete data encourage the use of these time series models, which will be promising key tools in future works on binary and count data.

For theoretical and substantive reasons, the analysis of discrete-valued times series would benefit from the specification of a unified framework able to encompass most of the models available in the literature. As a matter of fact, it is not trivial to explore whether the models that we have discussed are nested, and, consequently, to derive stochastic properties that simultaneously hold across models. In addition, model comparison becomes complicated when direct relationships among different models are unknown.

Concerning probabilistic properties, up to the present time, strict stationarity and ergodicity have not been established explicitly for a number of the models reviewed in this chapter (GLARMA and M-GARMA for discrete variables, for example). In principle, the theoretical tools presented in Appendix Section “Technical Details” would be sufficient to show stability conditions for such models, as well as any general framework encompassed in (8.1) and (8.3), but the derivations of such stationarity conditions might not be immediate and far from obvious, as shown in Sect. 8.5 for the GARMA and log-AR models. Then, this would be a useful step further for the literature.

Another aspect which may be interesting to consider is related to the inferential assumptions reported in Sect. 8.6, which could be generalized to distributions other than Poisson and Negative Binomial and for several different models encompassed in (8.1) and (8.3). Lastly, model selection procedures could also be further investigated. We view these aspects as promising topics for future research.

Appendix

Technical Details

Markov Chain Specification

In order to derive strict stationarity and ergodicity conditions, the problem is reformulated in terms of Markov chain theory. Let us consider an observation driven model in the most general form:

$$Y_t \mid \mathcal{F}_{t-1} \sim q(\cdot; \mu_t) \tag{8.23}$$

$$\mu_t = c_\delta(Y_{0:t-1}) \tag{8.24}$$

where we adopt the shorthand notation Y_t for the process and, as before, y_t its realization. The function q is simply the density function which comes from (8.1), whereas c_δ is some function which describes the form of the dependence from the observation. In general, $Y_{s:t} = (Y_s, Y_{s+1}, \dots, Y_t)$ where $s \leq t$. The symbol δ is the vector of parameters of the model. Of course, the initial values $\mu_{0:p-1}$ are supposed to be known. The model in (8.24) can be rewritten as:

$$\mu_t = g_\delta(Y_{t-p:t-1}, \mu_{t-p:t-1}).$$

This way of writing the observation driven model [15] gives a Markov p -structure for μ_t and then implies that the vector $\mu_{t-p:t-1}$ forms the state of a Markov chain indexed by t . In this case it is possible to prove stationarity and ergodicity of $\{Y_t\}_{t \in \mathbb{N}}$ by first showing these properties for the multivariate Markov chain $\{\mu_{t-p:t-1}\}_{t \geq p}$, then shifting the results back to the time series model $\{Y_t\}_{t \in \mathbb{N}}$.

Some useful definitions for theorems based upon the theory of Markov chains asserted throughout the paper are introduced. Define a general Markov chain $X = \{X_t\}_{t \in \mathbb{N}}$ on a state space S with σ -algebra \mathcal{F} and define $P^t(x, A) = P(X_t \in A \mid X_0 = x)$ for $A \in \mathcal{F}$ as the t -step transition probability starting from state $X_0 = x$.

Definition 8.1 A Markov chain X is φ -irreducible if there exists a non-trivial measure φ on \mathcal{F} such that, whenever $\varphi(A) > 0$, $P^t(x, A) > 0$ for some $t = t(x, A)$, for all $x \in S$.

Also, the definition of aperiodicity as stated in [44] is needed. Define a period $d(\alpha) = \gcd \{t \geq 1 : P^t(\alpha, \alpha) > 0\}$.

Definition 8.2 An irreducible Markov chain X is aperiodic if $d(x) \equiv 1, x \in X$.

Definition 8.3 A set $A \in \mathcal{F}$ is called a small set if there exists an $m > 1$, a non-trivial measure ν on \mathcal{F} , and a $\lambda > 0$ such that for all $x \in A$ and all $C \in \mathcal{F}$, $P^m(x, C) \geq \lambda \nu(C)$.

Let $E_x(\cdot)$ denote the expectation under the probability $P_x(\cdot)$ induced on the path space of the chain defined by $\Omega = \prod_{t=0}^{\infty} X_t$ with respect to $\mathcal{F}^{\infty} = \bigvee_{t=0}^{\infty} \mathcal{B}(X_t)$ when the initial state $X_0 = x$; where $\mathcal{B}(X_t)$ is the Borel σ -field on X_t .

Theorem 8.7 (Drift Conditions) *Suppose that $X = \{X_t\}_{t \in \mathbb{N}}$ is φ -irreducible on S . Let $A \subset S$ be small, and suppose that there exist $b \in (0, \infty)$, $\epsilon > 0$, and a function $V : S \rightarrow [0, \infty)$ such that for all $x \in S$,*

$$E_x [V(X_1)] \leq V(x) - \epsilon + b\mathbf{1}_{\{x \in A\}}, \tag{8.25}$$

then X is positive Harris recurrent.

The function V is called the Lyapunov function or energy function.

Positive Harris recurrent chains possess a unique stationary probability distribution π . Moreover, if X_0 is distributed according to π , then the chain X is a stationary process. If the chain is also aperiodic, then X is ergodic, in which case if the chain is initialized according to some other distribution, then the distribution of X_t will converge to π as $t \rightarrow \infty$.

A stronger form of ergodicity, called geometric ergodicity, arises if (8.25) is replaced by the condition

$$E_x [V(X_1)] \leq \beta V(x) + b\mathbf{1}_{\{x \in A\}} \tag{8.26}$$

for some $\beta \in (0, 1)$ and some $V : S \rightarrow [1, \infty)$. Indeed, (8.26) implies (8.25). Eventually, stationarity and ergodicity for the GARMA model would be accomplished if at least one of the sufficient condition (8.25), (8.26) above is fulfilled.

Unfortunately, a problem can occur when the distribution in (8.23) is not continuous (that is, Bernoulli, Poisson, . . .). In fact, in these cases the Markov chain $\{\mu_{t-p:t-1}\}_{n \geq p}$ may not be φ -irreducible. This occurs whenever Y_t can only take a countable set of values and the state space $\mu_{t-p:t-1}$ is \mathbb{R}^p . Then, given a particular initial vector $\mu_{0:p-1}$, the set of possible values for μ_t is countable. Definition 8.1 is not satisfied. For this reason, additional theoretical tools are required:

- Perturbation approach
- Feller conditions.

Perturbation Approach

First, define the perturbed form of an observation driven time series model:

$$Y_t^{(m)} \mid Y_{0:t-1}^{(m)} \sim q(\cdot; \mu_t^{(m)}) \tag{8.27}$$

$$\mu_t^{(m)} = g_{\delta,t}(Y_{0:t-1}^{(m)}, mZ_{0:t-1}), \tag{8.28}$$

where $Z_t \sim \phi$ are independent, identically distributed random perturbations having density function ϕ , $m > 0$ is a scale factor associated with the perturbation, and $g_{\delta,t}(\cdot, mZ_{0:t-1})$ is a continuous function of $Z_{0:t-1}$ such that $g_{\delta,t}(y, 0) = g_{\delta,t}(y)$ for any y . The value $\mu_0^{(m)}$ is a fixed constant that is taken to be independent of m , so that $\mu_0^{(m)} = \mu_0$. The perturbed model is constructed to be φ -irreducible, so that one can apply usual drift conditions to prove its stationarity.

Then, it can be proved that the likelihood of the parameter vector δ calculated using (8.28) converges uniformly to the likelihood calculated using the unperturbed model as $m \rightarrow 0$. More precisely, the joint density of the observations $Y = Y_{0:t}^{(m)}$ and first t perturbations $Z = Z_{0:t-1}$, conditional on the parameter vector δ , the perturbation scale m , and the initial value μ_0 , is:

$$f(Y, Z \mid \delta, m, \mu_0) = f(Z \mid \delta, m, \mu_0) \times f(Y \mid Z, \delta, m, \mu_0) \\ = \left[\prod_{k=0}^{t-1} \phi(Z_k) \right] \prod_{k=0}^t f\left(Y_k^{(m)}; \mu_k(mZ)\right)$$

where $\mu_k(mZ)$ is the value of $\mu_k^{(m)}$ induced by the perturbation vector mZ through (8.28), with $\mu_0(mZ) = \mu_0$. The likelihood function for the parameter vector δ implied by the perturbed model is the marginal density of Y integrating over Z , i.e.,

$$\mathcal{L}_m(\delta) = f(Y \mid \delta, m, \mu_0) = \int f(Y, Z \mid \delta, m, \mu_0) dZ.$$

Let the likelihood function without the perturbations be denoted by \mathcal{L} , so that

$$\mathcal{L}(\delta) = \prod_{k=0}^t f\left(Y_k^{(m)}; \mu_k(0)\right).$$

Theorem 8.8 *Under regularity conditions 1 and 2 below, the likelihood function \mathcal{L}_m based on the perturbed model (8.27)–(8.28) converges uniformly on any compact set K to the likelihood function \mathcal{L} based on the original model, i.e.,*

$$\sup_{\delta \in K} |\mathcal{L}_m(\delta) - \mathcal{L}(\delta)| \xrightarrow{m \rightarrow 0} 0$$

for any fixed sequence of observations $y_{0:t}$ and conditional on the initial value μ_0 .

So if \mathcal{L} is continuous in δ and has a finite number of local maxima and a unique global maximum on K , the maximum-likelihood estimate of δ based on \mathcal{L}_m converges to that based on \mathcal{L} . The proof is in [42]. Regularity Conditions:

1. For any fixed y the function $q(y; \mu)$ is bounded and Lipschitz continuous in μ , uniformly in $\delta \in K$.

2. For each t , $\mu_t(mZ)$ is Lipschitz in some bounded neighbourhood of zero, uniformly in $\delta \in K$.

Regularity condition 1 holds, e.g., for $q(y; \mu)$ equal to a Poisson or binomial density with mean μ , or a negative binomial density with mean μ and precision parameter φ . $\mu_t(mZ)$ can easily be constructed to satisfy condition 2. One can choose to use the perturbed model (with fixed and sufficiently small perturbation scale m) instead of the original model, without significantly affecting finite-sample parameter estimates, in order to get the strong theoretical properties associated with stationarity and ergodicity.

Although, it has been shown that the perturbed and original models are closely related, and although one can use drift conditions to show the stationarity and ergodicity properties of the perturbed model, this approach does not yield stationarity and ergodicity properties for the original model. In fact, this approach addresses consistency of parameter estimation for the perturbed model when $t \rightarrow \infty$ for fixed m and then shows that as $m \rightarrow 0$ the finite sample estimates (for a fixed number of observations t) of the perturbed model approach those of the original one. In order to show real proprieties of the original model one should consider both limits $t \rightarrow \infty$ together with $m \rightarrow 0$ in which a substantial technical difficulty associated with interchanging the limits arises. For this reason, the Feller properties introduced in the next section are needed.

Feller Conditions

To deal with the lack of the φ -irreducibility condition, the Feller properties can be used instead.

Definition 8.4 A chain evolving on a complete separable metric space S is said to be “weak Feller” if $P(x, \cdot)$ satisfies $P(x, \cdot) \Rightarrow P(y, \cdot)$ as $x \rightarrow y$, for any $y \in S$ and where \Rightarrow indicates convergence in distribution.

In the absence of φ -irreducibility, the “weak Feller” condition can be combined with a drift condition (8.25) or (8.26) to show the existence of a stationary distribution [55]:

Theorem 8.9 *Suppose that S is a locally compact complete separable metric space with \mathcal{F} the Borel σ -field on S , and the Markov chain $\{X_t\}_{t \in \mathbb{N}}$ with transition kernel P is weak Feller. Let $A \in \mathcal{F}$ be compact, and suppose that there exist $b \in (0, \infty)$, $\varepsilon > 0$, and a function $V : S \rightarrow [0, \infty)$ such that for all $x \in S$, the drift condition (8.25) holds. Then there exists a stationary distribution for P .*

Uniqueness of the stationary distribution can be established using the “asymptotic strong Feller” property, defined in [35]. Before doing it, further definitions are required:

Definition 8.5 Let S be a Polish (complete, separable, metrizable) space. A “totally separating system of metrics” $\{d_t\}_{t \in \mathbb{N}}$ for S is a set of metrics such that for any $x, y \in S$ with $x \neq y$, the value $d_t(x, y)$ is non-decreasing in t and $\lim_{t \rightarrow \infty} d_t(x, y) = 1$.

Definition 8.6 A metric d on S implies the following distance between probability measures μ_1 and μ_2 :

$$\|\mu_1 - \mu_2\|_d = \sup_{\text{Lip}_d \phi = 1} \left(\int \phi(x) \mu_1(dx) - \int \phi(x) \mu_2(dx) \right) \tag{8.29}$$

where

$$\text{Lip}_d \phi = \sup_{x, y \in S: x \neq y} \frac{|\phi(x) - \phi(y)|}{d(x, y)}$$

is the minimal Lipschitz constant for ϕ with respect to d .

Definition 8.7 A chain is “asymptotically strong Feller” if, for every fixed $x \in S$, there is a totally separating system of metric $\{d_t\}$ for S and a sequence $t_n > 0$ such that

$$\lim_{\delta \rightarrow \infty} \limsup_{t \rightarrow \infty} \sup_{y \in B(x, \delta)} \|P^{t_n}(x, \cdot) - P^{t_n}(y, \cdot)\|_{d_t} = 0$$

where $B(x, \delta)$ is the open ball of radius δ centred at x , as measured using some metric defining the topology of S .

Definition 8.8 A “reachable” point $x \in S$ means that for all open sets A containing x , $\sum_{t=1}^{\infty} P^t(y, A) > 0$ for all $y \in S$.

Theorem 8.10 *Suppose that S is a Polish space and the Markov chain $\{X_t\}_{t \in \mathbb{N}}$ with transition kernel P is asymptotically strong Feller. If there is a reachable point $x \in S$ then P can have at most one stationary distribution.*

This is an extension of [35]. The results of this section lay the foundation for showing the convergence and asymptotic properties of maximum likelihood estimators for the discrete-valued observation driven models.

Coupling Construction

Introduce a kernel \bar{H} from $(\mathcal{X}^2, \mathcal{X}^{\otimes 2})$ to $(\mathcal{Y}^2, \mathcal{Y}^{\otimes 2})$ satisfying the following conditions on the marginals: for all $(x, x') \in \mathcal{X}^2$ and $A \in \mathcal{Y}$,

$$\bar{H}((x, x'); A \times \mathcal{Y}) = H(x, A), \quad \bar{H}((x, x'); \mathcal{Y} \times A) = H(x', A). \tag{8.30}$$

Let $C \in \mathcal{Y}^{\otimes 2}$ such that $\bar{H}((x, x'); C) \neq 0$ and the chain $\left\{ Z_t = (X_t, X'_t, U_t) \right\}_{t \in \mathbb{Z}}$ on the “extended” space $(\mathcal{X}^2 \times 0, 1, \mathcal{X}^{\otimes 2} \otimes \mathcal{P}(0, 1))$ with transition kernel \bar{Q} implicitly

defined as follows. Given $Z_t = (x, x', u) \in \mathbf{X}^2 \times \{0, 1\}$, draw (Y_{t+1}, Y'_{t+1}) according to $\bar{H}((x, x'); \cdot)$ and set

$$X_{t+1} = f_{Y_{t+1}}(x), \quad X'_{t+1} = f_{Y'_{t+1}}(x'),$$

$$U_{t+1} = \mathbf{1}_C(Y_{t+1}, Y'_{t+1}),$$

$$Z_{t+1} = (X_{t+1}, X'_{t+1}, U_{t+1}).$$

The conditions on the marginals of \bar{H} , given by (8.30) also imply conditions on the marginals of \bar{Q} : for all $A \in \mathcal{X}$ and $z = (x, x', u) \in \mathbf{X}^2 \times \{0, 1\}$,

$$\bar{Q}(z; A \times \mathbf{X} \times \{0, 1\}) = Q(x, A), \quad \bar{Q}(z; \mathbf{X} \times A \times \{0, 1\}) = Q(x', A). \quad (8.31)$$

For $z = (x, x', u) \in \mathbf{X}^2 \times \{0, 1\}$, write

$$\alpha(x, x') = \bar{Q}(z; \mathbf{X}^2 \times \{1\}) = \bar{H}((x, x'); C) \neq 0. \quad (8.32)$$

The quantity $\alpha(x, x')$ is thus the probability of the event $\{U_1 = 1\}$ conditionally on Z_0 , taken on $Z_0 = z$. Denote by Q^\sharp the kernel on $(\mathbf{X}^2, \mathcal{X}^{\otimes 2})$ defined by: for all $z = (x, x', u) \in \mathbf{X}^2 \times \{0, 1\}$ and $A \in \mathcal{X}^{\otimes 2}$,

$$Q^\sharp((x, x'); A) = \frac{\bar{Q}(z; A \times \{1\})}{\bar{Q}(z; \mathbf{X}^2 \times \{1\})}$$

so that using (8.32),

$$\bar{Q}(z; A \times \{1\}) = \alpha(x, x') Q^\sharp((x, x'); A). \quad (8.33)$$

This shows that $Q^\sharp((x, x'); \cdot)$ is the distribution of (X_1, X'_1) conditionally on $(X_0, X'_0, U_1) = (x, x', 1)$.

Assumptions and Results of the Alternative Markov Chain Approach Without Irreducibility

In what follows, if $(\mathbf{E}, \mathcal{E})$ is a measurable space, ξ a probability distribution on $(\mathbf{E}, \mathcal{E})$, and R a Markov kernel on $(\mathbf{E}, \mathcal{E})$, denote by \mathbb{P}_ξ^R the probability induced on $(\mathbf{E}^{\mathbb{N}}, \mathcal{E}^{\otimes \mathbb{N}})$ by a Markov chain with transition kernel R and initial distribution ξ . Denote by \mathbb{E}_ξ^R the associated expectation. Consider the following assumptions.

(A1) The Markov kernel Q is weak Feller. Moreover, there exist a compact set $C \in \mathcal{X}, (b, \varepsilon) \in \mathbb{R}_*^+ \times \mathbb{R}_*^+$ and a function $V : \mathbf{X} \rightarrow \mathbb{R}^+$ such that

$$QV \leq V - \varepsilon + b\mathbf{1}_C.$$

(A2) The Markov kernel Q has a reachable point.

(A3) There exists a kernel \bar{Q} on $(\mathcal{X}^2 \times \{0, 1\}, \mathcal{X}^{\otimes 2} \otimes \mathcal{P}(\{0, 1\}))$, a kernel Q^\sharp on $(\mathcal{X}^2, \mathcal{X}^{\otimes 2})$, and a measurable function $\alpha : \mathcal{X}^2 \rightarrow [1, \infty)$, and real numbers $(D, \zeta_1, \zeta_2, \rho) \in (\mathbb{R}^+)^3 \times (0, 1)$ such that for all $(x, x') \in \mathcal{X}^2$,

$$1 - \alpha(x, x') \leq d(x, x')W(x, x') \tag{8.34}$$

$$E_{\delta_x \otimes \delta_{x'}}^{Q^\sharp} [d(X_t, X'_t)] \leq D\rho^t d(x, x') \tag{8.35}$$

$$E_{\delta_x \otimes \delta_{x'}}^{Q^\sharp} [d(X_t, X'_t)W(X_t, X'_t)] \leq D\rho^t d^{\zeta_1}(x, x')W^{\zeta_2}(x, x'). \tag{8.36}$$

Moreover, for all $x \in \mathcal{X}$, there exists $\gamma_x > 0$ such that

$$\sup_{x' \in B(x, \gamma_x)} W(x, x') < \infty$$

Some practical conditions from checking (8.35) and (8.36) in (A3) can be denoted.

Lemma 8.1 *Assume that either (i) or (ii) or (iii) (defined below) holds.*

(i) *There exist $(\rho, \beta) \in (0, 1) \times \mathbb{R}$ such that for all $(x, x') \in \mathcal{X}^2$*

$$d(X_1, X'_1) \leq \rho d(x, x'), \quad P_{\delta_x \otimes \delta_{x'}}^{Q^\sharp} - a.s. \tag{8.37}$$

$$Q^\sharp W \leq W + \beta \tag{8.38}$$

(ii) *Equation (8.35) holds and W is bounded.*

(iii) *Equation (8.35) holds and there exist $0 < \alpha < \alpha'$ and $\beta \in \mathbb{R}^+$ such that for all $(x, x') \in \mathcal{X}^2$*

$$d(x, x') \leq W^\alpha(x, x')$$

$$Q^\sharp W^{1+\alpha'} \leq W^{1+\alpha'} + \beta$$

Then, (8.35) and (8.36) hold.

Assumption (A1) implies, by Tweedie [55], that the Markov kernel Q admits at least one stationary distribution. Assumptions (A2)–(A3) are then used to show that this stationary distribution is unique.

Note that assumptions (A1)–(A2) are the same as those of Theorems 8.9 and 8.10 of Appendix Section “Feller Conditions” and they can be proved for each observation driven model as has been done for the GARMA model in Sect. 8.5.1;

assumption (A3) weakens the Lipschitz condition (8.18) by introducing a function W in (8.34). This allows to treat models which do not satisfy the Lipschitz condition (8.18); for example the log-linear Poisson autoregression of [28], see Sect. 8.5.2.

Theorem 8.11 *Assume that (A1)–(A3) hold. Then, the Markov kernel Q in (8.19) admits a unique invariant probability measure.*

Proposition 8.1 *Assume that the Markov kernel Q admits a unique invariant probability measure. Then, there exists a strict-sense stationary ergodic process on \mathbb{Z} , $\{Y_t\}_{t \in \mathbb{Z}}$, the solution to the recursion (8.19).*

These results can be found in [23].

