Iowa Research Online
The University of Iowa's Institutional Repository

University of Iowa
Iowa Research Online

Theses and Dissertations

2011

Likelihood-based inference for antedependence (Markov) models for categorical longitudinal data

Yunlong Xie
University of Iowa

Copyright 2011 YUNLONG XIE

This dissertation is available at Iowa Research Online: http://ir.uiowa.edu/etd/1193

Recommended Citation
Xie, Yunlong. "Likelihood-based inference for antedependence (Markov) models for categorical longitudinal data." PhD (Doctor of Philosophy) thesis, University of Iowa, 2011.
http://ir.uiowa.edu/etd/1193.

Follow this and additional works at: http://ir.uiowa.edu/etd

Part of the Statistics and Probability Commons

LIKELIHOOD-BASED INFERENCE FOR ANTEDEPENDENCE (MARKOV) MODELS FOR CATEGORICAL LONGITUDINAL DATA

by

Yunlong Xie

An Abstract

Of a thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Statistics in the
Graduate College of The
University of Iowa

July 2011

Thesis Supervisor: Professor Dale L. Zimmerman

ABSTRACT

Antedependence (AD) of order p, also known as the Markov property of order p, is a property of index-ordered random variables in which each variable, given at least p immediately preceding variables, is independent of all further preceding variables. Zimmerman and Núñez-Antón (2010) present statistical methodology for fitting and performing inference for AD models for continuous (primarily normal) longitudinal data. But analogous AD-model methodology for categorical longitudinal data has not yet been well developed. In this thesis, we derive maximum likelihood estimators of transition probabilities under antedependence of any order, and we use these estimators to develop likelihood-based methods for determining the order of antedependence of categorical longitudinal data. Specifically, we develop a penalized likelihood method for determining variable-order antedependence structure, and we derive the likelihood ratio test, score test, Wald test and an adaptation of Fisher's exact test for p^{th}-order antedependence against the unstructured (saturated) multinomial model. Simulation studies show that the score (Pearson's Chi-square) test performs better than all the other methods for complete and monotone missing data, while the likelihood ratio test is applicable for data with arbitrary missing pattern. But since the likelihood ratio test is oversensitive under the null hypothesis, we modify it by equating the expectation of the test statistic to its degrees of freedom so that it has actual size closer to nominal size. Additionally, we modify the likelihood ratio tests for use in testing for p^{th}-order antedependence against q^{th}-order antedependence, where $q > p$, and for testing nested variable-order antedependence models. We extend the methods to deal with data having a monotone or arbitrary missing pattern. For antedependence models of constant order

p, we develop methods for testing transition probability stationarity and strict stationarity and for maximum likelihood estimation of parametric generalized linear models that are transition probability stationary AD(p) models. The methods are illustrated using three data sets.

KEY WORDS: Antedependence; Categorical longitudinal data; Wald test; Score test; Likelihood ratio test; Penalized likelihood; Monotone missing (or monotone drop-ins); EM algorithm.

Abstract Approved: _____
Thesis Supervisor

_____
Title and Department

_____
Date

LIKELIHOOD-BASED INFERENCE FOR ANTEDEPENDENCE (MARKOV) MODELS FOR CATEGORICAL LONGITUDINAL DATA

by

Yunlong Xie

A thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Statistics in the
Graduate College of The
University of Iowa

July 2011

Thesis Supervisor: Professor Dale L. Zimmerman

Copyright by
YUNLONG XIE
2011
All Rights Reserved

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Yunlong Xie

has been approved by the Examining Committee
for the thesis requirement for the Doctor of
Philosophy degree in Statistics at the July 2011
graduation.

Thesis Committee: _____
Dale L. Zimmerman, Thesis Supervisor

_____
Kung-Sik Chan

_____
Richard L. Dykstra

_____
Joseph B. Lang

_____
Joseph E. Cavanaugh

In memory of my paternal grandmother, Guifen Dong and
my maternal grandfather, Chaoming Liu.

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation to my major professor Dr. Dale L. Zimmerman for his inspiring guidance, constructive suggestions and enthusiastic encouragement during my graduate study. I am also very grateful to my committee members (alphabetically), Dr. Joe Cavanaugh, Dr. Kung-Sik Chan, Dr. Richard Dykstra, and Dr. Joseph B. Lang for their precious help.

More specifically, I appreciate Dr. Zimmerman for his guidence in antedependence (Markov) models methodology on longitudinal data, Dr. Cavanaugh and Dr. Lang for their help in categorical data analysis, and Dr. Chan and Dr. Dykstra for their help in probability and statistical inference.

I am deeply appreciative of all the professors in the department for their excellent teaching and the staff for their kind assistance.