Main Proofs

Proof of Theorem 8.2

Following [42], Theorem 8.9 is applied to the chain $\{g(\mu_t)\}_{t \in \mathbb{N}}$ to show that it has a stationary distribution; this implies the same result for the chain $\{\mu_t\}_{t \in \mathbb{N}}$. The state space $S = \mathbb{R}$ of $\{g(\mu_t)\}_{t \in \mathbb{N}}$ is a locally compact complete separable metric space with Borel σ -field. A drift condition for $\{g(\mu_t)\}_{t \in \mathbb{N}}$ is given under the conditions of Theorem 8.1, for the compact set $A = [-M, M]$ (the drift condition holds when the perturbation $m = 0$). All that remains is to show that the chain $\{g(\mu_t)\}_{t \in \mathbb{N}}$ is weak Feller. Let $X_t = g(\mu_t)$. For $X_0 = x$, the GARMA model can be rewritten as

$$X_1(x) = \gamma + \phi(g(Y_0^*(g^{-1}(x))) - \gamma) + \theta(g(Y_0^*(g^{-1}(x))) - x).$$

Since g^{-1} is continuous, $Y_0(g^{-1}(x)) \Rightarrow Y_0(g^{-1}(x'))$ as $x \rightarrow x'$. Since the $*$ that maps Y_0 to the domain of g is continuous, it follows that $Y_0^*(g^{-1}(x)) \Rightarrow Y_0^*(g^{-1}(x'))$ as $x \rightarrow x'$. Since g is continuous, then $g(Y_0^*(g^{-1}(x))) \Rightarrow g(Y_0^*(g^{-1}(x')))$. So $X_1(x) \Rightarrow X_1(x')$ as $x \rightarrow x'$, showing the weak Feller property. This ends the proof.

Proof of Theorem 8.4

The proof is based on the results of Appendix Sections “Coupling Construction”–“Assumptions and Results of the Alternative Markov Chain Approach Without Irreducibility”. The conditions (A1)–(A2) for the log-linear Poisson autoregression are proved as in Sect. 8.5.1 for the GARMA model. We report the proof of (A3).

Lemma 8.2 *If $|a + b| \vee |a| \vee |b| < 1$, then (A3) holds.*

Proof Define \bar{Q} as the transition kernel Markov chain $\{Z_t\}_{t \in \mathbb{Z}}$ with $Z_t = (X_t, X'_t, U_t)$ in the following way. Given $Z_t = (x, x', u)$, if $x \leq x'$, draw independently $Y_{t+1} \sim \mathcal{P}(e^x)$ and $V_{t+1} \sim \mathcal{P}(e^{x'} - e^x)$ and set $Y'_{t+1} = Y_{t+1} + V_{t+1}$. Otherwise, draw independently $Y'_{t+1} \sim \mathcal{P}(e^{x'})$ and $V_{t+1} \sim \mathcal{P}(e^x - e^{x'})$ and set $Y_{t+1} = Y'_{t+1} + V_{t+1}$.

$$\begin{aligned} X_{t+1} &= d + a x + b \ln(Y_{t+1} + 1), \\ X'_{t+1} &= d + a x' + b \ln(Y'_{t+1} + 1), \\ U_{t+1} &= \mathbf{1}_{Y_{t+1}=Y'_{t+1}} = \mathbf{1}_{V_{t+1}=0}, \\ Z_{t+1} &= (X_{t+1}, X'_{t+1}, U_{t+1}) \end{aligned}$$

where \bar{Q} satisfies the marginal condition (8.31). Moreover, define for all $x^\sharp = (x, x') \in \mathbb{X}^2, Q^\sharp(x^\sharp, \cdot)$ as the law of (X_1, X'_1) where

$$\begin{aligned} X_1 &= d + a x + b \ln(Y + 1), \quad Y \sim \mathcal{P}(e^{x \wedge x'}), \\ X'_1 &= d + a x' + b \ln(Y + 1), \end{aligned} \tag{8.39}$$

and set for all $x^\sharp = (x, x') \in \mathbb{R}^2$,

$$\alpha(x^\sharp) = \left\{ \exp -e^{x \vee x'} + e^{x \wedge x'} \right\}.$$

Then, \bar{Q} and Q^\sharp satisfy (8.33). Using twice $1 - e^{-u} \leq u$, it follows that

$$\begin{aligned} 1 - \alpha(x^\sharp) &= 1 - \left\{ \exp -e^{x \vee x'} + e^{x \wedge x'} \right\} \leq e^{x \vee x'} - e^{x \wedge x'} \\ &= e^{x \vee x'} (1 - e^{-|x-x'|}) \leq W(x, x') |x - x'| \end{aligned}$$

with $W(x, x') = e^{|x \vee x'|}$ so that (8.34) holds true. To check (8.35) and (8.36), Lemma 8.1 is applied, by checking option (i). Note first that

$$\mathbb{P}_{\delta_x \otimes \delta_{x'}}^{Q^\sharp} \{ |X_1 - X'_1| = |a||x - x'| \} = 1, \tag{8.40}$$

so that (8.37) is satisfied. To check (8.38), it can be shown that

$$\lim_{|x \vee x'| \rightarrow \infty} \frac{Q^\sharp W(x, x')}{W(x, x')} = 0 \tag{8.41}$$

and for all $M > 0$,

$$\sup_{|x \vee x'| \leq M} Q^\sharp W(x, x') < \infty \tag{8.42}$$

Without loss of generality, assume $x \leq x'$. Using (8.39) provides

$$Q^\sharp W(x, x') = \mathbb{E} \left(e^{|X_1| \vee |X'_1|} \right) \leq \mathbb{E} \left(e^{|X_1|} \right) + \mathbb{E} \left(e^{|X'_1|} \right). \quad (8.43)$$

First, consider the second term of the right-hand side of (8.43),

$$\mathbb{E} \left(e^{|X'_1|} \right) \leq e^{|d|} \mathbb{E} \left(e^{ax' + b \ln(1+Y)} \right). \quad (8.44)$$

Noting that if u and v have different signs or if $v = 0$, then $|u + v| \leq |u| \vee |v|$. Otherwise, $|u + v| = (u + v)\mathbf{1}_{v>0} \vee (-u - v)\mathbf{1}_{v<0}$. This implies that

$$e^{|u+v|} \leq e^{|u|} + e^{|v|} + e^{u+v}\mathbf{1}_{v>0} + e^{-u-v}\mathbf{1}_{v<0}.$$

and plugging this into (8.44),

$$\begin{aligned} \mathbb{E}(e^{|X'_1|}) &\leq e^{|d|} \left(e^{|a||x'|} + \mathbb{E}[(1+Y)^{|b|}] + e^{ax'} \mathbb{E}[(1+Y)^b] \mathbf{1}_{b>0} \right. \\ &\quad \left. + e^{-ax'} \mathbb{E}[(1+Y)^{-b}] \mathbf{1}_{b<0} \right). \end{aligned}$$

Note that for all $\gamma \in [0, 1]$,

$$\mathbb{E}[(1+Y)^\gamma] \leq [\mathbb{E}(1+Y)]^\gamma = (1+e^x)^\gamma \leq 1 + e^{\gamma x} \leq 1 + e^{\gamma x'}.$$

Moreover, since $|b| \in [0, 1]$, $b\mathbf{1}_{b>0} \in [0, 1]$ and $-b\mathbf{1}_{b<0} \in [0, 1]$. Therefore,

$$\begin{aligned} \mathbb{E}(e^{|X'_1|}) &\leq e^{|d|} \left(e^{|a||x'|} + 1 + e^{|b||x|} + e^{ax'} (1 + e^{bx'}) \mathbf{1}_{b>0} + e^{-ax'} (1 + e^{-bx'}) \mathbf{1}_{b<0} \right) \\ &\leq e^{|d|} \left(e^{|a||x'|} + 1 + e^{|b||x|} + e^{|a||x'|} + e^{|a+b||x'|} \right) \\ &\leq e^{|d|} \left(1 + 4e^{\gamma(|x| \vee |x'|)} \right), \end{aligned}$$

where $\gamma = |a| \vee |b| \vee |a+b| < 1$. The first term of the right hand side of (8.43) is treated as the second term by setting $x' = x$. So

$$\mathbb{E}(e^{|X_1|}) \leq e^{|d|} \left(1 + 4e^{\gamma(|x| \vee |x'|)} \right),$$

so that using (8.43),

$$Q^\sharp W(x, x') \leq 2e^{|d|} \left(1 + 4e^{\gamma(|x| \vee |x'|)} \right).$$

Since $\gamma \in (0, 1)$ and $W(x, x') = e^{|\gamma \vee |x'|}$, and (8.43) clearly implies (8.41) and (8.42), this proves (A3) and together with (A1)–(A2) provides stationarity conditions for the process $\{Y_t\}$ of Theorem 8.4.

Computational Aspects

The replication code for the application in Sect. 8.7 is available at https://github.com/mirkoarmillotta/covid_code. First, a function for the log-likelihood and the gradient of the log-linear Poisson autoregression is provided. The code for the other models works in a similar way and it is available upon request. Then, a function to perform the QMLE is presented. Finally, we give the code for the COVID-19 example and the relative plots. The code to perform the PIT is due to [17] and it is available in the reference therein.

References

1. Ahmad, A., Francq, C.: Poisson QMLE of count time series models. *J. Time Ser. Anal.* **37**, 291–314 (2016)
2. Al-Osh, M.A., Alzaid, A.A.: First-order integer-valued autoregressive (INAR (1)) process. *J. Time Ser. Anal.* **8**, 261–275 (1987)
3. Alzaid, A.A., Al-Osh, M.: An integer-valued pth-order autoregressive structure (INAR (p)) process. *J. Appl. Probab.* **27**, 314–324 (1990)
4. Basawa, I.V., Prakasa Rao, B.L.S.: *Statistical Inference for Stochastic Processes. Probability and Mathematical Statistics*. Academic, Cambridge [Harcourt Brace Jovanovich, Publishers], London (1980)
5. Benjamin, M., Rigby, R., Stasinopoulos, D.M.: Generalized autoregressive moving average models. *J. Amer. Stat. Assoc.* **98**(461), 214–223 (2003)
6. Billingsley, P.: *Probability and Measure*, 3rd edn. Wiley, Hoboken (1995)
7. Bollerslev, T.: Generalized autoregressive conditional heteroskedasticity. *J. Econ.* **31**(3), 307–327 (1986)
8. Box, G.E.P., Jenkins, G.M.: *Time Series Analysis: Forecasting and Control*. Holden Day, San Francisco (1970)
9. Box, G.E.P., Jenkins, G.M.: *Time Series Analysis: Forecasting and Control*. Prentice-Hall, Hoboken (1976)
10. Breen, W., Glosten, L.R., Jagannathan, R.: Economic significance of predictable variations in stock index returns. *J. Finance* **44**(5), 1177–1189 (1989)
11. Brockwell, P.J., Davis, R.A.: *Time Series: Theory and Methods*. Springer Series in Statistics. Springer, Berlin (1991)
12. Christou, V., Fokianos, K.: Quasi-likelihood inference for negative binomial time series models. *J. Time Ser. Anal.* **35**, 55–78 (2014)
13. Christou, V., Fokianos, K.: On count time series prediction. *J. Stat. Comput. Simul.* **85**(2), 357–373 (2015)
14. Clark, N.J., Kaiser, M.S., Dixon, P.M.: A spatially correlated auto-regressive model for count data (2018). Preprint arXiv:1805.08323
15. Cox, D.R.: Statistical analysis of time series: some recent developments. *Scand. J. Stat.* **8**, 93–115 (1981)

16. Creal, D., Koopman, S.J., Lucas, A.: Generalized autoregressive score models with applications. *J. Appl. Econ.* **28**(5), 777–795 (2013)
17. Czado, C., Gneiting, T., Held, L.: Predictive model assessment for count data. *Biometrics* **65**(4), 1254–1261 (2009)
18. Davis, R.A., Liu, H.: Theory and inference for a class of nonlinear models with application to time series of counts. *Stat. Sinica* **26**(4), 1673–1707 (2016)
19. Davis, R.A., Dunsmuir, W.T.M., Streett, S.B.: Observation-driven models for Poisson counts. *Biometrika* **90**(4), 777–790 (2003)
20. Davis, R.A., Dunsmuir, W.T.M., Streett, S.B.: Maximum likelihood estimation for an observation driven model for Poisson counts. *Methodol. Comput. Appl. Probab.* **7**(2), 149–159 (2005)
21. Davis, R.A., Holan, S.H., Lund, R., Ravishanker, N.: *Handbook of Discrete-valued Time Series*. CRC Press, Boca Raton (2016)
22. Diaconis, P., Freedman, D.: Iterated random functions. *SIAM Rev.* **41**(1), 45–76 (1999)
23. Douc, R., Doukhan, P., Moulines, E.: Ergodicity of observation-driven time series models and consistency of the maximum likelihood estimator. *Stoch. Process. Appl.* **123**(7), 2620–2647 (2013)
24. Douc, R., Fokianos, K., Moulines, E.: Asymptotic properties of quasi-maximum likelihood estimators in observation-driven time series models. *Electron. J. Stat.* **11**(2), 2707–2740 (2017)
25. Dunsmuir, W., Scott, D.: The GLARMA package for observation-driven time series regression of counts. *J. Stat. Softw.* **67**(7), 1–36 (2015)
26. Engle, R.F.: Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* **50**(4), 987–1007, 06 (1982)
27. Ferland, R., Latour, A., Oraichi, D.: Integer-valued GARCH process. *J. Time Ser. Anal.* **27**, 923–942 (2006)
28. Fokianos, K., Tjøstheim, D.: Log-linear Poisson autoregression. *J. Multivar. Anal.* **102**, 563–578 (2011)
29. Fokianos, K., Kedem, B., et al.: Regression theory for categorical time series. *Stat. Sci.* **18**(3), 357–376 (2003)
30. Fokianos, K., Rahbek, A., Tjøstheim, D.: Poisson autoregression. *J. Amer. Stat. Assoc.* **104**, 1430–1439 (2009)
31. Fokianos, K., Støve, B., Tjøstheim, D., Doukhan, P.: Multivariate count autoregression. *Bernoulli* **26**(1), 471–499 (2020)
32. Gneiting, T., Raftery, A.E.: Strictly proper scoring rules, prediction, and estimation. *J. Amer. Stat. Assoc.* **102**(477), 359–378 (2007)
33. Gneiting, T., Balabdaoui, F., Raftery, A.E.: Probabilistic forecasts, calibration and sharpness. *J. Roy. Stat. Soc. Ser. B (Stat. Methodol.)* **69**(2), 243–268 (2007)
34. Gorgi, P.: Beta-negative binomial auto-regressions for modelling integer-valued time series with extreme observations. *J. Roy. Stat. Soc. Ser. B (Stat. Methodol.)* **82**, 1325–1347 (2020)
35. Hairer, M., Mattingly, J.C.: Ergodicity of the 2D Navier-Stokes equations with degenerate stochastic forcing. *Ann. Math.* **164**(3), 993–1032 (2006)
36. Harvey, A.C.: *Dynamic Models for Volatility and Heavy Tails: With Applications to Financial and Economic Time Series*. Cambridge University Press, Cambridge (2013)
37. Hayashi, F.: *Econometrics*. Princeton University Press, Princeton (2000)
38. Heyde, C.C.: A general approach to optimal parameter estimation. In: *Quasi-Likelihood and Its Application*. Springer Series in Statistics. Springer, New York (1997)
39. Ho, S.-L., Xie, M., Goh, T.N.: A comparative study of neural network and Box-Jenkins ARIMA modeling in time series prediction. *Comput. Ind. Eng.* **42**(2–4), 371–375 (2002)
40. Kauppi, H., Saikkonen, P.: Predicting U.S. recessions with dynamic binary response models. *Rev. Econ. Stat.* **90**(4), 777–791 (2008)
41. Li, W.K.: Time series models based on generalized linear models: some further results. *Biometrics* **50**(2), 506–511 (1994)
42. Matteson, D.S., Woodard, D.B., Henderson, S.G.: Stationarity of generalized autoregressive moving average models. *Electron. J. Stat.* **5**, 800–828 (2011)

43. McCullagh, P., Nelder, J.A.: *Generalized Linear Models*, 2nd edn. Chapman & Hall, Boca Raton (1989)
44. Meyn, S., Tweedie, R.L., Glynn, P.W.: *Markov Chains and Stochastic Stability*, 2nd edn. Cambridge University Press, Cambridge (2009)
45. Moysiadis, T., Fokianos, K.: On binary and categorical time series models with feedback. *J. Multivar. Anal.* **131**, 209–228 (2014)
46. Neumann, M.H.: Absolute regularity and ergodicity of Poisson count processes. *Bernoulli* **17**(4), 1268–1284 (2011)
47. Rydberg, T.H., Shephard, N.: Dynamics of trade-by-trade price movements: decomposition and models. *J. Financ. Econ.* **1**(1), 2–25 (2003)
48. Scotto, M.G., Weiß, C.H., Gouveia, S.: Thinning-based models in the analysis of integer-valued time series: a review. *Stat. Modell.* **15**(6), 590–618 (2015)
49. Sen, P., Roy, M., Pal, P.: Application of ARIMA for forecasting energy consumption and GHG emission: a case study of an Indian pig iron manufacturing organization. *Energy* **116**, 1031–1038 (2016)
50. Startz, R.: Binomial autoregressive moving average models with an application to U.S. recessions. *J. Business Econ. Stat.* **26**(1), 1–8 (2008)
51. Steutel, F.W., van Harn, K.: Discrete analogues of self-decomposability and stability. *Ann. Probab.* **7**(5), 893–899 (1979)
52. Taylor, S.: *Modeling Financial Time Series*. Wiley, Hoboken (1986)
53. Teräsvirta, T.: Specification, estimation, and evaluation of smooth transition autoregressive models. *J. Amer. Stat. Assoc.* **89**(425), 208–218 (1994)
54. Tong, H., Lim, K.S.: Threshold autoregression, limit cycles and cyclical data—with discussion. *J. Roy. Stat. Soc. Ser. B Stat. Methodol.* **42**(3), 245–292 (1980)
55. Tweedie, R.L.: Invariant measures for Markov chains with no irreducibility assumptions. *J. Appl. Probab.* **25**(A), 275–285 (1988)
56. Walker, G.T.: On periodicity in series of related terms. *Proc. R. Soc. Lond. Ser. A Contain. Papers Math. Phys. Char.* **131**(818), 518–532 (1931)
57. Wang, Y., Wang, J., Zhao, G., Dong, Y.: Application of residual modification approach in seasonal ARIMA for electricity demand forecasting: a case study of China. *Energy Policy* **48**, 284–294 (2012)
58. Weiß, C.H., Feld, M.H.-J.M., Khan, N.M., Sunecher, Y.: Inarma modeling of count time series. *Stats* **2**(2), 284–320 (2019)
59. Yule, G.U.: On a method of investigating periodicities disturbed series, with special reference to Wolfer’s sunspot numbers. *Philos. Trans. R. Soc. Lond. Ser. A Contain. Pap. Math. Phys. Char.* **226**(636–646), 267–298 (1927)
60. Zeger, S.L.: A regression model for time series of counts. *Biometrika* **75**, 621–629 (1988)
61. Zeger, S.L., Liang, K.-Y.: Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**, 121–130 (1986)
62. Zheng, T., Xiao, H., Chen, R.: Generalized ARMA models with martingale difference errors. *J. Econ.* **189**(2), 492–506 (2015)

Chapter 9

Advances in Maximum Likelihood Estimation of Fixed-Effects Binary Panel Data Models



Francesco Valentini, Claudia Pigni, and Francesco Bartolucci

9.1 Introduction

Panel data play a major role in applied research and the related literature has been rapidly growing in recent decades. The use of this type of data has become common practice in a wide range of applications across many fields [see, among others, 39, 43, 52, 75].

This data structure, consisting of longitudinal observations over time for every unit in the sample, is more informative than cross-sectional data. As such, it makes it possible to formulate statistical models that account for unobserved heterogeneity, which comes into play whenever the behavior of an individual is influenced by characteristics that cannot be directly measured and controlled for by the analyst, such as individual preferences, innate abilities, or risk attitudes. On the contrary, unobserved heterogeneity often gives rise to identification issues with cross-sectional data, especially if correlated with the model covariates. This happens, for instance, when the individual latent trait represents the degree of labor market attachment and preference for a high value career in a model for female labor supply, as this latent trait strongly affects both the response variable and one of its main determinants, such as family composition. In this case, the presence of unobserved heterogeneity compromises the correct identification of the effect of fertility on labor force participation [49].

F. Valentini (✉) · C. Pigni

Department of Economics and Social Sciences, Marche Polytechnic University, Ancona, Italy
e-mail: f.valentini@pm.univpm.it; c.pigni@univpm.it

F. Bartolucci

Department of Economics, University of Perugia, Perugia, Italy
e-mail: francesco.bartolucci@unipg.it

© Springer Nature Switzerland AG 2023

M. Kateri, I. Moustaki (eds.), *Trends and Challenges in Categorical Data Analysis*,
Statistics for Social and Behavioral Sciences,
https://doi.org/10.1007/978-3-031-31186-4_9

275

In order to properly account for individual-specific latent traits, applications whose aim is to uncover causal links largely rely on the use of panel data and of the so-called fixed-effects approach, which consists in modeling the individual unobserved heterogeneity through fixed parameters to be estimated. This strategy is frequently adopted in applied econometrics [5, 52, 75], while the mainstream approach in applied statistics is based on random-effects models [54, 69, 72]. Fixed-effects models have the advantage of avoiding distributional assumptions on the unobserved heterogeneity parameters and allow them to depend on the covariates in a nonparametric way. Practitioners make extensive use of linear models and several methodological approaches have been developed in this context to account for unobserved heterogeneity. We refer the reader to the books cited at the beginning of this section for a thorough description of estimation and inferential procedures for linear panel data models.

With binary, discrete, or somehow limited response variables, the linear panel data model must be reformulated. One possibility is to rely on a Generalized Linear Model (GLM) formulation [57] based on a distribution belonging to the exponential family for the response variable and a suitable link function relating the expected value of this distribution to the linear predictor. However, methodological issues arise with the application of the fixed-effects approach because the Maximum Likelihood (ML) estimator is inconsistent due to the presence of incidental parameters, that is, nuisance parameters whose number increases with the sample size [55, 61].

An additional challenge, especially in economic analyses with binary data, is related to the dynamic formulation that includes the lagged dependent variable in the set of covariates. This formulation allows us to identify the so-called *true state dependence* [48], that is, how the experience of an event in the past affects the probability of the same event occurring in the future. This effect is separate from that of the unobserved heterogeneity, which has an influence on the probability of that same event at all times. Dynamic binary choice models are adopted in a wide range of economic applications aimed at investigating individual decisions, such as labor market participation [50, 53], portfolio choices and financial conditions of households [2, 27, 45], migrants' remittances [24], and firms' access to credit [62].

Fixed-effects binary panel data models, with both a static and a dynamic formulation, have been developed in the recent econometric literature, which has been growing rapidly in the last two decades. Early surveys are provided by Arellano [6] and Arellano and Hahn [9], while the recent paper by Fernández-Val and Weidner [42] extensively reviews approaches and methods concerning only models for long panels, where the number of time occasions is relatively large. However, models for fixed- T panel data are still interesting due to the large availability of data sets of this type. For instance, most national household and workforce surveys are based on a rotating sampling scheme where subjects are interviewed a limited number of times.

The present chapter provides an extensive review of estimation approaches to fixed-effects binary choice models for longitudinal data where subjects are observed on a limited number of time occasions. This survey covers the two main groups of fixed-effects estimation approaches, differing as to how they deal with the

inconsistency of the ML estimator arising from the incidental parameter problem: target-corrected estimators on the one hand, aimed at reducing the order of the bias, and conditional inference on the other, based on conditioning on sufficient statistics, which directly eliminates the source of this bias and is specific to the logit model. In addition, special attention will be devoted to dynamic models, which have been the focus of the majority of the most recent contributions.

The chapter is organized as follows. Section 9.2 establishes notation for general panel data models for continuous responses. Section 9.3 presents binary choice panel data models. Section 9.4 describes the fixed-effects approach and discusses the related incidental parameter problem. Section 9.5 surveys estimation approaches based on bias reduction techniques and Sect. 9.6 reviews the conditional inference approach. Section 9.7 presents a simulation study comparing the finite sample performance of some of the reviewed approaches for the fixed-effects dynamic logit model and Sect. 9.8 illustrates an empirical application on female labor supply. Section 9.9 provides a brief review of the software packages available to estimate the models described in this work. Finally, Sect. 9.10 provides concluding remarks.

9.2 Preliminaries

We consider n units, indexed with $i = 1, \dots, n$, observed at time occasions $t = 1, \dots, T$. This is the case of a balanced panel dataset, where all units are observed at the same time occasions. For simplicity, we do not consider explicitly unbalanced panel datasets, where each subject i is observed for a specific number T_i of time occasions, but the approaches considered throughout the chapter can be directly extended to this case.

A general specification for a linear panel data model is of the form

$$y_{it} = \alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta} + \varepsilon_{it}, \quad (9.1)$$

where y_{it} is the dependent variable and α_i represents the individual traits that cannot be directly observed or measured and then allows us to account for unit-specific time-invariant heterogeneity. Moreover, \mathbf{x}_{it} is a $k \times 1$ vector of exogenous covariates and $\boldsymbol{\beta}$ is the corresponding vector of regression parameters. Finally, ε_{it} is a random variable with zero mean and variance equal to σ_ε^2 , representing idiosyncratic shocks.

Different specifications and estimation approaches for (9.1) can be adopted according to the hypotheses formulated on the individual unobserved heterogeneity parameter α_i . In a nutshell, the fixed-effects approach consists in treating these individual intercepts as fixed parameters to be estimated. In practice, this usually amounts to treating the subject identifier as a categorical variable or, equivalently, applying the so-called within-group transformation to eliminate the individual intercepts. Alternatively, α_i can be eliminated by taking the first differences of (9.1). The main advantage is that consistency of the OLS estimator of $\boldsymbol{\beta}$ in these formulations does not require any distributional assumption for the α_i , which

are also allowed to be correlated with the model covariates in a nonparametric way. In contrast, if α_i is assumed to be a random variable, usually distributed as $N(0, \sigma_\alpha^2)$, expression (9.1) defines a random-effects model. A Generalized Least Squares or ML estimator of β can be easily derived and its consistency relies on the independence between α_i and the model covariates.

The inclusion of the lagged response variable in the set of regressors in (9.1) complicates matters, as the time-constant dependence between the response variable and the unobserved heterogeneity parameter gives rise to an endogeneity problem. Within the random-effects approach, applying sequential factorization to write the likelihood recursively leaves us with the problem of handling the correlation between y_{i0} and α_i , the solutions to which will be briefly mentioned in the next section with reference to binary choice models. With the fixed-effects formulation, the endogeneity problem is tackled by the instrumental variable approach, and its generalizations based on the Generalized Method of Moments [4, 7, 8, 25].

The econometric literature on linear panel data models is indeed vast; we refer the reader to [5, 75], and [52], among others, for details on model formulations and related inferential strategies, as they are not the object of this review. However, some related extensions are worth mentioning as they are pervasive in the applied statistical literature. The first is represented by Linear Mixed Models (LMMs) for longitudinal data [54, 72]. LMMs can be seen as a generalization of random-effects models, in that slope parameters are also considered as random variables, further to the intercept α_i . Frequentist estimation is usually carried out by ML under the same independence assumption between the unobserved heterogeneity and covariates as the traditional random-effects model.

Another approach in a different direction is the generalization represented by finite-mixture models [59], in which every random effect is assumed to be discrete. This approach can be regarded as semi-parametric because a discrete distribution may adequately approximate any continuous distribution. This only partially relaxes the distributional assumption on the random effects, which are still required to be independent of the regressors. An extension in this respect is the concomitant variable approach [37] applied to finite-mixture models [73], where the probability of each mixture component is allowed to depend on individual time-constant covariates.

9.3 Binary Choice Panel Data Models

A convenient way to represent panel data models for binary dependent variables is the latent variable formulation, according to which the response variable for individual i at time t , y_{it} , is assumed to depend on a latent continuous variable y_{it}^* , representing the propensity of an event to occur and is conceived as a function

of a linear index. It is assumed that

$$\begin{aligned} y_{it} &= \mathbb{1}\{y_{it}^* \geq \tau\}, \\ y_{it}^* &= \alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta} + \varepsilon_{it}, \end{aligned}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function, so that the outcome y_{it} assumes values 1 or 0 according to whether or not the latent variable crosses the threshold τ , which is usually fixed at 0. As for the linear model, α_i represents the unobserved heterogeneity and \mathbf{x}_{it} collects the exogenous explanatory variables associated with the parameter vector $\boldsymbol{\beta}$. Finally, ε_{it} is a zero mean and constant variance error component representing the effect of idiosyncratic shocks.

It is straightforward to extend the latent variable formulation to accommodate dynamic models, as it amounts to augmenting the set of explanatory variables by the lagged dependent variable. We have that

$$y_{it} = \mathbb{1}\{\alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta} + y_{i,t-1}\gamma + \varepsilon_{it} > 0\}, \quad (9.2)$$

where γ measures the *true state dependence*, as defined by Heckman [48]. Throughout this work we will consider γ as time-invariant. We also assume y_{i0} to be the known initial observation for the i -th subject. For ease of exposition, what follows is based on the static formulation even though most results are still valid for dynamic models.

The initial latent variable formulation implies that the probability of $y_{it} = 1$ given α_i and \mathbf{x}_{it} is

$$p(y_{it} = 1|\alpha_i, \mathbf{x}_{it}) = F(\alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta}), \quad (9.3)$$

where $F(\cdot)$ denotes a general functional form for the inverse link function depending on the distributional assumption formulated on the idiosyncratic component. For instance, by assuming a standard normal distribution for ε_{it} , we have the probit model according to which $p(y_{it} = 1|\alpha_i, \mathbf{x}_{it}) = \Phi(\alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta})$, where $\Phi(\cdot)$ is the standard normal cdf, while from a standard logistic cdf the logit model derives, according to which

$$p(y_{it} = 1|\alpha_i, \mathbf{x}_{it}) = \frac{\exp(\alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta})}{1 + \exp(\alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta})}.$$

Notice that in both the probit and logit models, the variance of the error term ε_{it} is known and equal to 1 or $\pi^2/3$, respectively. Differently from the linear model, the variance of ε_{it} is not identified and therefore cannot be treated as a parameter to be estimated. This is because any positive value for such variance is coherent with the same frequency of 0s and 1s observed in the sample.

As in the case of continuous response variables, modeling and estimation approaches differ according to the hypotheses on the unobserved heterogeneity α_i .

Assuming that every α_i is a random variable, typically with distribution $N(0, \sigma_\alpha^2)$, also yields a random-effects model and specular extensions to include random slopes are represented by Generalized Linear Mixed Models (GLMMs); see [69]. Parameters can be estimated by ML and, as for the linear model, consistency relies on the independence between α_i and the covariates. A partial extension is considered by Mundlak [60] and Chamberlain [30] for models where α_i is allowed to depend on the model covariates in a linear manner. In the same vein is the concomitant variable approach to latent class analysis for categorical data [13, 44], where the mass probabilities associated with the support points are allowed to depend parametrically on individual characteristics.

The random-effects approach to model discrete responses is widely employed in applied statistics [54], and is therefore not treated in this chapter. Nevertheless, it is worth recalling that random-effects models, differently from fixed-effects formulations, require appropriate handling of the initial observation y_{i0} , as it is correlated with the unobserved heterogeneity parameters α_i , whereas it is usually considered as given within fixed-effects approaches. A solution to the so-called “initial-conditions” problem was first put forward by Heckman [49], while a popular alternative is provided by Wooldridge [74]. Discussions and comparisons between these and other, less conventional, approaches are given by Arulampalam and Stewart [11], Akay [1], Rabe-Hesketh and Skrondal [65], and Skrondal and Rabe-Hesketh [67], while a review of the most recent extensions of Heckman’s approach is given by Lucchetti and Pigni [56].

The remainder of the chapter first discusses the fixed-effects approach and the incidental parameter problem that arises with nonlinear binary choice models and then reviews the most recent solutions.

9.4 Fixed-Effects Approach and Incidental Parameter Problem

In the following, we clarify how the incidental parameter problem arises within the ML estimation of binary fixed-effects models.

As anticipated in Sect. 9.1, the fixed-effects approach consists in modeling the time-invariant individual unobserved characteristics by means of fixed individual intercepts to be estimated along with the regression parameters. A feasible approach is to include a set of individual dummy variables among the covariates. It is worth stressing that, within the fixed-effects approach, it is possible to identify only parameters related to the time-varying explanatory variables. In fact, it would be impossible to simultaneously identify the individual intercept and the parameters for time-constant covariates.

In principle, the ML framework can be seen as a standard technique to obtain an estimate of the whole set of parameters. It is possible to build the log-likelihood function for the sample relying on the formulation in Eq. (9.3), as

$$\ell(\alpha_1, \dots, \alpha_n, \boldsymbol{\beta}) = \sum_{i=1}^n \ell_i(\alpha_i, \boldsymbol{\beta}), \tag{9.4}$$

$$\ell_i(\alpha_i, \boldsymbol{\beta}) = \sum_{t=1}^T \log p(y_{it} | \mathbf{x}_{it}; \alpha_i, \boldsymbol{\beta}),$$

where we define $\ell_i(\alpha_i, \boldsymbol{\beta})$ as the individual log-likelihood and where $p(y_{it} | \mathbf{x}_{it}; \alpha_i, \boldsymbol{\beta}) = F(\alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta})^{y_{it}} [1 - F(\alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta})]^{1-y_{it}}$. The ML estimator of the parameters is then obtained by maximizing the log-likelihood function in Eq. (9.4) w.r.t. the parameters $(\alpha_1, \dots, \alpha_n, \boldsymbol{\beta})'$.

The aforementioned technique consists in an optimization problem involving a $(k + n)$ -dimensional space and it is computationally cumbersome with a large number of individuals. In order to overcome this problem, we can obtain the same results by relying on the maximization of the concentrated log-likelihood, also known as profile log-likelihood. In this case, the ML estimator of $\boldsymbol{\beta}$ is derived by concentrating out the α_i , meaning that for any value of $\boldsymbol{\beta}$ the individual intercepts are evaluated at their ML estimates and then as the solution of

$$\hat{\alpha}_i(\boldsymbol{\beta}) = \operatorname{argmax}_{\alpha_i} \sum_{t=1}^T \log p(y_{it} | \mathbf{x}_{it}; \alpha_i, \boldsymbol{\beta}), \tag{9.5}$$

$$\hat{\boldsymbol{\beta}} = \operatorname{argmax}_{\boldsymbol{\beta}} \sum_{i=1}^n \sum_{t=1}^T \log p(y_{it} | \mathbf{x}_{it}; \hat{\alpha}_i(\boldsymbol{\beta}), \boldsymbol{\beta}). \tag{9.6}$$

Here each individual intercept estimate $\hat{\alpha}_i(\boldsymbol{\beta})$ is a function of $\boldsymbol{\beta}$ but it is easy to compute since it depends only on data specific to individual i . In this way, we face n optimization problems in a unidimensional space. Starting from a given value of $\boldsymbol{\beta}$, it suffices to iterate optimization problems in Eqs. (9.5) and (9.6) until convergence.

The maximization of the concentrated log-likelihood is also useful to understand how the incidental parameter problem [61] arises. From Eq. (9.5), it is clear that only T observations contribute to the ML estimator of each individual intercept so that it is impossible to consistently estimate the parameters $\alpha_1, \dots, \alpha_n$. Furthermore, since the parameters are not orthogonal [see 55], the bias in $\hat{\alpha}_i(\boldsymbol{\beta})$ spreads to the estimator of the slope parameters, $\hat{\boldsymbol{\beta}}$.

The estimation bias in $\hat{\alpha}_i(\boldsymbol{\beta})$ disappears only if $T \rightarrow \infty$. Therefore, the ML estimator $\hat{\boldsymbol{\beta}}$ is not consistent when T is fixed and only $n \rightarrow \infty$, meaning that, in general, $\operatorname{plim}_{n \rightarrow \infty} \hat{\boldsymbol{\beta}} \equiv \boldsymbol{\beta}_* \neq \boldsymbol{\beta}_0$, where $\boldsymbol{\beta}_0$ is the true parameter value.

The problem mentioned above is clarified through an intuitive example provided by Lancaster [55] on continuous data. Consider a random variable z_{it} following a

Gaussian distribution such that $z_{it} \sim N(\delta_i, \sigma_0^2)$. The ML estimator for the mean parameters is given by $\hat{\delta}_i = T^{-1} \sum_{t=1}^T z_{it}$ and for the variance parameter is $\hat{\sigma}^2 = (nT)^{-1} \sum_{i=1}^n \sum_{t=1}^T (z_{it} - \hat{\delta}_i)^2$, which is inconsistent for $n \rightarrow \infty$ and fixed T since it converges to $(T - 1)\sigma_0^2/T$. This is due to the limited number of observations over T for each $\hat{\delta}_i$.

In order to characterize the bias of the ML estimator due to the incidental parameter problem for binary choice models, it is useful to recall the description provided by Arellano and Hahn [9]. Under suitable regularity conditions [see 47, Section 7], we have that

$$\beta_* = \beta_0 + \frac{\mathbf{B}}{T} + O\left(\frac{1}{T^2}\right), \tag{9.7}$$

where the leading term of the bias, \mathbf{B} , is of order T^{-1} and the remainder is of order T^{-2} . Moreover, if $n, T \rightarrow \infty$, $\hat{\beta}$, centered on its probability limit, is asymptotically normal, namely $\sqrt{nT}(\hat{\beta} - \beta_*) \xrightarrow{d} N(\mathbf{0}, \mathbf{\Omega})$. Under these general conditions, the ML estimator is asymptotically biased even if T grows at the same rate as n . In particular, for $n/T \rightarrow \rho$, with $\rho \in (0, +\infty)$, we have

$$\sqrt{nT}(\hat{\beta} - \beta_0) = \sqrt{nT}(\hat{\beta} - \beta_*) + \sqrt{nT} \frac{\mathbf{B}}{T} + O\left(\sqrt{\frac{n}{T^3}}\right) \xrightarrow{d} N(\mathbf{B}\sqrt{\rho}, \mathbf{\Omega}).$$

The bias also arises in the expected score function of the concentrated log-likelihood, described below. The following results are provided by Arellano and Hahn [9]. Let us denote with $\ell_i(\hat{\alpha}_i(\beta), \beta)$ the concentrated log-likelihood for subject i , and the individual score function for β , $\mathbf{s}_{\beta i}(\hat{\alpha}_i(\beta), \beta)$, as

$$\begin{aligned} \ell_i(\hat{\alpha}_i(\beta), \beta) &= \sum_{t=1}^T \log p(y_{it} | \mathbf{x}_{it}; \hat{\alpha}_i(\beta), \beta), \\ \mathbf{s}_{\beta i}(\hat{\alpha}_i(\beta), \beta) &= \frac{\partial \ell_i(\hat{\alpha}_i(\beta), \beta)}{\partial \beta}. \end{aligned}$$

The bias of the ML estimator of β arises because the expected score function, evaluated at $\beta = \beta_0$ with $n \rightarrow \infty$ and T fixed, does not converge in probability to $\mathbf{0}$ as $\hat{\alpha}_i(\beta_0)$ does not converge to the true value α_{i0} . The bias of the expected score function is:

$$E\left[\frac{1}{T} \mathbf{s}_{\beta i}(\hat{\alpha}_i(\beta_0), \beta_0)\right] = \frac{\mathbf{S}_i(\beta_0)}{T} + o\left(\frac{1}{T}\right), \tag{9.8}$$

where $\mathbf{S}_i(\beta_0)/T$ is the component of order T^{-1} and the remainder is of order smaller than T^{-1} .

Finally, it is possible to consider the bias of the profile likelihood. Consider now the infeasible concentrated log-likelihood function for subject i , which is given by $\ell_i(\bar{\alpha}_i(\boldsymbol{\beta}), \boldsymbol{\beta})$, where $\bar{\alpha}_i(\boldsymbol{\beta})$ is the ML estimate of α_i when $T \rightarrow \infty$, so that $\bar{\alpha}_i(\boldsymbol{\beta}_0)$ may be substituted with α_{i0} . The bias in the expected concentrated likelihood can be characterized as

$$\mathbb{E} \left[\frac{1}{T} \ell_i(\hat{\alpha}_i(\boldsymbol{\beta}), \boldsymbol{\beta}) - \frac{1}{T} \ell_i(\bar{\alpha}_i(\boldsymbol{\beta}), \boldsymbol{\beta}) \right] = \frac{L_i(\boldsymbol{\beta})}{T} + o\left(\frac{1}{T}\right),$$

so that this difference tends to disappear as $T \rightarrow \infty$ because it contains a component of order T^{-1} given by $L_i(\boldsymbol{\beta})/T$.

Two different streams of literature have been developed in order to overcome the incidental parameter problem, giving rise to target corrections and conditional inference. Within the first approach, the literature has investigated how to reduce the leading bias component of the ML estimator by correcting either the estimator [38, 40, 47], the score [29], or the concentrated likelihood [19]. The second approach is based on conditioning the response probabilities on sufficient statistics for the incidental parameters. Main contributions in this field are those by Chamberlain [30] for the static version of the model and by Chamberlain [31], Honoré and Kyriazidou [51], and Bartolucci and Nigro [14, 15] for the dynamic setup.

9.5 Target-Corrected Estimators

Target corrections are classified by Arellano and Hahn [9] in three main categories: bias-corrected estimators, correction of the moment equation, and corrected objective-function estimators. The general idea underlying this approach is to mitigate the bias of the ML estimator, reducing its order from T^{-1} to T^{-2} . The advantage of this approach is given by its wide applicability. Indeed, the results are easy to adapt to binary choice models, both static and dynamic, regardless of the functional form assumed for the error term.

In this section, we consider recent contributions and provide a unified framework that embeds them in the classification by Arellano and Hahn [9]. For this reason, we avoid technical details, for which we refer the reader to [6] and [9]. In particular, we first discuss analytic and jackknife corrections of the ML estimator, for which the main contributions have been provided by Hahn and Newey [47], Fernández-Val [40], Hahn and Kuersteiner [46], and Dhaene and Jochmans [38]. We then move to the correction of the first order conditions of a modified likelihood function as proposed by Carro [29], and finally we deal with modifications of the (profile) likelihood function, discussing [10, 23], and [19].

9.5.1 Bias Correction of the ML Estimator

Bias corrections of the ML estimator of β can be obtained analytically or by applying a jackknife method. The former can be achieved by deriving \mathbf{B} in Eq. (9.7) and using the sample counterpart, so as to obtain

$$\hat{\beta}_{BC} = \hat{\beta} - \frac{\hat{\mathbf{B}}}{T}, \tag{9.9}$$

where $\hat{\mathbf{B}} = \hat{\mathbf{B}}(\hat{\beta})$. The analytical expression for $\hat{\mathbf{B}}$ can be derived by an asymptotic expansion of the ML estimator not reported here, and for which we refer the reader to the original works. In particular, [47] derive the analytical bias correction for a general static formulation of nonlinear panel data models. Hahn and Kuersteiner [46] extend this result to a general dynamic formulation for nonlinear panel models and derive regularity conditions for the estimator, under the assumptions that covariates are independent across subjects and are also stationary, meaning that their probability distribution does not change when shifted in time. Finally, [40] derives an analytical bias formula for binary static and dynamic choice models with predetermined regressors.

Hahn and Newey [47] argue that a relevant part of the incidental parameters bias is removed so that the asymptotic distribution of $\hat{\beta}_{BC}$ is centered on β_0 as long as T grows faster than $n^{1/3}$. In fact, if $\sqrt{nT}(\hat{\mathbf{B}} - \mathbf{B})/T \xrightarrow{P} \mathbf{0}$, the confidence intervals of the bias-corrected estimator will be centered on the true value of the parameter β_0 only if $T/n^{1/3} \rightarrow \infty$, since it can be proved that

$$\sqrt{nT}(\hat{\beta}_{BC} - \beta_0) = \sqrt{nT}(\hat{\beta} - \beta_*) - \sqrt{\frac{n}{T}}(\hat{\mathbf{B}} - \mathbf{B}) + O\left(\sqrt{\frac{n}{T^3}}\right) \xrightarrow{d} N(\mathbf{0}, \mathbf{\Omega}).$$

Because $\hat{\beta}$ is used to compute $\hat{\mathbf{B}}$, the bias of the ML estimator could spread to $\hat{\mathbf{B}}$, especially when T is small. Therefore [47] propose iterating the procedure in Eq. (9.9) by updating the estimation of $\hat{\mathbf{B}}$ at the s -th step so that $\hat{\beta}_{BC}^{(s)} = \hat{\beta} - \hat{\mathbf{B}}(\hat{\beta}_{BC}^{(s-1)})$. The resulting estimator $\hat{\beta}_{BC}^{(\infty)}$ is also shown to have better finite sample properties.

An alternative way to perform the bias correction is to rely on the panel jackknife estimator. This is proposed by Hahn and Newey [47] and based on the general jackknife formula provided by Quenouille [63]. The bias correction is formed by using the variation in the ML estimators obtained using samples where each time occasion is dropped sequentially. Specifically, we have

$$\hat{\beta}_{JK} = T\hat{\beta} - \frac{T-1}{T} \sum_{t=1}^T \hat{\beta}^{(t)},$$

where $\hat{\beta}(t)$ is the ML estimator computed after subtracting the t -th observation from the sample. Hahn and Newey [47] show that the order of the bias of $\hat{\beta}_{JK}$ is reduced to T^{-2} .

The previous version of the panel jackknife estimator cannot be directly extended to handle dynamic formulations. Dhaene and Jochmans [38] provide two different estimators based on the split-panel jackknife (SPJ) for dynamic nonlinear models. They consider sub-panels consisting of a reduced number of consecutive observations for each subject in order to preserve the dynamic structure of the data.

Consider a subset of consecutive observations $\mathcal{S} \subset \{1, \dots, T\}$ such that $|\mathcal{S}| \geq T_{min}$, where $|\mathcal{S}|$ denotes the cardinality of the subset \mathcal{S} and T_{min} is the least number of observations for which the ML estimator exists. Let us denote with $\theta = (\beta', \gamma)'$ the vector of slope and state dependence parameters of the dynamic model (9.2) and $\ell(\theta)$ the corresponding profile likelihood where nuisance parameters $\alpha_1, \dots, \alpha_n$ have been concentrated out. Dhaene and Jochmans [38] show that a consistent estimator of the leading component of the bias of the ML estimator is

$$\frac{|\mathcal{S}|}{T - |\mathcal{S}|} (\hat{\theta}_{\mathcal{S}} - \hat{\theta}),$$

where $\hat{\theta}_{\mathcal{S}}$ is the ML estimator of θ based on the subset of observations considered in \mathcal{S} and $\hat{\theta}$ is the ML estimator based on the full data set.

This split-panel jackknife estimator described above could suffer from the arbitrary choice of the sub-panel \mathcal{S} . Define an integer number $G \geq 2$ such that $T \geq G \cdot T_{min}$ and suppose to split the panel into a collection of G non-overlapping sub-panels $\{\mathcal{S}_1, \dots, \mathcal{S}_G\}$.

A consistent estimator for the bias can be obtained by averaging the estimators $\hat{\theta}_{\mathcal{S}_g}$ over the subsets \mathcal{S}_g , for $g = 1, \dots, G$, so that the estimate of the leading bias component becomes

$$\frac{1}{G - 1} (\bar{\theta}_{\mathcal{S}} - \hat{\theta}), \quad \bar{\theta}_{\mathcal{S}} = \sum_{g=1}^G \frac{|\mathcal{S}_g|}{T} \hat{\theta}_{\mathcal{S}_g}.$$

Then the bias corrected estimator, $\hat{\theta}_{JK}$, is derived by subtracting from the ML estimator the estimated bias as follows:

$$\hat{\theta}_{SPJ} = \frac{G}{G - 1} \hat{\theta} - \frac{1}{G - 1} \bar{\theta}_{\mathcal{S}}. \tag{9.10}$$

In cases where splitting the sample in G sub-panels would lead to an arbitrary choice of the partitions, the authors propose to average the estimator over many sets of sub-panels characterized by different cardinalities [see 38, for details].

A peculiarity of this approach is that it allows for generated regressors, which typically arise when handling endogeneity and sample selection. It is also important to mention that consistency of $\hat{\theta}_{SPJ}$ requires strong regularity conditions such

as stationarity of covariates, a sufficient degree of mixing, and independence of observations across subjects.

Dhaene and Jochmans [38] also prove that the split-panel jackknife procedure can be used to correct the profile likelihood function instead of the estimator. A consistent bias-adjusted estimator is then obtained by maximizing the resulting modified function. Finally, an interesting extension of the split-panel jackknife estimator, as well as analytical corrections, to the case of both individual and time incidental parameters is provided by Fernández-Val and Weidner [41].

9.5.2 Bias Correction of the Score and Likelihood Functions

An alternative approach to reduce the bias of the ML estimator is to correct the estimating equation, because the incidental parameter problem affects the first order conditions of the concentrated log-likelihood, as shown in expression (9.8). Bias corrections of the score function have been proposed by Arellano [6], Carro [29], Arellano and Hahn [9], and Fernández-Val [40].

The proposals of [6] and [29] have been most frequently considered in later works as a benchmark. They are based on an earlier work by Cox and Reid [33], whose solution to mitigate the bias arising from the incidental parameter problem is to find a reparametrization such that the nuisance parameters are information orthogonal to the other parameters of interest. Arellano applies this general idea to fixed-effects static binary choice models and expresses the modification in terms of the model’s original parameters. Carro [29] extends [6]’s work to dynamic binary choice models and shows that the order of bias of the ML estimator is reduced from $O(T^{-1})$ to $O(T^{-2})$. Carro’s proposal relies on the solution of the modified estimating equation in the following expression

$$\begin{aligned}
 \mathbf{M}s_{\theta,i}(\boldsymbol{\theta}) &= s_{\theta i}(\hat{\alpha}_i(\boldsymbol{\theta}), \boldsymbol{\theta}) - \frac{1}{2} \frac{1}{s_{\alpha\alpha_i}(\hat{\alpha}_i(\boldsymbol{\theta}), \boldsymbol{\theta})} \left[s_{\theta\alpha\alpha_i}(\hat{\alpha}_i(\boldsymbol{\theta}), \boldsymbol{\theta}) \right. \\
 &\quad \left. + s_{\alpha\alpha\alpha_i}(\hat{\alpha}_i(\boldsymbol{\theta}), \boldsymbol{\theta}) \frac{\partial \hat{\alpha}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right] \\
 &+ \frac{\partial}{\partial \alpha_i} \left\{ \frac{1}{E[s_{\alpha\alpha_i}(\hat{\alpha}_i(\boldsymbol{\theta}), \boldsymbol{\theta})]} E[s_{\theta\alpha_i}(\hat{\alpha}_i(\boldsymbol{\theta}), \boldsymbol{\theta})] \right\} = \mathbf{0}, \quad (9.11)
 \end{aligned}$$

where the subscripts in the score function $s(\cdot)$ denote the derivatives of the profile likelihood w.r.t. a parameter (e.g., $s_{\alpha\alpha_i} = \partial^2 \ell(\boldsymbol{\theta}, \alpha_i(\boldsymbol{\theta})) / \partial \alpha_i^2$). The estimator of $\boldsymbol{\theta}$ that satisfies Eq. (9.11) exhibits a bias of $O(T^{-2})$ and shares the same asymptotic properties of the ML estimator.

Along with the modification of first order conditions, there is a class of approaches dealing with bias-corrected estimators of the objective functions. Bester and Hansen [23] make use of a penalty function for the unconstrained likelihood

function, differently from other approaches exploiting the profile likelihood. Let us define the penalized objective function by

$$Q(\alpha_1, \dots, \alpha_n, \theta) = \sum_{i=1}^n \ell_i(\alpha_i, \theta) - \pi_i(\alpha_i, \theta),$$

whose maximand is the bias-adjusted estimator. A crucial role is played by the penalty function $\pi_i(\alpha_i, \theta)$, defined as

$$\pi_i(\alpha_i, \theta) = \frac{1}{2} \left[\text{trace} \left(-\hat{\mathbf{I}}_{\alpha_i}^{-1} \hat{\mathbf{V}}_{\alpha_i} \right) - 1 \right],$$

which can easily be extended to accommodate multiple fixed effects. The terms $\hat{\mathbf{I}}_{\alpha_i}$ and $\hat{\mathbf{V}}_{\alpha_i}$ are the information for the parameter α_i and a heteroskedasticity- and autocorrelation-robust estimator for the variance of the expected score, respectively. Formally, we have

$$\hat{\mathbf{I}}_{\alpha_i} = \frac{1}{T} s_{\alpha\alpha_i}(\alpha_i, \theta),$$

$$\hat{\mathbf{V}}_{\alpha_i} = \frac{1}{T} \sum_{l=-m}^m \sum_{t=\max(1, 1+l)}^{\min(T, T+l)} s_{\alpha_i, t}(\alpha_i, \theta) s_{\alpha_i, t-l}(\alpha_i, \theta)',$$

where m is a bandwidth parameter and the additional subscript in the score function indicates the t -th observation-specific contribution to the individual score. The main advantages of this approach are its wide applicability for static and dynamic models and the computational easiness, since it only requires the calculation of the score function and the Hessian matrix. [Bester and Hansen \[23\]](#) argue about the asymptotic equivalence of their approach with those previously discussed, highlighting the trade-off between the generality of their proposal and the better finite sample properties of other model-specific bias-corrected estimators like those of [\[40\]](#) and [\[29\]](#).

[Arellano and Hahn \[10\]](#) propose two corrections for the profile likelihood. The first one is called “trace-based” correction. It is not restricted to the likelihood setting and is extremely close to the methodology proposed by [Bester and Hansen \[23\]](#). Their second proposal is the “determinant-based” correction that exploits the log determinants of $\hat{\mathbf{I}}_{\alpha_i}$ and $\hat{\mathbf{V}}_{\alpha_i}$.

A similar approach has been put forward by [Bartolucci et al. \[19\]](#) where the authors propose to derive the bias-reduced estimator by maximizing an adjusted profile likelihood function. In this work, the objective function is the Modified Profile Likelihood (MPL), the logarithm of which is given by

$$\ell_{M,i}(\hat{\alpha}_i(\theta), \theta) = \ell_i(\hat{\alpha}_i(\theta), \theta) + M_i(\hat{\alpha}_i(\theta), \theta),$$

where $M_i(\cdot)$ denotes the adjustment function of the profile log-likelihood. In particular, the correction exploits the modification proposed by Severini [66] which is given by

$$M_i(\hat{\alpha}_i(\boldsymbol{\theta}), \boldsymbol{\theta}) = \frac{1}{2} \log | -s_{\alpha_i \alpha_i}(\hat{\alpha}_i(\boldsymbol{\theta}), \boldsymbol{\theta}) | - \log | \mathbf{I}_{\alpha_i \alpha_i}(\hat{\alpha}_i, \hat{\boldsymbol{\theta}}; \hat{\alpha}_i(\boldsymbol{\theta}), \boldsymbol{\theta}) | ,$$

where $s_{\alpha_i \alpha_i}$ denotes the second derivative of the concentrated log-likelihood w.r.t. the parameter α_i and

$$\mathbf{I}_{\alpha_i \alpha_i}(\hat{\alpha}_i, \hat{\boldsymbol{\theta}}; \hat{\alpha}_i(\boldsymbol{\theta}), \boldsymbol{\theta}) = E[s_{\alpha_i}(\hat{\alpha}_i, \hat{\boldsymbol{\theta}})s_{\alpha_i}(\hat{\alpha}_i(\boldsymbol{\theta}), \boldsymbol{\theta})] ,$$

where, by analogous notation, s_{α_i} denotes the $\partial \ell_i(\hat{\alpha}_i(\boldsymbol{\theta}), \boldsymbol{\theta})/\partial \alpha_i$.

The above framework is general and, specifically for the dynamic binary choice model, the first term of the adjustment can be derived analytically as

$$-s_{\alpha_i \alpha_i}(\boldsymbol{\theta}, \hat{\alpha}_i(\boldsymbol{\theta})) = \sum_t \left\{ \frac{f(\tilde{\mu}_{it})^2}{F(\tilde{\mu}_{it})[1 - F(\tilde{\mu}_{it})]} - C(\tilde{\mu}_{it}) \right\} ,$$

where

$$C(\tilde{\mu}_{it}) = [y_{it} - F(\tilde{\mu}_{it})] \left\{ \frac{f(\tilde{\mu}_{it})}{F(\tilde{\mu}_{it})[1 - F(\tilde{\mu}_{it})]} - \frac{f(\tilde{\mu}_{it})^2 [1 - 2F(\tilde{\mu}_{it})]}{F(\tilde{\mu}_{it})^2 [1 - F(\tilde{\mu}_{it})]^2} \right\} .$$

In this formulation, $f(\cdot)$ denotes the density derived from the distribution function $F(\cdot)$ and $\tilde{\mu}_{it} = \hat{\alpha}_i(\boldsymbol{\theta}) + \mathbf{x}'_{it}\boldsymbol{\beta} + y_{i,t-1}\gamma$ is the linear predictor we obtain considering $\alpha_i = \hat{\alpha}_i(\boldsymbol{\theta})$.

Unfortunately, for the term $\mathbf{I}_{\alpha_i \alpha_i}(\hat{\alpha}_i, \hat{\boldsymbol{\theta}}; \hat{\alpha}_i(\boldsymbol{\theta}), \boldsymbol{\theta})$ we do not have a closed-form expression. Bartolucci et al. [19] propose two different ways to obtain it. The first exploits all the possible configurations of the vector $(y_{i1}, \dots, y_{iT})'$, weighting the product of the scores over the probability assigned to each vector configuration. However, this technique is convenient only when the time dimension of the data set is moderate. The second proposal consists in using a Monte Carlo approximation relying on simulations from the model and has a moderate computational cost.

9.6 Conditional Inference

The conditional inference approach is based on conditioning the joint probability of the response configuration on sufficient statistics for the individual effects. Doing so eliminates the unobserved heterogeneity and thus overcomes the incidental parameter problem. For binary choice models, consider the joint probability for the individual outcome configuration $\mathbf{y}_i = (y_{i1}, \dots, y_{iT})'$, denoted by $p(\mathbf{y}_i|\alpha_i, \mathbf{X}_i)$,

where $\mathbf{X}_i = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT})$ is the matrix collecting the related set of covariates. Consider now a statistic h_i with probability distribution $p(h_i|\alpha_i, \mathbf{X}_i)$. If conditioning the joint probability of \mathbf{y}_i on h_i leads to a conditional probability that is independent of α_i , then h_i is said to be a sufficient statistic for the incidental parameters:

$$p(\mathbf{y}_i|\mathbf{X}_i, h_i) = \frac{p(\mathbf{y}_i, h_i|\alpha_i, \mathbf{X}_i)}{p(h_i|\alpha_i, \mathbf{X}_i)}. \tag{9.12}$$

Andersen [3] shows that the maximand of the log-likelihood function based on the conditional probability in Eq. (9.12) is a consistent estimator for the parameters of interest. Although this idea looks simple and intuitive, it may happen that a sufficient statistic for α_i does not exist or it is not trivial to identify for general binary choice models [52].

A specification admitting a sufficient statistic is the logit model [28, 52]. The probability function for the logit model can be written as

$$p(y_{it}|\alpha_i, \mathbf{x}_{it}) = \frac{\exp[y_{it}(\alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta})]}{1 + \exp(\alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta})},$$

and the joint probability for \mathbf{y}_i is

$$p(\mathbf{y}_i|\alpha_i, \mathbf{X}_i) = \prod_{t=1}^T p(y_{it}|\alpha_i, \mathbf{x}_{it}) = \frac{\exp\left[\alpha_i y_{i+} + \left(\sum_{t=1}^T y_{it} \mathbf{x}_{it}\right)' \boldsymbol{\beta}\right]}{\prod_{t=1}^T [1 + \exp(\alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta})]}, \tag{9.13}$$

where the sum (or total) score $y_{i+} = \sum_{t=1}^T y_{it}$ is the sufficient statistic for the incidental parameter α_i [30]. This can be proven in a simple way, following [28]. The conditional probability of the configuration \mathbf{y}_i given α_i , \mathbf{X}_i , and y_{i+} can be expressed as

$$p(\mathbf{y}_i|\alpha_i, \mathbf{X}_i, y_{i+}) = \frac{p(y_{i+}|\alpha_i, \mathbf{X}_i, \mathbf{y}_i)p(\mathbf{y}_i|\alpha_i, \mathbf{X}_i)}{p(y_{i+}|\alpha_i, \mathbf{X}_i)}.$$

Since the sum score y_{i+} is the sum of the elements in \mathbf{y}_i , then $p(y_{i+}|\alpha_i, \mathbf{X}_i, \mathbf{y}_i) = 1$ by definition. Therefore, it is possible to write

$$p(\mathbf{y}_i|\alpha_i, \mathbf{X}_i, y_{i+}) = \frac{p(\mathbf{y}_i|\alpha_i, \mathbf{X}_i)}{p(y_{i+}|\alpha_i, \mathbf{X}_i)}. \tag{9.14}$$

The numerator in Eq. (9.14) is given by Eq. (9.13), whereas the denominator is given by the sum of the probabilities of observing each possible vector configuration of binary responses $\mathbf{z} = (z_1, \dots, z_T)'$ such that $z_+ = y_{i+}$, where $z_+ = \sum_{t=1}^T z_t$,

obtaining

$$p(y_{i+}|\alpha_i, \mathbf{X}_i) = \frac{\sum_{\mathbf{z}:z_+=y_{i+}} \exp(\alpha_i z_+) \exp \left[\left(\sum_{t=1}^T z_t \mathbf{x}_{it} \right)' \boldsymbol{\beta} \right]}{\prod_{t=1}^T [1 + \exp(\alpha_i + \mathbf{x}'_{it} \boldsymbol{\beta})]} .$$

Finally, it is possible to compute the conditional probability $p(\mathbf{y}_i|\mathbf{X}_i, y_{i+})$, independently of the parameter α_i , as follows

$$\begin{aligned} p(\mathbf{y}_i|\alpha_i, \mathbf{X}_i, y_{i+}) &= \frac{p(\mathbf{y}_i|\alpha_i, \mathbf{X}_i)}{p(y_{i+}|\alpha_i, \mathbf{X}_i)} = \\ &= \frac{\exp \left[\left(\sum_{t=1}^T y_{it} \mathbf{x}_{it} \right)' \boldsymbol{\beta} \right]}{\sum_{\mathbf{z}:z_+=y_{i+}} \exp \left[\left(\sum_{t=1}^T z_t \mathbf{x}_{it} \right)' \boldsymbol{\beta} \right]} = p(\mathbf{y}_i|\mathbf{X}_i, y_{i+}) . \end{aligned} \quad (9.15)$$

Equation (9.15) defines the conditional logit model shown in [58] and [30]. Then, we express the conditional log-likelihood function, as the sum of the logarithm of the individual conditional probabilities:

$$\ell(\boldsymbol{\beta}) = \sum_i \mathbb{1}\{0 < y_{i+} < T\} \log p(\mathbf{y}_i|\mathbf{X}_i, y_{i+}) . \quad (9.16)$$

Note that we exclude individuals characterized by a sum score of 0 or T , because their conditional log-probability is equal to 0 by construction. The function in Eq. (9.16) can be maximized with respect to $\boldsymbol{\beta}$ by the Newton-Raphson algorithm, obtaining the Conditional ML (CML) estimator $\hat{\boldsymbol{\beta}}_{CML}$.

Differently from the static case, conditional inference for dynamic models is more difficult because a useful sufficient statistic for α_i is not always available. Different approaches have been implemented in order to overcome this problem. Consider the model defined by Eq. (9.2) where ε_{it} is logistically distributed, corresponding to the Dynamic Logit (DL) model [see 52, Chapter 7], implying that

$$p(y_{it}|\alpha_i, \mathbf{x}_{it}, y_{i0}, \dots, y_{i,t-1}) = \frac{\exp [y_{it}(\alpha_i + \mathbf{x}'_{it} \boldsymbol{\beta} + y_{i,t-1}\gamma)]}{1 + \exp(\alpha_i + \mathbf{x}'_{it} \boldsymbol{\beta} + y_{i,t-1}\gamma)} , \quad (9.17)$$

and y_{i0} is assumed to be known. In this case, the probability for the response configuration \mathbf{y}_i is

$$p(\mathbf{y}_i|\alpha_i, \mathbf{X}_i, y_{i0}) = \frac{\exp \left[y_{i+}\alpha_i + \left(\sum_{t=1}^T y_{it} \mathbf{x}_{it} \right)' \boldsymbol{\beta} + y_{i+}\gamma \right]}{\prod_{t=1}^T [1 + \exp(\alpha_i + \mathbf{x}'_{it} \boldsymbol{\beta} + y_{i,t-1}\gamma)]} , \quad (9.18)$$

where $y_{i*} = \sum_{t=1}^T y_{i,t-1}y_{i,t}$. It can be proven that the sum score y_{i+} is no longer a sufficient statistic for the incidental parameters as in the static specification [31].

A first feasible, even though quite restrictive, solution for the DL model is given by Chamberlain [31]. Consider the model given in Eq. (9.17) where the exogenous variables \mathbf{x}_{it} are excluded and $T \geq 3$. Under this setup, consider the example with $T = 3$, where the probability for the observations (y_{i0}, \dots, y_{i3}) is independent of α_i conditional on $y_{i1} + y_{i2} = 1$. By maximizing the resulting conditional log-likelihood function

$$\sum_{i=1}^n \mathbb{1}\{y_{i1} + y_{i2} = 1\} (y_{i1}[(y_{i0} - y_{i3})\gamma] - \log\{1 + \exp[(y_{i0} - y_{i3})\gamma]\}) ,$$

it is possible to obtain a \sqrt{n} consistent estimator of γ . For additional details, we refer the reader to the proof in Section 7.5.3 of [52].

Chamberlain [31]’s approach has been extended by Honoré and Kyriazidou [51] allowing for the presence of exogenous explanatory variables. The authors show how to identify and estimate β and γ regardless of the parameters α_i . Given $T = 3$, this can be done by maximizing a weighted conditional log-likelihood function given by

$$\sum_{i=1}^n \mathbb{1}\{y_{i1} + y_{i2} = 1\} K\left(\frac{\mathbf{x}_{i2} - \mathbf{x}_{i3}}{\sigma_n}\right) \log p(\mathbf{y}_i | \alpha_i, \mathbf{X}_i, y_{i0}, y_{i1} + y_{i2} = 1, y_{i3}) ,$$

where $K(\cdot)$ is a kernel density function, to be carefully chosen, used in order to weigh observations. In particular, weights are inversely proportional to the magnitude of the difference $(\mathbf{x}_{i2} - \mathbf{x}_{i3})$, σ_n is a fixed bandwidth that depends on n , and

$$p(\mathbf{y}_i | \alpha_i, \mathbf{X}_i, y_{i0}, y_{i1} + y_{i2} = 1, y_{i3}) = \frac{\exp\{y_{i1}[(\mathbf{x}_{i1} - \mathbf{x}_{i2})'\beta + (y_{i0} - y_{i3})\gamma]\}}{1 + \exp\{(\mathbf{x}_{i1} - \mathbf{x}_{i2})'\beta + (y_{i0} - y_{i3})\gamma\}} .$$

Although the proposed estimator is proven to be consistent and asymptotically normal, it shows some drawbacks. The convergence rate, due to the presence of the kernel density function, is slower than \sqrt{n} . Moreover, the conditions exploited for identification, namely that $y_{i1} + y_{i2} = 1$, and the weight given by the kernel, limit the number of statistic units that actually contribute to the likelihood, affecting the efficiency of the estimator. These conditions are also tightened by the presence of discrete covariates which are required, at individual level, to assume the same value in periods 2 and 3 [see 51, Section 2]. Moreover, the condition imposed on the covariates rules out time dummies. [51] also provide identification for $T \geq 3$ and more than one lag of the dependent variable.

As shown above, conditional inference for the DL model leads to restrictive conditions on the covariates for the identification. In order to overcome this shortcoming, [14] proposed an approximation based on the Quadratic Exponential

(QE) model, which derives from the multivariate binary data distribution introduced by Cox [32]. A similar approach was proposed by Bartolucci and Pennoni [16] for the two-parameter logistic model.

The QE model is particularly useful as it closely resembles the DL model. In general, the QE model describes the joint distribution of binary variables, which are here represented by the responses in the T time occasions. In the QE models, probabilities depend on both the main effects, which are the parameters associated with the regressors in the panel data case, and the so-called bivariate interaction effects [34], where interactions are between two of the binary responses. In our case, the parameter associated with the interaction term captures the true state dependence. We refer the reader to [32] and [14] for a detailed derivation of the model and its specification to account for subject-specific unobserved heterogeneity, which is beyond the scope of this work.

The QE directly defines the conditional probability for the vector \mathbf{y}_i as

$$p(\mathbf{y}_i | \delta_i, \mathbf{X}_i, y_{i0}) = \frac{\exp \left[y_{i+} \delta_i + \left(\sum_{t=1}^T y_{it} \mathbf{x}_{it} \right)' \boldsymbol{\eta}_1 + y_{iT} (\phi + \mathbf{x}'_{iT} \boldsymbol{\eta}_2) + y_{i*} \psi \right]}{\sum_{\mathbf{z}} \exp \left[z_+ \delta_i + \left(\sum_{t=1}^T z_t \mathbf{x}_{it} \right)' \boldsymbol{\eta}_1 + z_T (\phi + \mathbf{x}'_{iT} \boldsymbol{\eta}_2) + z_{i*} \psi \right]}, \tag{9.19}$$

where the notation for the parameters is different in order to distinguish them from the DL model, so that δ_i is the individual parameter for the unobserved heterogeneity, $\boldsymbol{\eta}_1$ is a vector of parameters related to the set of the strictly exogenous regressors, ϕ and $\boldsymbol{\eta}_2$ are nuisance parameters, and ψ measures the state dependence. The denominator is given by the sum of all possible binary response vectors $\mathbf{z} = (z_1, \dots, z_T)'$, where $z_+ = \sum_{t=1}^T z_t$ and $z_{i*} = y_{i0} z_1 + \sum_{t=2}^T z_{t-1} z_t$.

The QE and the DL models share many properties. First of all, the parameter ψ can be interpreted as the log-odds ratio between pairs of consecutive response variables $(y_{i,t-1}, y_{it})$ for every i and t . The same definition holds for the state dependence parameter γ in the DL model. Moreover, the DL and the QE models coincide when $\gamma = \psi = 0$.

Further to the above similarities, the most important feature of the QE model is that it admits a sufficient statistic for the incidental parameters, namely the sum score y_{i+} . Conditioning the probability in Eq. (9.19) on the sum score leads to

$$p(\mathbf{y}_i | \delta_i, \mathbf{X}_i, y_{i0}, y_{i+}) = \frac{\exp \left[\left(\sum_{t=1}^T y_{it} \mathbf{x}_{it} \right)' \boldsymbol{\eta}_1 + y_{iT} (\phi + \mathbf{x}'_{iT} \boldsymbol{\eta}_2) + y_{i*} \psi \right]}{\sum_{\mathbf{z}: z_+ = y_{i+}} \exp \left[\left(\sum_{t=1}^T z_t \mathbf{x}_{it} \right)' \boldsymbol{\eta}_1 + z_T (\phi + \mathbf{x}'_{iT} \boldsymbol{\eta}_2) + z_{i*} \psi \right]}, \tag{9.20}$$

which does not depend on the incidental parameters δ_i .

Consistent estimators of parameters $(\eta'_1, \eta'_2, \phi, \psi)'$ can be obtained via the maximization of a conditional likelihood function built summing the individual probabilities in Eq. (9.20). Moreover, the estimator has a rate of convergence of \sqrt{n} and is asymptotically normally distributed. The model specification is also more flexible than those provided by previous contributions, since it allows for time dummies and it is valid for $T \geq 2$ beyond the initial observation.

An interesting feature of the QE model is given by the fact that it can be exploited as an approximation in order to estimate the parameter of a DL model, as argued by Bartolucci and Nigro [15], who derive a Pseudo Conditional ML (PCML) estimator. The starting point is the log-probability of the DL in Eq. (9.18) given by

$$\begin{aligned} \log p(\mathbf{y}_i | \alpha_i, \mathbf{X}_i, y_{i0}) = & \\ y_{i+} \alpha_i + \left(\sum_{t=1}^T y_{it} \mathbf{x}_{it} \right)' \boldsymbol{\beta} + y_{i*} \gamma - \sum_{t=1}^T \log [1 + \exp(\alpha_i + \mathbf{x}'_{it} \boldsymbol{\beta} + y_{i,t-1} \gamma)] . & \end{aligned} \tag{9.21}$$

The non-linear component in Eq. (9.21) is approximated by a first-order Taylor's expansion around $\alpha_i = \bar{\alpha}_i$, $\boldsymbol{\beta} = \bar{\boldsymbol{\beta}}$, and $\gamma = 0$ as

$$\begin{aligned} \sum_{t=1}^T \log [1 + \exp(\alpha_i + \mathbf{x}'_{it} \boldsymbol{\beta} + y_{i,t-1} \gamma)] \approx \sum_{t=1}^T \left\{ \log [1 + \exp(\bar{\alpha}_i + \mathbf{x}'_{it} \bar{\boldsymbol{\beta}})] \right. \\ \left. + \bar{q}_{i1} [\alpha_i - \bar{\alpha}_i + \mathbf{x}'_{it} (\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})] \right\} + \bar{q}_{i1} y_{i0} \gamma + \sum_{t>1} \bar{q}_{it} y_{i,t-1} \gamma , \end{aligned}$$

where $\bar{\alpha}_i$ and $\bar{\boldsymbol{\beta}}$ are fixed values for α_i and $\boldsymbol{\beta}$ and

$$\bar{q}_{it} = \frac{\exp(\bar{\alpha}_i + \mathbf{x}'_{it} \bar{\boldsymbol{\beta}})}{1 + \exp(\bar{\alpha}_i + \mathbf{x}'_{it} \bar{\boldsymbol{\beta}})} .$$

The last expression corresponds to a static logit formulation for $p(y_{it} = 1 | \alpha_i, \mathbf{x}_{it})$ at the fixed value of the parameters. Therefore, replacing the non-linear term with its expansion in Eq. (9.21) and restoring the exponential form leads to the approximated probability given by

$$\begin{aligned} p^*(\mathbf{y}_i | \alpha_i, \mathbf{X}_i, y_{i0}) & \\ = \frac{\exp \left[y_{i+} \alpha_i + \left(\sum_{t=1}^T y_{it} \mathbf{x}_{it} \right)' \boldsymbol{\beta} + y_{i*} \gamma - \sum_{t=1}^T \bar{q}_{it} y_{i,t-1} \gamma \right]}{\sum_{\mathbf{z}} \exp \left[z_{+} \alpha_i + \left(\sum_{t=1}^T z_t \mathbf{x}_{it} \right)' \boldsymbol{\beta} + z_{i*} \gamma - \sum_{t=1}^T \bar{q}_{it} z_{i,t-1} \gamma \right]} . & \end{aligned}$$

The last equation corresponds to a modified version of the QE model, which can be exploited for the estimation of the parameters of the DL model. As shown above, the QE model admits the sum score as a sufficient statistic for parameters α_i (denoted by δ_i in the QE parametrization). Hence, by conditioning y_i also on the sufficient statistic y_{i+} , we obtain

$$\begin{aligned}
 & p^*(\mathbf{y}_i | \mathbf{X}_i, y_{i0}, y_{i+}) \\
 &= \frac{\exp \left[\left(\sum_{t=1}^T y_{it} \mathbf{x}_{it} \right)' \boldsymbol{\beta} + y_{i*} \gamma - \sum_{t=1}^T \bar{q}_{it} y_{i,t-1} \gamma \right]}{\sum_{\mathbf{z}: z_+ = y_{i+}} \exp \left[\left(\sum_{t=1}^T z_t \mathbf{x}_{it} \right)' \boldsymbol{\beta} + z_{i*} \gamma - \sum_{t=1}^T \bar{q}_{it} z_{i,t-1} \gamma \right]}, \quad (9.22)
 \end{aligned}$$

which is independent of α_i . Finally, the probability in Eq. (9.22) enters the likelihood function and the estimation procedure involves two steps:

1. A preliminary estimate of $\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\beta}}$, is obtained by CML estimation of the static logit model [30]. The probabilities \bar{q}_{it} are evaluated at $\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}$ and at $\tilde{\alpha}_i$ equal to the ML estimate of α_i under the static logit model, where the concentrated likelihood is a function of $\tilde{\boldsymbol{\beta}}$.
2. The conditional log-likelihood for (9.22), given the preliminary estimates, is maximized w.r.t. the set of parameters $\boldsymbol{\theta}$ and is

$$\ell^*(\boldsymbol{\theta} | \tilde{\boldsymbol{\beta}}) = \sum_{i=1}^n \mathbb{1}\{0 < y_{i+} < T\} \log [P_{\boldsymbol{\theta} | \tilde{\boldsymbol{\beta}}}^*(\mathbf{y}_i | \mathbf{X}_i, y_{i0}, y_{i+})],$$

and the subscript $\boldsymbol{\theta} | \tilde{\boldsymbol{\beta}}$ denotes the dependence on $\tilde{\boldsymbol{\beta}}$ through the probabilities \bar{q}_{it} .

Asymptotic properties of the PCML estimator exhibit some peculiarities discussed in [15]. The proposed estimator is proven to be consistent for the parameters $\boldsymbol{\theta}$ when the true value of the state dependence parameter is $\gamma_0 = 0$. Moreover, the PCML estimator is biased for the DL parameters and its bias is proportional to the magnitude of the state dependence parameter γ . However, simulation results suggest that the PCML estimator provides a good approximation of the DL parameters. Moreover, a modified version of the QE model has been derived by Bartolucci et al. [20] to build a statistical test procedure for the null hypothesis of absence of state dependence $H_0 : \gamma = 0$. The proposed test outperforms the classical *t-test*, based on the basic QE or the PCML estimator, in terms of size and power.

9.7 Simulation Study

This section describes the Monte Carlo study aimed at comparing the finite sample performance of a set of the most recent estimators designed to overcome the

incidental parameter problem. The study is focused on the DL model with fixed effects, for which the conditional inference approach is a viable alternative to bias correction techniques.

9.7.1 Simulation Design

The data are generated from a DL model formulated as

$$y_{i0} = \mathbb{1}\{\alpha_i + x_{i0}\beta + \varepsilon_{i0} > 0\}, \quad (9.23)$$

$$y_{it} = \mathbb{1}\{\alpha_i + x_{it}\beta + y_{i,t-1}\gamma + \varepsilon_{it} > 0\}, \quad t = 1, \dots, T, \quad (9.24)$$

for $i = 1, \dots, n$. The variable x_{it} is an exogenous regressor generated from a Gaussian distribution with zero mean and variance $\pi^2/3$, and ε_{it} is a random variable following a logistic distribution. The parameter β is equal to 1 and γ takes values in $\{0, 0.25, 0.5, 1, 2\}$, in order to evaluate different degrees of state dependence. Individual intercepts α_i are generated as in [51], that is, $\alpha_i = (1/4) \sum_{t=0}^3 x_{it}$. Although fixed-effects estimators do not require any assumption concerning the unobserved heterogeneity, the adopted design for α_i allows intercepts and the covariate to be correlated, which is one of the main advantages of the fixed-effects approach. Finally, the sample sizes and the time lengths considered are $n = 250, 500, 1000$ and $T = 3, 4, 6, 8, 12$. The number of Monte Carlo replications is 1000.

We consider the following estimators based on both the target-adjusted and the conditional approaches: (i) the MPL put forward by [Bartolucci et al. \[19\]](#) and illustrated in Sect. 9.5.2; (ii) the SPJ bias-correction proposed by [Dhaene and Jochmans \[38\]](#) and defined in Eq. (9.10); (iii) the estimator proposed by [Honoré and Kyriazidou \[51\]](#), denoted HK (see Sect. 9.6), where the bandwidth parameter is set to $\sigma_n = 8 \cdot n^{-1/5}$; and (iv) the PCML estimator by [Bartolucci and Nigro \[15\]](#) presented and defined in Eq. (9.22). Moreover, these four approaches are compared with the ML estimator and with the Infeasible ML estimator (INF). The latter is based on including the true values of the individual intercepts in the model as an additional regressor and the slope parameter is estimated by ML. It is therefore not affected by the incidental parameter problem and serves as a benchmark. [\[38\]](#) propose a simulation study with the same design, where they compare a wide set of target-corrected estimators, including those proposed by [Carro \[29\]](#) and [\[40\]](#). For this reason, here we only focus on the comparison between conditional and recent bias reduction techniques, referring the reader to the comprehensive study by [Dhaene and Jochmans \[38\]](#) for a comparison with earlier approaches.

9.7.2 Simulation Results

This section describes the main results of the simulation study. As in the original contribution of [51], the true value of the state dependence parameter is set to $\gamma = 0.5$ for the benchmark design. Tables 9.1 and 9.2 show the statistics of the six estimators considered for the parameters β and γ , respectively.

For each sample size, the mean bias, the median bias, the Root Mean Square Error (RMSE) and the Median Absolute Error (MAE) are reported in the tables, where

$$\text{RMSE} = \sqrt{\frac{\sum_{j=1}^{1000} (\hat{\xi}_j - \xi_0)^2}{1000}}; \quad \text{MAE} = \text{median}_{j=1, \dots, 1000} (|\hat{\xi}_j - \xi_0|),$$

and where ξ_0 denotes the true value of the parameter in turn evaluated and $\hat{\xi}_j$ denotes the value of the related estimators in replication j . The full set of additional results is reported in Tables 9.5, 9.6, 9.7, 9.8, 9.9, 9.10, 9.11 and 9.12 of the Appendix.

First of all, we confirm that the incidental parameter problem severely affects the ML estimator, as can be evinced from the comparison with the INF estimator. As expected, the bias is considerable regardless of the sample size and it decreases only slightly with the time series length. Moreover, the bias appears to be larger for the estimator of the state dependence parameter, $\hat{\gamma}$. Note that, as we could expect, the INF estimator performs well and its finite sample bias is always negligible for the whole set of parameters.

The behavior of the other estimators considered is not homogeneous across designs. As the theory would suggest, the finite sample performance of target-corrected estimators is sensitive to T as the larger the number of time occasions, the smaller the magnitude of the bias. In fact, for a given n , the biases of MPL and SPJ estimators, which can only be computed for $T \geq 6$, shrink as the time series grows, even though the latter requires $T \geq 8$ to produce a sizable bias reduction.

As for the CML approaches, results show that the bias of the HK estimator is small for both parameters. The bias of $\hat{\beta}$ tends to reduce as both n and T increase, while the bias of $\hat{\gamma}$ is stable for each configuration, even though the MAE decreases when the sample size increases. The PCML estimator shows the best finite sample performance across almost all configurations. In fact, mean and median bias are not only the smallest but these quantities are negligible for the estimator of the regression parameter β and the state dependence parameter γ . In addition, the RMSE and MAE decrease as n and T increase.

In order to better understand these phenomena, we report some of the results shown in Tables 9.1 and 9.2 by means of graphical illustrations. Figures 9.1 and 9.2 report, in boxplots, the bias of the six estimators of β and γ , respectively, in four of the scenarios considered, namely $n = 500, 1000$ and $T = 6, 12$. From the two figures, it is evident that an increase in n is not able to shift the distribution of the bias of the ML and target-corrected estimators toward zero, meaning that the magnitude of the bias is not affected by the number of individuals, as theory suggests. On the

Table 9.1 Simulation results under Benchmark design based on Eqs.(9.23) and (9.24) with parameters $\beta = 1, \gamma = 0.5$

Results for $\hat{\beta}$													
n	T	Mean Bias						RMSE					
		INF	ML	MPL	SPJ	HK	PCML	INF	ML	MPL	SPJ	HK	PCML
250	3	0.007	0.817	0.002	–	0.149	0.027	0.087	0.865	0.179	–	0.375	0.145
	4	0.010	0.572	0.013	–	0.057	0.015	0.074	0.600	0.080	–	0.165	0.102
	6	0.008	0.327	0.027	–0.152	0.029	0.009	0.057	0.343	0.070	0.209	0.102	0.069
	8	0.008	0.224	0.021	–0.121	0.017	0.007	0.048	0.237	0.061	0.153	0.071	0.059
	12	0.004	0.134	0.010	–0.057	0.010	0.004	0.041	0.144	0.046	0.077	0.052	0.045
500	3	0.002	0.777	–0.059	–	0.069	0.006	0.064	0.800	0.106	–	0.200	0.095
	4	0.004	0.552	0.007	–	0.031	0.005	0.051	0.566	0.055	–	0.110	0.069
	6	0.001	0.316	0.020	–0.140	0.015	0.002	0.041	0.324	0.050	0.172	0.068	0.048
	8	0.001	0.216	0.015	–0.117	0.010	0.002	0.033	0.222	0.040	0.131	0.049	0.038
	12	0.003	0.131	0.008	–0.057	0.006	0.002	0.026	0.135	0.030	0.067	0.035	0.029
1000	3	0.002	0.764	–0.019	–	0.046	0.002	0.044	0.776	0.095	–	0.138	0.066
	4	0.002	0.546	0.005	–	0.024	0.002	0.035	0.553	0.038	–	0.082	0.048
	6	0.001	0.314	0.019	–0.135	0.011	0.001	0.029	0.319	0.038	0.151	0.050	0.035
	8	0.002	0.215	0.014	–0.112	0.008	0.001	0.024	0.218	0.031	0.120	0.037	0.028
	12	0.000	0.129	0.007	–0.054	0.005	0.001	0.019	0.132	0.022	0.059	0.025	0.020
n	T	Median Bias						MAE					
		INF	ML	MPL	SPJ	HK	PCML	INF	ML	MPL	SPJ	HK	PCML
250	3	0.006	0.782	–0.039	–	0.078	0.009	0.061	0.782	0.072	–	0.168	0.090
	4	0.003	0.560	0.010	–	0.044	0.007	0.047	0.560	0.051	–	0.101	0.065
	6	0.005	0.319	0.023	–0.145	0.021	0.006	0.035	0.319	0.049	0.148	0.062	0.045
	8	0.006	0.217	0.021	–0.119	0.019	0.006	0.037	0.217	0.043	0.120	0.048	0.039
	12	0.006	0.134	0.011	–0.058	0.011	0.005	0.029	0.134	0.031	0.060	0.036	0.030
500	3	0.000	0.759	–0.055	–	0.041	0.001	0.041	0.759	0.063	–	0.115	0.060
	4	0.003	0.545	0.006	–	0.025	0.002	0.034	0.545	0.037	–	0.071	0.045
	6	0.001	0.314	0.020	–0.139	0.013	0.001	0.027	0.314	0.034	0.140	0.044	0.032
	8	–0.000	0.214	0.014	–0.116	0.008	0.000	0.021	0.214	0.027	0.116	0.033	0.026
	12	0.003	0.131	0.009	–0.058	0.004	0.003	0.017	0.131	0.021	0.058	0.023	0.020
1000	3	–0.001	0.750	–0.043	–	0.028	–0.003	0.030	0.750	0.054	–	0.079	0.043
	4	0.003	0.539	0.003	–	0.022	–0.000	0.024	0.539	0.025	–	0.052	0.032
	6	0.001	0.312	0.018	–0.133	0.009	–0.000	0.020	0.312	0.025	0.133	0.032	0.022
	8	0.001	0.214	0.015	–0.112	0.006	0.001	0.017	0.214	0.021	0.112	0.024	0.019
	12	0.000	0.130	0.007	–0.054	0.005	0.001	0.013	0.130	0.014	0.054	0.017	0.014

INF: Infeasible Likelihood Estimator, ML: Maximum Likelihood Estimator, MPL: Modified Profile Likelihood [19], SPJ: Split-panel Jackknife [38], HK: DL estimator [51], PCML: Pseudo Conditional Maximum Likelihood Estimator [15]

contrary, when T moves from 6 to 12, we observe that the bias distribution is not only centered in values closer to 0, but also shrinks denoting a sizable reduction in the variability of the estimators.

Table 9.2 Simulation results under Benchmark design based on Eqs.(9.23) and (9.24) with parameters $\beta = 1, \gamma = 0.5$

Results for $\hat{\gamma}$													
n	T	Mean Bias						RMSE					
		INF	ML	MPL	SPJ	HK	PCML	INF	ML	MPL	SPJ	HK	PCML
250	3	-0.007	-2.617	-0.202	-	-0.053	-0.000	0.149	2.690	0.370	-	0.690	0.438
	4	-0.000	-1.576	-0.208	-	-0.036	0.017	0.126	1.615	0.302	-	0.355	0.280
	6	-0.003	-0.907	-0.110	0.770	-0.058	0.003	0.103	0.933	0.210	0.845	0.232	0.190
	8	-0.001	-0.649	-0.074	0.286	-0.080	-0.012	0.094	0.669	0.165	0.357	0.182	0.150
	12	-0.002	-0.397	-0.028	0.098	-0.057	-0.003	0.071	0.415	0.115	0.162	0.130	0.112
500	3	-0.000	-2.588	-0.229	-	-0.058	0.003	0.107	2.621	0.293	-	0.397	0.288
	4	-0.002	-1.580	-0.214	-	-0.051	0.009	0.089	1.600	0.263	-	0.257	0.199
	6	-0.001	-0.909	-0.114	0.743	-0.050	0.001	0.070	0.921	0.168	0.782	0.166	0.133
	8	-0.000	-0.630	-0.059	0.298	-0.041	0.006	0.058	0.640	0.116	0.330	0.126	0.104
	12	0.004	-0.387	-0.019	0.104	-0.052	0.006	0.049	0.395	0.078	0.140	0.097	0.077
1000	3	0.002	-2.548	-0.233	-	-0.031	0.021	0.075	2.563	0.286	-	0.273	0.197
	4	-0.001	-1.577	-0.217	-	-0.052	0.009	0.065	1.587	0.245	-	0.193	0.142
	6	-0.002	-0.901	-0.109	0.740	-0.043	0.007	0.051	0.907	0.138	0.758	0.113	0.090
	8	0.002	-0.628	-0.058	0.288	-0.038	0.006	0.041	0.634	0.092	0.305	0.093	0.074
	12	-0.001	-0.397	-0.030	0.091	-0.046	-0.004	0.034	0.402	0.064	0.111	0.079	0.058
n	T	Median Bias						MAE					
		INF	ML	MPL	SPJ	HK	PCML	INF	ML	MPL	SPJ	HK	PCML
250	3	-0.010	-2.563	-0.197	-	-0.050	0.015	0.108	2.563	0.233	-	0.394	0.300
	4	-0.000	-1.570	-0.211	-	-0.061	-0.000	0.082	1.570	0.228	-	0.253	0.189
	6	-0.007	-0.921	-0.121	0.760	-0.058	-0.012	0.072	0.921	0.157	0.760	0.154	0.130
	8	-0.001	-0.650	-0.079	0.287	-0.088	-0.006	0.071	0.650	0.102	0.287	0.128	0.109
	12	-0.002	-0.400	-0.030	0.097	-0.054	-0.004	0.048	0.400	0.079	0.112	0.089	0.078
500	3	-0.000	-2.586	-0.225	-	-0.054	0.000	0.073	2.586	0.227	-	0.259	0.199
	4	-0.000	-1.575	-0.215	-	-0.054	0.011	0.058	1.575	0.219	-	0.177	0.131
	6	0.000	-0.906	-0.111	0.732	-0.058	0.005	0.046	0.906	0.118	0.732	0.114	0.097
	8	-0.001	-0.629	-0.059	0.293	-0.041	0.007	0.038	0.629	0.081	0.293	0.087	0.071
	12	0.007	-0.392	-0.028	0.102	-0.060	-0.002	0.028	0.392	0.054	0.104	0.075	0.048
1000	3	0.001	-2.530	-0.227	-	-0.033	0.022	0.050	2.530	0.228	-	0.180	0.130
	4	0.001	-1.571	-0.215	-	-0.051	0.012	0.042	1.571	0.215	-	0.132	0.094
	6	-0.003	-0.899	-0.105	0.734	-0.044	0.008	0.035	0.899	0.108	0.734	0.078	0.062
	8	0.002	-0.630	-0.058	0.286	-0.040	0.005	0.028	0.630	0.066	0.286	0.062	0.051
	12	0.001	-0.400	-0.031	0.088	-0.050	-0.006	0.022	0.400	0.045	0.089	0.058	0.039

See notes in Table 9.1

It is interesting to note that the conditional approach has better finite sample performance than target-adjusted estimators when $T \leq 8$, despite the drawbacks that should be expected from the theoretical results. In particular, the HK estimator is consistent with a rate of convergence that is slower than \sqrt{n} , while the PCML

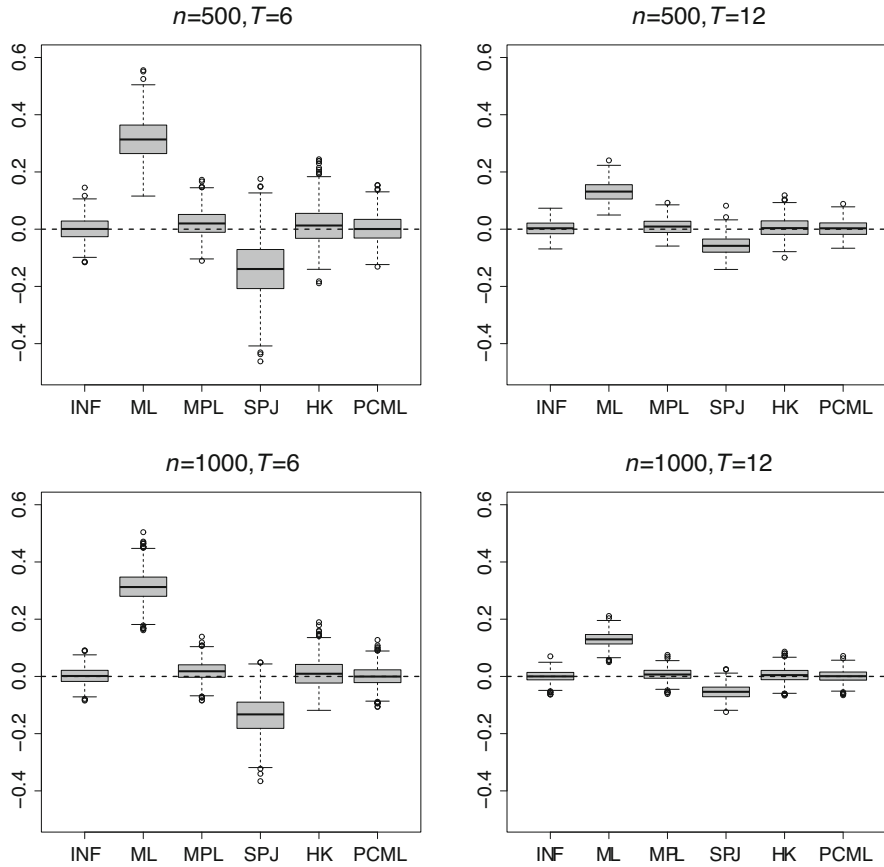


Fig. 9.1 Distribution of the bias: estimators of β . Notes: INF: Infeasible Likelihood Estimator, ML: Maximum Likelihood Estimator, MPL: Modified Profile Likelihood [19], SPJ: Split-panel Jackknife [38], HK: DL estimator [51], PCML: Pseudo Conditional Maximum Likelihood Estimator [15]

estimator is consistent only for $\gamma = 0$. Nevertheless, the second has the smallest bias across all the scenarios with $\gamma = 0.5$.

The finite sample behavior of the PCML estimator is depicted by the simulation results based on the designs with different values of γ . Because the PCML estimator is not consistent with $\gamma \neq 0$, a way to evaluate its performance is to quantify the relative improvement in terms of bias over the ML estimator. This can be achieved by computing the Δ index proposed by Bartolucci et al. [19], that is,

$$\Delta(\cdot) = \frac{|MB(ML)| - |MB(\cdot)|}{|MB(ML)| - |MB(INF)|},$$

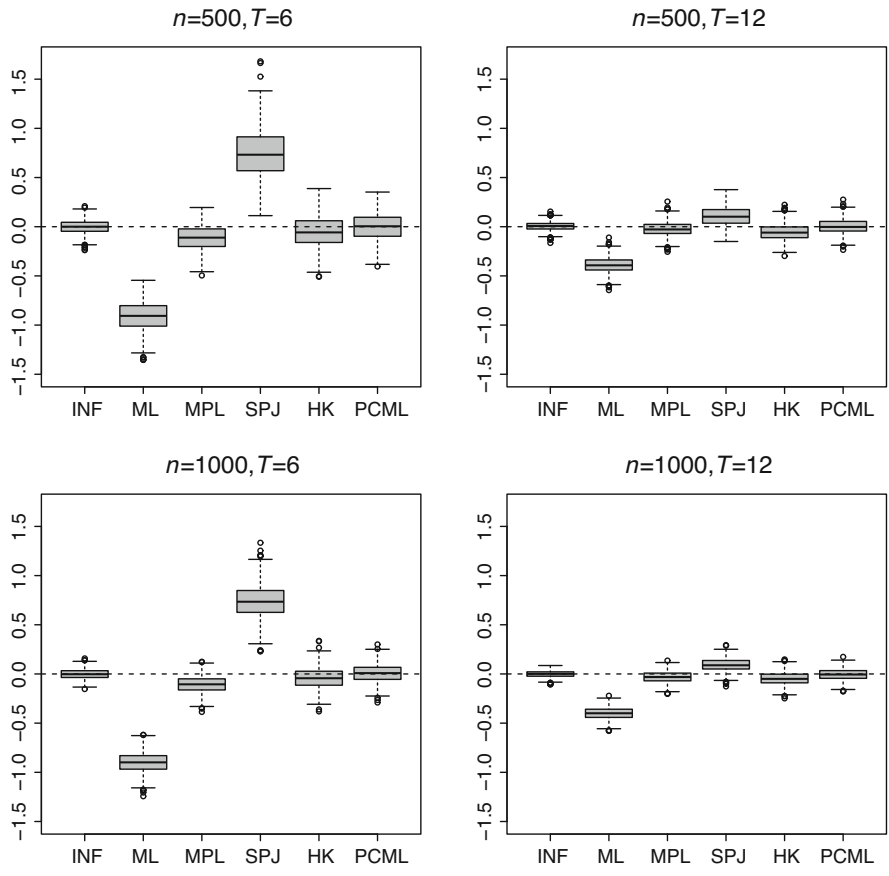


Fig. 9.2 Distribution of the bias: estimators of γ . Note: see Fig. 9.1

where $|MB(\cdot)|$ is the absolute value of the median bias of the estimators considered in the simulation study. This index can be seen as a measure of the relative performance of an estimator with respect to the INF and where the ML estimator represents the benchmark. Table 9.3 reports the results for the estimators considered and for the set of values of γ .

The effect of a variation in the true value of the state dependence parameter is different for $\hat{\beta}$ and $\hat{\gamma}$. In fact, the Δ index is stable for all the four estimators of β as the true value of γ grows. Moreover, the relative bias among the estimators is in line with the previous results. On the contrary, the behavior of $\hat{\gamma}$ varies according to the approach considered. In fact, target-adjusted estimators are more sensitive to the degree of state dependence than CML estimators. Specifically, the performance of the MPL method worsens as γ increases, while the SPJ estimator stably exhibits a larger bias across all values of γ .

Table 9.3 Simulation results under the Benchmark design. Relative Performance (Δ), $n = 500$

γ	T	$\hat{\beta}$				$\hat{\gamma}$			
		MPL	SPJ	HK	PCML	MPL	SPJ	HK	PCML
0	3	0.915	–	0.993	0.994	1.004	–	0.986	1.003
	6	0.959	0.568	1.002	1.005	0.912	0.195	1.001	0.999
	12	0.974	0.621	1.022	1.014	0.996	0.686	1.016	0.992
0.25	3	0.919	–	0.903	0.998	0.972	–	0.970	0.998
	6	0.954	0.504	0.980	1.003	0.893	0.212	0.959	0.999
	12	0.949	0.581	0.965	0.995	0.941	0.746	0.911	1.000
0.5	3	0.928	–	0.947	1.000	0.913	–	0.979	1.000
	6	0.938	0.557	0.961	1.000	0.878	0.192	0.936	0.995
	12	0.953	0.571	0.996	1.001	0.946	0.753	0.862	1.012
1	3	0.954	–	0.890	1.005	0.819	–	0.961	0.985
	6	0.953	0.629	0.957	1.023	0.840	0.219	0.898	1.004
	12	0.952	0.596	0.961	1.002	0.914	0.821	0.737	0.997
2	3	0.975	–	0.829	1.006	0.614	–	0.939	0.915
	6	0.924	0.649	0.930	0.993	0.700	0.207	0.782	0.991
	12	0.935	0.662	0.982	0.978	0.858	0.889	0.492	0.977

See notes in Table 9.1

As for the CML estimators, the Δ index for the HK approach is always close to 1 but it shrinks when T and γ are both large. Surprisingly, the PCML estimator is also the best for a strong state dependence, since its Δ index is not only stable across the different designs but it often outperforms the INF approach. This result is in line with the findings of [15], where the bias of the PCML estimator is reported for a set of simulations with different levels of state dependence.

9.8 Empirical Application

In this section, we compare the performance of the estimators presented in Sect. 9.7 by empirically investigating the relationship between the presence of children in the household, accounting also for their age, and female labor force participation using a sample drawn from the Panel Study of Income Dynamics (PSID); see [22]. The dataset consists of $n = 1908$ married women between 19 and 59 years old in 1980, followed for $T = 6$ time occasions, ranging from 1979 to 1985. Similar applications of dynamic binary choice models have been proposed by Hyslop [53], Carro [29], Fernández-Val [40], and Dhaene and Jochmans [38].

The evaluation of the impact of fertility on female labor supply requires taking into account state dependence and unobserved heterogeneity. In particular, it is reasonable to assume that the occupation status at time t is strongly influenced by whether the woman was in the workforce at time $t - 1$, other things being equal, and that there might be unobservable factors, such as some degree of labor market

attachment and career preferences, which can likely affect both decisions about having children and labor supply.

For the data at issue, we specify a DL model as in (9.17), where the response variable is equal to 1 if a woman in a given year is in the labor force and 0 otherwise. The set of explanatory variables is similar to that of the seminal contribution of [53]. Along with the lagged dependent variable, we include a set of covariates containing the woman's age, its square, the husband's income, and three variables that report whether there are children aged between 0 – 2, 3 – 5, and 6 – 17 in the household, which will be denoted by $k2$, $k5$, and $k17$, respectively.

Table 9.4 reports the estimation results (coefficients with standard errors in parentheses) for the DL model obtained by the following estimators: ML, MPL proposed by Bartolucci et al. [19], SPJ shown in [38], HK proposed by Honoré and Kyriazidou [51], and PCML by Bartolucci and Nigro [15]. In this sample, 244 women do not participate in the labor market at all, 950 always participate, and 714 change their occupational status at least once.

Results show that the parameter estimates differ depending on the approach adopted, as also emerges from the simulation results in Sect. 9.7. In fact, the ML estimates of the parameter γ are between 25 and 50% smaller than those obtained with the alternative estimators and, in general, the whole set of estimated coefficients shows marked differences with respect to the others, due to the bias generated by the incidental parameter problem. As concerns the other estimators considered, results are coherent with those in the empirical literature on fertility and the female labor supply. In particular, there is a strong state dependence in the woman labor force participation, whereas there is a negative and statistically significant effect of the presence of children in the household on labor supply, albeit decreasing as the children's age increases.

Table 9.4 Empirical application: labor force participation

	γ	<i>Age</i>	<i>Age</i> ²	<i>Income</i>	<i>k2</i>	<i>k5</i>	<i>k17</i>
ML	0.577	0.189	-0.002	-0.012	-1.272	-0.886	-0.272
	(0.082)	(0.126)	(0.002)	(0.006)	(0.141)	(0.149)	(0.136)
MPL	1.350	0.171	-0.002	-0.010	-0.933	-0.576	-0.198
	(0.081)	(0.093)	(0.001)	(0.004)	(0.114)	(0.113)	(0.101)
SPJ	2.200	-0.015	-0.002	-0.006	-1.083	-0.713	-0.136
	(0.173)	(0.497)	(0.007)	(0.013)	(0.420)	(0.452)	(0.440)
HK	1.081	-1.332	0.018	-0.031	-1.010	-0.895	-0.411
	(0.208)	(0.723)	(0.011)	(0.015)	(0.367)	(0.408)	(0.498)
PCML	1.713	0.151	-0.002	-0.009	-0.909	-0.555	-0.173
	(0.103)	(0.083)	(0.001)	(0.004)	(0.099)	(0.102)	(0.094)

In this sample, 244 women do not participate in the labor market at all, 950 always participate, and 714 change their occupational status at least once

ML: Maximum Likelihood Estimator, MPL: Modified Profile Likelihood [19], SPJ: Split-panel Jackknife [38], HK: DL estimator [51], PCML: Pseudo Conditional Maximum Likelihood Estimator [15]

It is worth noting that the PCML and MPL estimators provide similar results in terms of coefficients and standard errors. With respect to the ML estimates, the state dependence coefficients are approximately three times larger and the coefficients relative to the number of the children aged in the three different ranges are smaller. In this setup, the PCML and MPL estimators are the most reliable, as the HK and SPJ approaches have some drawbacks. Other than a slow rate of convergence, the HK estimator is here based on a reduced sample size because of the presence of discrete predictors. Moreover, the covariates Age and Age^2 do not meet the regularity conditions for the identification of continuous explanatory variables proposed in the original work of [51]. Finally, the SPJ estimator may here exhibit some bias due to the limited number of time occasions of the dataset, which is $T = 6$ further to the initial observation.

9.9 Software

This section briefly recalls the main software components available for the estimation of binary choice models with fixed-effects. What follows is far from being exhaustive and will be mainly focused on packages available in R [64] and Stata.

For what concerns R, ML estimation can be easily performed by the command `glm()` provided by the `stats` package. However, when the number of subjects, and therefore that of the individual intercepts, is large, computation of the ML estimator can be efficiently dealt with by the `glmMML` package [26], which also provides routines for the estimation of random-effects models.

Many packages provide the different target-corrected estimators described in Sect. 9.5. In R, analytical bias-corrected estimators of [47] and [40] are provided by package `bife` [68]. Moreover, the MPL estimator [19] is provided by the package `panelMPL`, available at https://ruggerobellio.weebly.com/uploads/5/1/5/0/51505127/panelmpl_0.23.tar.gz. In Stata, package `XTSPJ` [70] performs the split-panel jackknife estimators proposed by Dhaene and Jochmans [38], while packages `LOGITFE` and `PROBITFE` [35, 36] provide a wide range of techniques such as analytical and jackknife corrections for models with individual and time fixed-effects [41].

Moving toward the conditional inference approach, different R routines provide the CML estimator [30] for the static logit model, such as function `clogit()` in package `survival` [71] and function `cquad_basic()` in package `cquad` [17]. Furthermore, `cquad` provides the estimators for the QE models [14, 20] and the PCML estimators for the dynamic logit model [15].

In Stata, the CML estimator can be performed by the commands `xtlogit` and `clogit`. Finally, module `CQUADR` [21] allows users to exploit the `cquad` package in Stata.

The R code for the replication of the simulations in Sect. 9.7 and the application in Sect. 9.8 is available at https://github.com/fravale/replication_fe_dyn_logit.

9.10 Conclusions

We reviewed recent fixed-effects approaches to the formulation and estimation of binary, static, and dynamic, panel data models. Fixed-effects models are popular in many applied fields, as they avoid distributional assumptions on the unobserved individual effects and allow them to be correlated with the model covariates. This chapter offers a unified perspective about the two main streams of literature focused on the inconsistency of the ML estimator arising from the incidental parameter problem, namely the target-adjusted estimators and the conditional inference, also by means of an extensive simulation study and an empirical application on female labor supply.

The main advantage of applying conditional inference is that CML estimators are fixed- T consistent, whereas target-corrected estimators are biased of order $O(T^{-2})$ and require T to grow faster than $n^{1/3}$ for confidence intervals of the ML estimator to be centered at their probability limit. This makes the CML a viable approach, especially for applications based on surveys collected with a rotating sampling scheme, such as most national household and workforce surveys. There are also some drawbacks when dynamic formulations are considered, as both the HK and PCML estimators do not share the same asymptotic properties as the CML estimator for static models. Nevertheless, our simulation study shows that they outperform target-corrected estimators, especially when T is small.

It is, however, worth recalling that the conditional inference approach is model specific, even though the approaches reviewed here can be easily extended for certain models for ordered and multinomial responses. In particular, the estimation of an ordered fixed-effects logit model with c response categories can be reduced to that of a fixed-effects binary logit model once the ordered response is dichotomized to generate $c - 1$ binary response variables [12, 30]. Moreover, the CML approach does not provide estimates of the individual intercepts. These are necessary to compute predictions and partial effects of explanatory variables on the response probability, which are often the object of interest in causal inference. In this respect, one solution is to adopt a mixed approach that exploits a bias-corrected estimator of the individual intercepts. This is proposed by Bartolucci and Pignini [18] and requires further investigation.

Acknowledgments F. Bartolucci acknowledges the financial support from the grant “Partial effects in econometric models for binary longitudinal data based on quadratic exponential distributions” of the University of Perugia (RICBASE2018).

Appendix

This appendix reports the additional results concerning the Monte Carlo simulation experiment described in Sect. 9.7 (Tables 9.5, 9.6, 9.7, 9.8, 9.9, 9.10, 9.11 and 9.12).

Table 9.5 Simulation results under the design based on Eqs. (9.23) and (9.24) with $\beta = 1, \gamma = 0$

Results for $\hat{\beta}$													
n	T	Mean Bias						RMSE					
		INF	ML	MPL	SPJ	HK	PCML	INF	ML	MPL	SPJ	HK	PCML
250	3	0.009	0.806	-0.025	-	0.100	0.023	0.088	0.848	0.150	-	0.275	0.133
	4	0.017	0.578	0.003	-	0.031	0.016	0.076	0.600	0.071	-	0.136	0.091
	6	0.007	0.330	0.023	-0.167	0.010	0.010	0.057	0.346	0.069	0.222	0.088	0.069
	8	-0.000	0.215	0.012	-0.126	0.005	0.002	0.048	0.227	0.055	0.151	0.065	0.054
	12	0.001	0.128	0.005	-0.062	0.002	0.000	0.038	0.137	0.042	0.078	0.048	0.042
500	3	0.004	0.771	-0.083	-	0.041	0.004	0.065	0.789	0.108	-	0.173	0.084
	4	0.001	0.548	-0.013	-	0.007	-0.000	0.051	0.562	0.083	-	0.098	0.069
	6	0.005	0.324	0.020	-0.142	0.008	0.006	0.040	0.331	0.049	0.174	0.061	0.047
	8	0.002	0.216	0.013	-0.124	0.004	0.002	0.034	0.222	0.041	0.137	0.050	0.039
	12	0.006	0.135	0.011	-0.055	0.006	0.006	0.030	0.141	0.035	0.065	0.038	0.034
1000	3	0.011	0.799	-0.021	-	0.033	0.020	0.043	0.810	0.100	-	0.117	0.065
	4	0.001	0.556	-0.004	-	0.010	0.005	0.035	0.562	0.037	-	0.072	0.047
	6	0.000	0.315	0.014	-0.143	0.003	0.000	0.029	0.319	0.036	0.157	0.046	0.034
	8	0.000	0.214	0.011	-0.121	0.003	0.000	0.024	0.217	0.030	0.127	0.035	0.028
	12	0.001	0.128	0.005	-0.059	0.001	0.001	0.020	0.130	0.021	0.064	0.025	0.021
n	T	Median Bias						MAE					
		INF	ML	MPL	SPJ	HK	PCML	INF	ML	HK	PCML	MPL	SPJ
250	3	0.009	0.780	-0.052	-	0.077	0.014	0.056	0.780	0.076	-	0.160	0.090
	4	0.011	0.570	0.001	-	0.021	0.007	0.051	0.570	0.042	-	0.077	0.051
	6	0.006	0.327	0.024	-0.161	0.009	0.008	0.038	0.327	0.046	0.164	0.054	0.047
	8	-0.002	0.213	0.010	-0.126	0.000	0.001	0.032	0.213	0.038	0.126	0.048	0.038
	12	-0.002	0.126	0.004	-0.064	0.000	-0.001	0.025	0.126	0.029	0.064	0.033	0.029
500	3	0.006	0.791	-0.072	-	0.011	0.010	0.044	0.791	0.072	-	0.111	0.066
	4	0.001	0.544	-0.007	-	0.001	-0.001	0.034	0.544	0.035	-	0.065	0.044
	6	0.006	0.321	0.019	-0.142	0.005	0.004	0.026	0.321	0.032	0.142	0.040	0.031
	8	0.002	0.215	0.012	-0.126	0.002	0.001	0.024	0.215	0.027	0.126	0.035	0.026
	12	0.008	0.134	0.011	-0.056	0.005	0.006	0.020	0.134	0.022	0.056	0.024	0.021
1000	3	0.014	0.804	-0.042	-	0.027	0.017	0.026	0.804	0.060	-	0.063	0.042
	4	0.001	0.554	-0.004	-	0.012	0.004	0.024	0.554	0.025	-	0.046	0.031
	6	0.001	0.315	0.014	-0.142	0.001	0.000	0.019	0.315	0.023	0.142	0.030	0.021
	8	0.000	0.214	0.012	-0.121	0.002	0.000	0.016	0.214	0.020	0.121	0.024	0.019
	12	0.000	0.128	0.006	-0.059	0.001	0.001	0.014	0.128	0.014	0.059	0.018	0.014

See notes in Table 9.1

Table 9.6 Simulation results under the design based on Eqs. (9.23) and (9.24) with $\beta = 1, \gamma = 0$

Results for $\hat{\gamma}$													
n	T	Mean Bias						RMSE					
		INF	ML	MPL	SPJ	HK	PCML	INF	ML	MPL	SPJ	HK	PCML
250	3	0.008	-2.729	0.055	-	0.000	0.039	0.160	2.796	0.350	-	0.539	0.407
	4	0.009	-1.695	-0.088	-	0.018	0.035	0.115	1.733	0.220	-	0.321	0.263
	6	-0.005	-0.974	-0.087	0.837	0.007	-0.001	0.104	0.996	0.185	0.908	0.202	0.172
	8	-0.005	-0.679	-0.056	0.345	-0.001	-0.006	0.090	0.700	0.158	0.404	0.158	0.150
	12	-0.001	-0.415	-0.022	0.122	-0.001	-0.000	0.068	0.431	0.111	0.179	0.113	0.110
500	3	-0.005	-2.759	0.003	-	-0.016	-0.012	0.101	2.794	0.173	-	0.378	0.287
	4	0.001	-1.695	-0.109	-	0.005	0.014	0.092	1.716	0.188	-	0.240	0.199
	6	-0.002	-0.975	-0.088	0.794	-0.000	-0.002	0.074	0.987	0.153	0.831	0.152	0.132
	8	0.001	-0.672	-0.050	0.341	0.001	0.001	0.063	0.681	0.111	0.369	0.112	0.102
	12	0.009	-0.401	-0.008	0.132	0.000	0.012	0.054	0.411	0.083	0.159	0.081	0.084
1000	3	-0.006	-2.748	-0.037	-	-0.018	-0.007	0.080	2.764	0.190	-	0.260	0.183
	4	0.001	-1.722	-0.130	-	-0.012	-0.009	0.064	1.732	0.177	-	0.168	0.141
	6	0.003	-0.970	-0.086	0.783	0.001	0.001	0.048	0.976	0.121	0.802	0.103	0.089
	8	-0.002	-0.674	-0.052	0.337	-0.001	-0.002	0.043	0.679	0.088	0.352	0.082	0.072
	12	-0.001	-0.415	-0.022	0.120	0.001	-0.001	0.035	0.419	0.058	0.135	0.058	0.054
n	T	Median Bias						MAE					
		INF	ML	MPL	SPJ	HK	PCML	INF	ML	MPL	SPJ	HK	PCML
250	3	-0.008	-2.703	0.017	-	0.009	0.061	0.123	2.703	0.184	-	0.367	0.263
	4	0.001	-1.680	-0.082	-	0.042	0.051	0.079	1.680	0.149	-	0.207	0.173
	6	-0.010	-0.976	-0.084	0.824	0.013	0.009	0.073	0.976	0.118	0.824	0.140	0.115
	8	-0.008	-0.669	-0.045	0.344	0.018	-0.000	0.058	0.669	0.105	0.344	0.107	0.105
	12	-0.002	-0.416	-0.021	0.114	-0.006	-0.000	0.046	0.416	0.071	0.124	0.075	0.072
500	3	-0.011	-2.759	0.000	-	-0.050	-0.003	0.074	2.759	0.097	-	0.241	0.209
	4	0.001	-1.683	-0.107	-	0.006	0.022	0.064	1.683	0.125	-	0.176	0.146
	6	-0.003	-0.978	-0.088	0.788	-0.002	-0.004	0.050	0.978	0.113	0.788	0.104	0.089
	8	0.001	-0.668	-0.046	0.336	0.003	0.003	0.040	0.668	0.076	0.336	0.078	0.068
	12	0.007	-0.402	-0.009	0.131	-0.001	0.010	0.041	0.402	0.049	0.131	0.051	0.050
1000	3	-0.002	-2.706	-0.000	-	0.004	0.001	0.059	2.706	0.075	-	0.193	0.139
	4	0.001	-1.716	-0.126	-	-0.012	-0.005	0.046	1.716	0.131	-	0.118	0.098
	6	0.001	-0.969	-0.086	0.778	0.001	-0.001	0.033	0.969	0.090	0.778	0.069	0.061
	8	-0.003	-0.669	-0.049	0.338	-0.000	0.001	0.029	0.669	0.059	0.338	0.056	0.048
	12	-0.001	-0.416	-0.022	0.120	-0.002	-0.001	0.023	0.416	0.040	0.120	0.040	0.033

See notes in Table 9.1

Table 9.7 Simulation results under the design based on Eqs. (9.23) and (9.24) with $\beta = 1, \gamma = 0.25$

Results for $\hat{\beta}$													
n	T	Mean Bias						RMSE					
		INF	ML	MPL	SPJ	HK	PCML	INF	ML	MPL	SPJ	HK	PCML
250	3	0.010	0.802	-0.025	-	0.116	0.017	0.089	0.844	0.159	-	0.318	0.127
	4	0.010	0.598	0.016	-	0.056	0.028	0.077	0.626	0.080	-	0.171	0.106
	6	0.004	0.314	0.016	-0.168	0.007	-0.000	0.052	0.327	0.062	0.226	0.082	0.062
	8	0.001	0.215	0.014	-0.124	0.006	0.002	0.047	0.226	0.053	0.149	0.064	0.051
	12	0.001	0.129	0.007	-0.061	0.004	0.001	0.040	0.139	0.045	0.079	0.052	0.044
500	3	0.009	0.789	-0.063	-	0.096	0.012	0.054	0.808	0.100	-	0.199	0.087
	4	0.005	0.561	0.006	-	0.030	0.009	0.049	0.573	0.052	-	0.108	0.066
	6	0.004	0.320	0.020	-0.155	0.014	0.004	0.041	0.328	0.049	0.187	0.063	0.047
	8	0.002	0.216	0.014	-0.120	0.008	0.003	0.034	0.222	0.040	0.134	0.049	0.038
	12	0.003	0.132	0.009	-0.053	0.008	0.004	0.027	0.137	0.030	0.063	0.037	0.029
1000	3	0.001	0.775	-0.031	-	0.033	0.006	0.044	0.787	0.092	-	0.122	0.068
	4	-0.001	0.542	-0.003	-	0.010	-0.002	0.036	0.548	0.037	-	0.074	0.047
	6	0.000	0.313	0.016	-0.143	0.006	-0.000	0.028	0.317	0.037	0.161	0.045	0.035
	8	0.001	0.214	0.013	-0.11	0.005	0.001	0.024	0.217	0.029	0.123	0.035	0.026
	12	0.000	0.128	0.006	-0.057	0.002	0.000	0.019	0.130	0.021	0.062	0.025	0.020
n	T	Median Bias						MAE					
		INF	ML	MPL	SPJ	HK	PCML	INF	ML	MPL	SPJ	HK	PCML
250	3	0.008	0.773	-0.051	-	0.058	0.004	0.056	0.773	0.073	-	0.145	0.080
	4	0.004	0.590	0.014	-	0.039	0.023	0.055	0.590	0.058	-	0.094	0.075
	6	-0.001	0.302	0.007	-0.167	0.008	-0.007	0.033	0.302	0.041	0.168	0.054	0.044
	8	0.002	0.211	0.011	-0.126	0.001	-0.001	0.031	0.211	0.034	0.127	0.040	0.033
	12	0.001	0.129	0.007	-0.060	0.004	0.001	0.026	0.129	0.028	0.061	0.034	0.027
500	3	0.003	0.773	-0.066	-	0.078	0.005	0.037	0.773	0.071	-	0.118	0.056
	4	0.003	0.556	0.005	-	0.024	0.008	0.033	0.556	0.035	-	0.069	0.044
	6	0.002	0.316	0.016	-0.158	0.008	-0.001	0.028	0.316	0.031	0.159	0.044	0.029
	8	0.003	0.217	0.016	-0.120	0.007	0.004	0.024	0.217	0.028	0.120	0.032	0.026
	12	-0.000	0.128	0.007	-0.054	0.005	0.001	0.017	0.128	0.019	0.054	0.025	0.019
1000	3	-0.002	0.771	-0.045	-	0.029	0.001	0.033	0.771	0.055	-	0.076	0.045
	4	-0.002	0.537	-0.004	-	0.005	-0.003	0.024	0.537	0.025	-	0.047	0.031
	6	-0.001	0.311	0.015	-0.145	0.005	-0.002	0.020	0.311	0.025	0.145	0.035	0.021
	8	-0.000	0.213	0.012	-0.117	0.003	0.000	0.016	0.213	0.020	0.117	0.024	0.018
	12	-0.000	0.128	0.006	-0.058	0.003	0.000	0.013	0.128	0.014	0.058	0.017	0.014

See notes in Table 9.1

Table 9.8 Simulation results under the design based on Eqs. (9.23) and (9.24) with $\beta = 1, \gamma = 0.25$

Results for $\hat{\gamma}$													
n	T	Mean Bias						RMSE					
		INF	ML	MPL	SPJ	HK	PCML	INF	ML	MPL	SPJ	HK	PCML
250	3	0.011	-2.675	-0.085	-	0.008	0.061	0.151	2.760	0.317	-	0.599	0.468
	4	0.006	-1.686	-0.182	-	-0.069	-0.018	0.132	1.722	0.275	-	0.350	0.272
	6	-0.007	-0.931	-0.093	0.790	-0.020	0.008	0.098	0.954	0.190	0.869	0.212	0.176
	8	0.003	-0.648	-0.051	0.329	-0.022	0.005	0.084	0.665	0.141	0.379	0.155	0.135
	12	-0.002	-0.404	-0.024	0.107	-0.025	-0.001	0.068	0.420	0.110	0.161	0.115	0.108
500	3	-0.017	-2.659	-0.118	-	-0.104	0.002	0.120	2.694	0.219	-	0.453	0.282
	4	0.002	-1.639	-0.165	-	-0.014	0.010	0.088	1.659	0.225	-	0.239	0.192
	6	-0.008	-0.941	-0.101	0.766	-0.033	-0.002	0.074	0.954	0.162	0.805	0.151	0.133
	8	0.001	-0.650	-0.054	0.316	-0.023	0.002	0.062	0.660	0.114	0.348	0.119	0.103
	12	-0.000	-0.406	-0.026	0.107	-0.034	-0.003	0.049	0.414	0.084	0.138	0.090	0.080
1000	3	0.001	-2.642	-0.128	-	-0.007	0.010	0.069	2.657	0.211	-	0.270	0.196
	4	-0.001	-1.647	-0.177	-	-0.035	-0.005	0.065	1.656	0.209	-	0.177	0.138
	6	-0.006	-0.937	-0.098	0.761	-0.024	0.003	0.047	0.943	0.132	0.781	0.109	0.095
	8	0.002	-0.647	-0.052	0.312	-0.020	0.005	0.043	0.652	0.088	0.327	0.084	0.072
	12	-0.001	-0.405	-0.025	0.107	-0.023	-0.001	0.034	0.409	0.059	0.123	0.064	0.054
n	T	Median Bias						MAE					
		INF	ML	MPL	SPJ	HK	PCML	INF	ML	MPL	SPJ	HK	PCML
250	3	0.010	-2.643	-0.095	-	0.002	0.056	0.107	2.643	0.197	-	0.356	0.296
	4	0.013	-1.671	-0.177	-	-0.091	-0.025	0.091	1.671	0.195	-	0.250	0.175
	6	-0.007	-0.946	-0.096	0.751	-0.019	0.001	0.074	0.946	0.134	0.751	0.145	0.118
	8	0.002	-0.643	-0.052	0.333	-0.024	0.006	0.061	0.643	0.094	0.333	0.106	0.094
	12	-0.001	-0.401	-0.022	0.112	-0.025	-0.000	0.046	0.401	0.075	0.124	0.078	0.075
500	3	-0.018	-2.679	-0.091	-	-0.098	0.024	0.079	2.679	0.120	-	0.315	0.179
	4	0.000	-1.631	-0.166	-	-0.016	0.014	0.059	1.631	0.171	-	0.159	0.130
	6	-0.007	-0.952	-0.108	0.751	-0.046	-0.008	0.048	0.952	0.125	0.751	0.112	0.088
	8	0.001	-0.650	-0.052	0.316	-0.023	0.004	0.042	0.650	0.079	0.316	0.082	0.066
	12	0.002	-0.404	-0.026	0.104	-0.038	-0.002	0.031	0.404	0.056	0.106	0.065	0.053
1000	3	0.002	-2.633	-0.110	-	-0.003	0.021	0.046	2.633	0.117	-	0.179	0.139
	4	-0.000	-1.649	-0.180	-	-0.039	-0.008	0.040	1.649	0.180	-	0.112	0.093
	6	-0.008	-0.926	-0.094	0.758	-0.022	0.007	0.032	0.926	0.097	0.758	0.072	0.064
	8	0.003	-0.645	-0.050	0.313	-0.025	0.005	0.028	0.645	0.058	0.313	0.059	0.049
	12	-0.000	-0.404	-0.025	0.106	-0.026	-0.001	0.022	0.404	0.038	0.106	0.044	0.036

See notes in Table 9.1

Table 9.9 Simulation results under the design based on Eqs. (9.23) and (9.24) with $\beta = 1, \gamma = 1$

Results for $\hat{\beta}$													
n	T	Mean Bias						RMSE					
		INF	ML	MPL	SPJ	HK	PCML	INF	ML	MPL	SPJ	HK	PCML
250	3	0.008	0.788	-0.005	-	0.188	0.020	0.098	0.837	0.153	-	0.521	0.150
	4	0.001	0.554	0.013	-	0.052	0.009	0.081	0.588	0.085	-	0.192	0.112
	6	0.011	0.341	0.036	-0.174	0.052	0.018	0.074	0.359	0.078	0.244	0.101	0.075
	8	0.004	0.220	0.019	-0.113	0.017	0.003	0.049	0.233	0.059	0.144	0.073	0.057
	12	0.004	0.136	0.012	-0.054	0.010	0.005	0.040	0.145	0.043	0.075	0.049	0.043
500	3	0.007	0.781	-0.040	-	0.114	0.011	0.064	0.805	0.103	-	0.249	0.100
	4	0.000	0.540	0.008	-	0.045	0.000	0.053	0.555	0.057	-	0.128	0.073
	6	0.006	0.320	0.026	-0.121	0.021	0.005	0.039	0.329	0.055	0.158	0.074	0.051
	8	0.007	0.232	0.027	-0.092	0.020	0.013	0.041	0.239	0.052	0.113	0.060	0.047
	12	0.001	0.133	0.009	-0.052	0.008	0.002	0.028	0.137	0.031	0.063	0.037	0.030
1000	3	0.001	0.762	-0.019	-	0.081	0.001	0.050	0.776	0.083	-	0.185	0.080
	4	-0.002	0.531	0.005	-	0.031	-0.005	0.038	0.539	0.040	-	0.091	0.051
	6	0.003	0.319	0.025	-0.124	0.022	0.004	0.029	0.323	0.041	0.145	0.055	0.035
	8	0.002	0.217	0.017	-0.103	0.011	0.002	0.026	0.221	0.034	0.112	0.040	0.029
	12	-0.002	0.128	0.005	-0.053	0.002	-0.002	0.020	0.131	0.022	0.059	0.027	0.022
n	T	Median Bias						MAE					
		INF	ML	MPL	SPJ	HK	PCML	INF	ML	MPL	SPJ	HK	PCML
250	3	0.005	0.763	-0.043	-	0.087	0.008	0.063	0.763	0.076	-	0.206	0.091
	4	-0.007	0.529	0.007	-	0.030	-0.005	0.060	0.529	0.055	-	0.107	0.075
	6	0.006	0.339	0.028	-0.171	0.058	0.018	0.049	0.339	0.056	0.174	0.079	0.048
	8	0.003	0.217	0.014	-0.113	0.012	-0.002	0.031	0.217	0.037	0.114	0.044	0.034
	12	-0.000	0.131	0.008	-0.061	0.008	0.003	0.026	0.131	0.030	0.061	0.034	0.031
500	3	0.006	0.762	-0.041	-	0.090	0.003	0.043	0.762	0.060	-	0.141	0.069
	4	-0.001	0.535	0.006	-	0.041	-0.004	0.036	0.535	0.040	-	0.080	0.051
	6	0.007	0.319	0.022	-0.123	0.021	0.000	0.025	0.319	0.036	0.128	0.048	0.033
	8	0.003	0.229	0.025	-0.088	0.022	0.010	0.028	0.229	0.034	0.088	0.042	0.032
	12	0.002	0.131	0.008	-0.054	0.007	0.001	0.020	0.131	0.022	0.054	0.025	0.021
1000	3	0.002	0.751	-0.033	-	0.060	-0.007	0.036	0.751	0.052	-	0.102	0.049
	4	-0.004	0.530	0.004	-	0.025	-0.008	0.025	0.530	0.029	-	0.057	0.036
	6	0.001	0.319	0.025	-0.123	0.017	0.004	0.019	0.319	0.028	0.123	0.035	0.024
	8	0.002	0.217	0.018	-0.104	0.010	0.001	0.019	0.217	0.023	0.104	0.025	0.019
	12	-0.001	0.128	0.006	-0.053	0.003	-0.002	0.014	0.128	0.015	0.053	0.019	0.015

See notes in Table 9.1

Table 9.10 Simulation results under the design based on Eqs. (9.23) and (9.24) with $\beta = 1, \gamma = 1$

Results for $\hat{\gamma}$													
n	T	Mean Bias						RMSE					
		INF	ML	MPL	SPJ	HK	PCML	INF	ML	MPL	SPJ	HK	PCML
250	3	-0.003	-2.389	-0.406	-	0.066	0.091	0.153	2.450	0.488	-	0.768	0.469
	4	0.010	-1.449	-0.301	-	-0.089	0.043	0.133	1.489	0.373	-	0.375	0.302
	6	0.013	-0.850	-0.137	0.800	-0.105	0.008	0.137	0.888	0.252	0.895	0.290	0.233
	8	0.005	-0.605	-0.077	0.227	-0.103	0.001	0.083	0.626	0.164	0.295	0.196	0.150
	12	-0.001	-0.361	-0.014	0.102	-0.085	0.018	0.069	0.379	0.111	0.169	0.149	0.112
500	3	0.000	-2.395	-0.438	-	-0.077	0.048	0.115	2.430	0.496	-	0.465	0.328
	4	0.003	-1.451	-0.303	-	-0.079	0.029	0.091	1.471	0.342	-	0.285	0.213
	6	0.010	-0.857	-0.147	0.677	-0.092	0.004	0.076	0.872	0.198	0.724	0.182	0.146
	8	0.006	-0.606	-0.079	0.243	-0.097	0.002	0.059	0.615	0.122	0.270	0.149	0.095
	12	-0.004	-0.387	-0.039	0.070	-0.102	-0.008	0.049	0.397	0.091	0.117	0.133	0.083
1000	3	-0.004	-2.378	-0.458	-	-0.084	0.031	0.076	2.393	0.494	-	0.331	0.218
	4	-0.002	-1.459	-0.311	-	-0.076	0.018	0.067	1.470	0.332	-	0.213	0.155
	6	0.003	-0.846	-0.139	0.695	-0.075	0.012	0.049	0.852	0.163	0.715	0.139	0.095
	8	0.001	-0.601	-0.075	0.240	-0.082	0.005	0.044	0.607	0.107	0.261	0.121	0.080
	12	0.002	-0.378	-0.030	0.075	-0.083	0.001	0.034	0.383	0.065	0.100	0.107	0.058
n	T	Median Bias						MAE					
		INF	ML	MPL	SPJ	HK	PCML	INF	ML	MPL	SPJ	HK	PCML
250	3	0.011	-2.376	-0.412	-	0.021	0.037	0.103	2.376	0.413	-	0.381	0.307
	4	0.005	-1.434	-0.299	-	-0.101	0.015	0.091	1.434	0.300	-	0.246	0.206
	6	0.014	-0.848	-0.131	0.838	-0.127	0.015	0.081	0.848	0.181	0.838	0.209	0.185
	8	0.006	-0.618	-0.092	0.214	-0.098	-0.009	0.052	0.618	0.123	0.218	0.133	0.106
	12	-0.005	-0.369	-0.022	0.110	-0.082	0.009	0.042	0.369	0.079	0.121	0.105	0.074
500	3	0.002	-2.388	-0.434	-	-0.096	0.039	0.077	2.388	0.434	-	0.300	0.205
	4	0.002	-1.446	-0.301	-	-0.097	0.026	0.062	1.446	0.301	-	0.196	0.142
	6	0.007	-0.861	-0.144	0.674	-0.094	-0.004	0.049	0.861	0.150	0.674	0.130	0.108
	8	0.008	-0.598	-0.075	0.240	-0.102	0.001	0.038	0.598	0.080	0.240	0.106	0.057
	12	-0.004	-0.385	-0.037	0.072	-0.104	-0.005	0.033	0.385	0.062	0.087	0.105	0.058
1000	3	0.001	-2.370	-0.436	-	-0.087	0.050	0.052	2.370	0.436	-	0.201	0.158
	4	-0.002	-1.458	-0.308	-	-0.080	0.013	0.047	1.458	0.308	-	0.138	0.105
	6	0.002	-0.844	-0.136	0.694	-0.071	0.012	0.035	0.844	0.137	0.694	0.095	0.068
	8	0.000	-0.604	-0.076	0.236	-0.085	0.003	0.029	0.604	0.082	0.236	0.091	0.054
	12	-0.001	-0.379	-0.030	0.078	-0.085	-0.001	0.022	0.379	0.044	0.078	0.086	0.038

See notes in Table 9.1

Table 9.11 Simulation results under the design based on Eqs. (9.23) and (9.24) with $\beta = 1, \gamma = 2$

Results for $\hat{\beta}$													
n	T	Mean Bias						RMSE					
		INF	ML	MPL	SPJ	HK	PCML	INF	ML	MPL	SPJ	HK	PCML
250	3	0.014	0.812	-0.006	-	0.351	0.038	0.105	0.875	0.130	-	0.698	0.188
	4	0.008	0.548	0.008	-	0.088	0.007	0.082	0.583	0.089	-	0.209	0.116
	6	0.004	0.334	0.030	-0.135	0.035	0.011	0.065	0.355	0.079	0.223	0.117	0.085
	8	0.025	0.266	0.051	-0.035	0.052	0.043	0.054	0.273	0.067	0.086	0.078	0.067
	12	0.000	0.141	0.011	-0.047	-0.002	-0.001	0.042	0.152	0.048	0.074	0.058	0.046
500	3	0.012	0.780	-0.011	-	0.190	0.022	0.074	0.813	0.122	-	0.370	0.133
	4	0.003	0.540	0.004	-	0.064	0.001	0.057	0.558	0.060	-	0.168	0.084
	6	0.002	0.325	0.025	-0.115	0.027	0.004	0.044	0.334	0.055	0.162	0.083	0.056
	8	0.002	0.230	0.024	-0.085	0.012	0.004	0.036	0.237	0.048	0.110	0.063	0.045
	12	0.001	0.141	0.012	-0.044	-0.002	-0.001	0.030	0.146	0.035	0.059	0.041	0.033
1000	3	-0.003	0.739	-0.036	-	0.116	-0.003	0.048	0.753	0.061	-	0.217	0.081
	4	0.000	0.528	0.001	-	0.047	-0.005	0.038	0.536	0.039	-	0.107	0.054
	6	-0.000	0.321	0.022	-0.097	0.016	-0.000	0.032	0.326	0.044	0.126	0.063	0.042
	8	0.004	0.231	0.024	-0.078	0.011	0.004	0.028	0.235	0.040	0.093	0.045	0.033
	12	0.001	0.143	0.014	-0.040	0.001	0.001	0.022	0.146	0.028	0.050	0.030	0.025
n	T	Median Bias						MAE					
		INF	ML	MPL	SPJ	HK	PCML	INF	ML	MPL	SPJ	HK	PCML
250	31	0.006	0.778	-0.022	-	0.206	0.013	0.065	0.778	0.078	-	0.255	0.115
	4	0.002	0.523	0.004	-	0.070	-0.004	0.055	0.523	0.062	-	0.134	0.077
	6	0.005	0.330	0.029	-0.118	0.029	0.008	0.044	0.330	0.053	0.146	0.070	0.055
	8	0.023	0.268	0.053	-0.034	0.064	0.042	0.036	0.268	0.053	0.061	0.064	0.042
	12	-0.000	0.139	0.010	-0.045	-0.003	-0.003	0.029	0.139	0.032	0.050	0.039	0.031
500	3	0.012	0.753	-0.031	-	0.139	-0.008	0.052	0.753	0.056	-	0.190	0.083
	4	-0.001	0.524	0.001	-	0.041	-0.007	0.037	0.524	0.041	-	0.095	0.056
	6	0.000	0.323	0.025	-0.113	0.023	0.002	0.031	0.323	0.035	0.119	0.053	0.037
	8	0.004	0.229	0.021	-0.084	0.014	0.004	0.023	0.229	0.031	0.086	0.042	0.030
	12	0.000	0.137	0.009	-0.047	-0.003	-0.003	0.020	0.137	0.021	0.048	0.028	0.022
1000	3	-0.003	0.741	-0.036	-	0.108	-0.003	0.035	0.741	0.043	-	0.128	0.057
	4	-0.002	0.527	0.002	-	0.045	-0.005	0.026	0.527	0.025	-	0.066	0.035
	6	-0.001	0.316	0.021	-0.096	0.014	-0.003	0.022	0.316	0.028	0.099	0.043	0.026
	8	0.004	0.231	0.024	-0.078	0.010	0.004	0.018	0.231	0.027	0.078	0.028	0.021
	12	0.001	0.142	0.013	-0.042	0.000	0.001	0.015	0.142	0.019	0.042	0.020	0.017

See notes in Table 9.1

Table 9.12 Simulation results under the design based on Eqs. (9.23) and (9.24) with $\beta = 1$, $\gamma = 2$

Results for $\hat{\gamma}$													
n	T	Mean Bias						RMSE					
		INF	ML	MPL	SPJ	HK	PCML	INF	ML	MPL	SPJ	HK	PCML
250	3	0.016	-2.012	-0.836	-	0.162	0.213	0.196	2.099	0.903	-	1.198	0.639
	4	0.018	-1.241	-0.504	-	-0.093	0.081	0.166	1.307	0.570	-	0.530	0.390
	6	0.003	-0.759	-0.227	0.610	-0.174	0.020	0.124	0.799	0.299	0.723	0.350	0.232
	8	0.065	-0.514	-0.080	0.296	-0.098	0.091	0.145	0.553	0.189	0.408	0.197	0.230
	12	0.002	-0.372	-0.054	0.043	-0.205	-0.006	0.091	0.399	0.148	0.166	0.261	0.141
500	3	0.004	-1.995	-0.778	-	0.007	0.209	0.154	2.047	0.821	-	0.742	0.507
	4	0.010	-1.255	-0.512	-	-0.151	0.062	0.115	1.286	0.545	-	0.375	0.282
	6	-0.000	-0.766	-0.234	0.610	-0.160	0.016	0.087	0.785	0.272	0.664	0.261	0.161
	8	0.000	-0.562	-0.127	0.166	-0.176	0.004	0.072	0.577	0.173	0.228	0.236	0.128
	12	0.001	-0.369	-0.052	0.039	-0.184	-0.007	0.058	0.381	0.105	0.116	0.214	0.093
1000	3	0.005	-2.050	-0.878	-	-0.108	0.139	0.093	2.071	0.912	-	0.441	0.301
	4	0.008	-1.261	-0.519	-	-0.143	0.048	0.080	1.273	0.535	-	0.291	0.179
	6	0.001	-0.763	-0.231	0.616	-0.165	0.011	0.061	0.773	0.252	0.642	0.219	0.120
	8	0.005	-0.554	-0.120	0.183	-0.159	0.008	0.054	0.562	0.146	0.216	0.198	0.091
	12	0.002	-0.366	-0.050	0.042	-0.158	-0.005	0.042	0.373	0.081	0.085	0.176	0.065
n	T	Median Bias						MAE					
		INF	ML	MPL	SPJ	HK	PCML	INF	ML	MPL	SPJ	HK	PCML
250	3	0.002	-2.019	-0.838	-	-0.084	0.145	0.135	2.019	0.838	-	0.640	0.388
	4	0.008	-1.254	-0.506	-	-0.147	0.062	0.119	1.254	0.506	-	0.389	0.276
	6	0.006	-0.756	-0.224	0.603	-0.190	0.031	0.084	0.756	0.226	0.604	0.264	0.156
	8	0.061	-0.576	-0.119	0.276	-0.088	0.056	0.069	0.576	0.132	0.276	0.113	0.131
	12	0.002	-0.375	-0.055	0.037	-0.223	-0.015	0.060	0.375	0.099	0.111	0.227	0.089
500	3	-0.017	-1.995	-0.780	-	-0.137	0.185	0.089	1.995	0.780	-	0.492	0.362
	4	0.011	-1.260	-0.515	-	-0.161	0.057	0.080	1.260	0.515	-	0.262	0.192
	6	-0.003	-0.768	-0.232	0.609	-0.170	0.010	0.058	0.768	0.232	0.609	0.197	0.107
	8	-0.002	-0.562	-0.126	0.168	-0.176	0.003	0.050	0.562	0.129	0.175	0.183	0.089
	12	0.001	-0.369	-0.053	0.042	-0.188	-0.010	0.038	0.369	0.074	0.082	0.189	0.064
1000	3	-0.004	-2.024	-0.850	-	-0.131	0.129	0.064	2.024	0.850	-	0.312	0.205
	4	0.008	-1.256	-0.516	-	-0.148	0.055	0.053	1.256	0.516	-	0.200	0.108
	6	0.003	-0.766	-0.233	0.616	-0.163	0.012	0.041	0.766	0.233	0.616	0.167	0.077
	8	0.004	-0.552	-0.118	0.183	-0.164	0.004	0.037	0.552	0.118	0.183	0.165	0.063
	12	0.002	-0.363	-0.046	0.041	-0.157	-0.002	0.028	0.363	0.055	0.056	0.157	0.046

See notes in Table 9.1

References

1. Akay, A.: Finite-sample comparison of alternative methods for estimating dynamic panel data models. *J. Appl. Econ.* **27**, 1189–1204 (2012)
2. Alessie, R., Hochguertel, S., Van Soest, A.: Ownership of stocks and mutual funds: a panel data analysis. *Rev. Econ. Stat.* **86**, 783–796 (2004)
3. Andersen, E.B.: Asymptotic properties of conditional maximum-likelihood estimators. *J. R. Stat. Soc. Ser. B* **32**, 283–301 (1970)
4. Anderson, T.W., Hsiao, C.: Estimation of dynamic models with error components. *J. Amer. Stat. Assoc.* **76**, 598–606 (1981)
5. Arellano, M.: *Panel Data Econometrics*. Oxford University Press, New York (2003)
6. Arellano, M.: Discrete choices with panel data. *Invest. Econ.* **27**, 423–458 (2003)
7. Arellano, M., Bond, S.: Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Rev. Econ. Stud.* **58**, 277–297 (1991)
8. Arellano, M., Bover, O.: Another look at the instrumental variable estimation of error-components models. *J. Econ.* **68**, 29–51 (1995)
9. Arellano, M., Hahn, J.: Understanding bias in nonlinear panel models: some recent developments. *Econ. Soc. Monogr.* **43**, 381 (2007)
10. Arellano, M., Hahn, J.: A likelihood-based approximate solution to the incidental parameter problem in dynamic nonlinear models with multiple effects. *Global Econ. Rev.* **45**, 251–274 (2016)
11. Arulampalam, W., Stewart, M.B.: Simplified implementation of the Heckman estimator of the dynamic probit model and a comparison with alternative estimators. *Oxf. Bull. Econ. Stat.* **71**, 659–681 (2009)
12. Baetschmann, G., Staub, K.E., Winkelmann, R.: Consistent estimation of the fixed effects ordered logit model. *J. R. Stat. Soc. Ser. A* **178**, 685–703 (2015)
13. Bandeen-Roche, K., Miglioretti, D.L., Zeger, S.L., Rathouz, P.J.: Latent variable regression for multiple discrete outcomes. *J. Amer. Stat. Assoc.* **92**, 1375–1386 (1997)
14. Bartolucci, F., Nigro, V.: A dynamic model for binary panel data with unobserved heterogeneity admitting a \sqrt{n} -consistent conditional estimator. *Econometrica* **78**, 719–733 (2010)
15. Bartolucci, F., Nigro, V.: Pseudo conditional maximum likelihood estimation of the dynamic logit model for binary panel data. *J. Econ.* **170**, 102–116 (2012)
16. Bartolucci, F., Pennoni, F.: On the approximation of the quadratic exponential distribution in a latent variable context. *Biometrika* **94**, 745–754 (2007)
17. Bartolucci, F., Pigini, C.: cquad: An R and Stata package for conditional maximum likelihood estimation of dynamic binary panel data models. *J. Stat. Softw.* **78**, 1–26 (2017)
18. Bartolucci, F., Pigini, C.: Partial effects estimation for fixed-effects logit panel data models. Technical Report, MPRA Paper No. 92251 (2019)
19. Bartolucci, F., Bellio, R., Salvan, A., Sartori, N.: Modified profile likelihood for fixed-effects panel data models. *Econ. Rev.* **35**, 1271–1289 (2016)
20. Bartolucci, F., Nigro, V., Pigini, C.: Testing for state dependence in binary panel data with individual covariates by a modified quadratic exponential model. *Econ. Rev.* **37**, 61–88 (2018)
21. Bartolucci, F., Pigini, C., Valentini, F.: CQUADR: Stata module to estimate Quadratic Exponential models running the cquad R package. Statistical Software Components, Boston College Department of Economics (2020)
22. Beale, A., Campbell, F., Dascola, M., Insolera, N., et al.: PSID main interview user manual: Release 2021. Institute for Social Research, University of Michigan (2021)
23. Bester, C.A., Hansen, C.: A penalty function approach to bias reduction in nonlinear panel models with fixed effects. *J. Business Econ. Stat.* **27**, 131–148 (2009)
24. Bettin, G., Lucchetti, R.: Steady streams and sudden bursts: persistence patterns in remittance decisions. *J. Populat. Econ.* **29**, 263–292 (2016)
25. Blundell, R., Bond, S.: Initial conditions and moment restrictions in dynamic panel data models. *J. Econ.* **87**, 115–143 (1998)

26. Broström, G.: *glimmML: Generalized Linear Models with Clustering*. R package version 1.1.1 (2020)
27. Brown, S., Ghosh, P., Taylor, K.: *The existence and persistence of household financial hardship*. Technical Report, Department of Economics, University of Sheffield (2012)
28. Cameron, A.C., Trivedi, P.K.: *Microeconometrics: Methods and Applications*. Cambridge University Press, New York (2005)
29. Carro, J.M.: Estimating dynamic panel data discrete choice models with fixed effects. *J. Econ.* **140**, 503–528 (2007)
30. Chamberlain, G.: Analysis of covariance with qualitative data. *Rev. Econ. Stud.* **47**, 225–238 (1980)
31. Chamberlain, G.: *Feedback in panel data models*. Technical Report, Harvard-Institute of Economic Research (1993)
32. Cox, D.R.: The analysis of multivariate binary data. *Appl. Stat.* **21**, 113–120 (1972)
33. Cox, D.R., Reid, N.: Parameter orthogonality and approximate conditional inference. *J. R. Stat. Soc. Ser. B* **49**, 1–39 (1987)
34. Cox, D.R., Wermuth, N.: A note on the quadratic exponential binary distribution. *Biometrika* **81**(2), 403–408 (1994)
35. Cruz-Gonzalez, M., Fernandez-Val, I., Weidner, M.: *LOGITFE: Stata module to compute analytical and jackknife bias corrections for fixed effects estimators of panel logit models with individual and time effects*. Statistical Software Components, Boston College Department of Economics (2016)
36. Cruz-Gonzalez, M., Fernandez-Val, I., Weidner, M.: *PROBITFE: Stata module to compute analytical and jackknife bias corrections for fixed effects estimators of panel probit models with individual and time effects*. Statistical Software Components, Boston College Department of Economics (2016)
37. Dayton, C.M., Macready, G.B.: Concomitant-variable latent-class models. *J. Amer. Stat. Assoc.* **83**, 173–178 (1988)
38. Dhaene, G., Jochmans, K.: Split-panel jackknife estimation of fixed-effect models. *Rev. Econ. Stud.* **82**, 991–1030 (2015)
39. Diggle, P.J., Heagerty, P., Liang, K.-Y., Zeger, S.: *Analysis of Longitudinal Data*. Oxford University Press, Oxford (2002)
40. Fernández-Val, I.: Fixed effects estimation of structural parameters and marginal effects in panel probit models. *J. Econ.* **150**, 71–85 (2009)
41. Fernández-Val, I., Weidner, M.: Individual and time effects in nonlinear panel models with large N , T . *J. Econ.* **192**, 291–312 (2016)
42. Fernández-Val, I., Weidner, M.: Fixed effects estimation of large- T panel data models. *Ann. Rev. Econ.* **10**, 109–138 (2018)
43. Fitzmaurice, G., Davidian, M., Verbeke, G., Molenberghs, G.: *Longitudinal Data Analysis*. CRC Press, Boca Raton (2008)
44. Formann, A.K.: Mixture analysis of multivariate categorical data with covariates and missing entries. *Comput. Stat. Data Anal.* **51**, 5236–5246 (2007)
45. Giarda, E.: Persistency of financial distress amongst Italian households: evidence from dynamic models for binary panel data. *J. Banking Finance* **37**, 3425–3434 (2013)
46. Hahn, J., Kuersteiner, G.: Bias reduction for dynamic nonlinear panel models with fixed effects. *Econ. Theory* **27**, 1152–1191 (2011)
47. Hahn, J., Newey, W.: Jackknife and analytical bias reduction for nonlinear panel models. *Econometrica* **72**, 1295–1319 (2004)
48. Heckman, J.J.: Heterogeneity and state dependence. In *Studies in Labor Markets*, pp. 91–140. University of Chicago Press, Chicago (1981)
49. Heckman, J.J.: The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process and some Monte Carlo evidence. In: Manski, C.F., McFadden, D. (eds.) *Structural Analysis of Discrete Data with Economic Applications*. MIT Press, Cambridge (1981)

50. Heckman, J.J., Borjas, G.J.: Does unemployment cause future unemployment? Definitions, questions and answers from a continuous time model of heterogeneity and state dependence. *Economica* **47**, 247–283 (1980)
51. Honoré, B.E., Kyriazidou, E.: Panel data discrete choice models with lagged dependent variables. *Econometrica* **68**, 839–874 (2000)
52. Hsiao, C.: *Analysis of Panel Data*. Cambridge University Press, New York (2014)
53. Hyslop, D.R.: State dependence, serial correlation and heterogeneity in intertemporal labor force participation of married women. *Econometrica* **67**, 1255–1294 (1999)
54. Jiang, J., Nguyen, T.: *Linear and Generalized Linear Mixed Models and Their Applications*. Springer Nature, New York (2021)
55. Lancaster, T.: The incidental parameter problem since 1948. *J. Econ.* **95**, 391–413 (2000)
56. Lucchetti, R., Pignini, C.: DPB: dynamic panel binary data models in gretl. *J. Stat. Softw.* **79**, 1–33 (2017)
57. McCullagh, P., Nelder, J.A.: *Generalized Linear Models*. Chapman & Hall, Boca Raton (1989)
58. McFadden, D.: Conditional logit analysis of qualitative choice behavior. In: Zarembka, P. (ed.) *Frontiers in Econometrics*, pp. 105–142. Academic, Cambridge (1974)
59. McLachlan, G.J., Peel, D.: *Finite Mixture Models*. Wiley, New York (2004)
60. Mundlak, Y.: On the pooling of time series and cross section data. *Econometrica* **46**, 69–85 (1978)
61. Neyman, J., Scott, E.L.: Consistent estimates based on partially consistent observations. *Econometrica* **16**, 1–32 (1948)
62. Pignini, C., Presbitero, A.F., Zazzaro, A.: State dependence in access to credit. *J. Finan. Stab.* **27**, 17–34 (2016)
63. Quenouille, M.H.: Notes on bias in estimation. *Biometrika* **43**, 353–360 (1956)
64. R Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna (2021)
65. Rabe-Hesketh, S., Skrondal, A.: Avoiding biased versions of Wooldridge’s simple solution to the initial conditions problem. *Econ. Lett.* **120**, 346–349 (2013)
66. Severini, T.A.: An approximation to the modified profile likelihood function. *Biometrika* **85**, 403–411 (1998)
67. Skrondal, A., Rabe-Hesketh, S.: Handling initial conditions and endogenous covariates in dynamic/transition models for binary data with unobserved heterogeneity. *J. R. Stat. Soc. Ser. C* **63**, 211–237 (2014)
68. Stammann, A., Heiss, F., McFadden, D.: Estimating fixed effects logit models with large panel data. In: *VfS Annual Conference 2016 (Augsburg): Demographic Change* (2016)
69. Stroup, W.W.: *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*. Taylor & Francis, Boca Raton (2012)
70. Sun, Y., Dhaene, G.: XTSPJ: Stata module for split-panel jackknife estimation. In: *Statistical Software Components*, Boston College Department of Economics (2019)
71. Therneau, T.M.: *A Package for Survival Analysis in R*. R package version 3.2-11 (2021)
72. Verbeke, G., Molenberghs, G.: *Linear Mixed Models for Longitudinal Data*. Springer, New York (2001)
73. Wedel, M.: Concomitant variables in finite mixture models. *Statistica Neerlandica* **56**, 362–375 (2002)
74. Wooldridge, J.M.: Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. *J. Appl. Econ.* **20**, 39–54 (2005)
75. Wooldridge, J.M.: *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge (2010)