ABSTRACT

Antedependence (AD) of order p, also known as the Markov property of order p, is a property of index-ordered random variables in which each variable, given at least p immediately preceding variables, is independent of all further preceding variables. Zimmerman and Núñez-Antón (2010) present statistical methodology for fitting and performing inference for AD models for continuous (primarily normal) longitudinal data. But analogous AD-model methodology for categorical longitudinal data has not yet been well developed. In this thesis, we derive maximum likelihood estimators of transition probabilities under antedependence of any order, and we use these estimators to develop likelihood-based methods for determining the order of antedependence of categorical longitudinal data. Specifically, we develop a penalized likelihood method for determining variable-order antedependence structure, and we derive the likelihood ratio test, score test, Wald test and an adaptation of Fisher's exact test for p^{th}-order antedependence against the unstructured (saturated) multinomial model. Simulation studies show that the score (Pearson's Chi-square) test performs better than all the other methods for complete and monotone missing data, while the likelihood ratio test is applicable for data with arbitrary missing pattern. But since the likelihood ratio test is oversensitive under the null hypothesis, we modify it by equating the expectation of the test statistic to its degrees of freedom so that it has actual size closer to nominal size. Additionally, we modify the likelihood ratio tests for use in testing for p^{th}-order antedependence against q^{th}-order antedependence, where $q > p$, and for testing nested variable-order antedependence models. We extend the methods to deal with data having a monotone or arbitrary missing pattern. For antedependence models of constant order

p, we develop methods for testing transition probability stationarity and strict stationarity and for maximum likelihood estimation of parametric generalized linear models that are transition probability stationary AD(p) models. The methods are illustrated using three data sets.

KEY WORDS: Antedependence; Categorical longitudinal data; Wald test; Score test; Likelihood ratio test; Penalized likelihood; Monotone missing (or monotone drop-ins); EM algorithm.

TABLE OF CONTENTS

LIST OF TABLES viii

LIST OF FIGURES x

CHAPTER

1 INTRODUCTION 1

1.1 Antedependence (Markov) model 1
1.2 Literature review 2
1.3 Overview 6

2 MAXIMUM LIKELIHOOD ESTIMATION 7

2.1 Maximum likelihood estimation of transition probabilities under given AD order 7
2.2 Maximum likelihood estimation of transition probabilities under two types of stationarity given AD order 20

3 MODEL SELECTION USING PENALIZED LOG-LIKELIHOOD . 32

3.1 Order selection 32

4 HYPOTHESIS TESTS FOR THE ORDER OF ANTEDEPENDENCE 36

4.1 AD(p) versus AD($n-1$) 36
4.1.1 Score test 36
4.1.2 Likelihood ratio test and its modification 38
4.1.3 Wald test 41
4.1.4 Adaptation of Freeman and Halton's exact test 48
4.1.5 Simulation study 50
4.2 AD(p) versus AD(q) for $0 \leq p < q \leq n-2$ 56
4.3 Nested variable-order AD models 59
4.4 Homogeneity in distribution of several groups 61

5 STATIONARITY UNDER AD(p) MODEL 65

5.1 Time-invariant transition probabilities under AD(p) for $1 \leq p \leq n-2$ . . . 65
5.1.1 Likelihood ratio and score tests . . . 65
5.1.2 Simulation . . . 66
5.1.3 Parametric generalized linear model stationary AD(p) structure . . . 68
5.2 Strict stationarity . . . 71
5.2.1 Likelihood ratio and score tests . . . 71
5.2.2 Simulation . . . 72

6 EXAMPLES . . . 75

6.1 Labor force data . . . 75
6.2 Wheeze data . . . 80
6.3 Toenail infection data . . . 82

7 CONCLUSION and DISCUSSION . . . 89

7.1 Conclusion with flowchart . . . 89
7.1.1 Flowchart . . . 90
7.1.2 Comparison of the tests . . . 91
7.1.3 Extension to multivariate categorical longitudinal data . 93
7.2 Discussion and open questions . . . 93

REFERENCES . . . 95

LIST OF TABLES

Table

2.1 Complete binary longitudinal data observed at three time points . . 8

2.2 Toy example for EM algorithm with missingness 16

2.3 Toy example for EM algorithm with Y_1 completed 19

2.4 Toy example for EM algorithm with Y_1 and Y_2 completed 20

4.1 Toy example for Wald test 45

4.2 Table 4.1 partitioned into two 2×2 tables for different values of Y_2 49

4.3 Rejection rates by Triad (Wald, likelihood ratio and score tests) . . 53

4.4 Rejection rates by modified likelihood ratio test (LRT1) 54

5.1 Empirical rejection rates for tests of transition stationarity for (5.2) 67

5.2 Empirical rejection rates for tests of two types of stationarity for (5.5) 74

6.1 Labor Force Data 76

6.2 P-values for testing for order of antedependence of the labor force data 78

6.3 P-values for testing for stationarity under AD(3) for labor force data 79

6.4 Stationary transition probabilities under AD(3) for labor force data 79

6.5 Link selection for AR(3) in the labor force data 79

6.6 Wheeze data 80

6.7 P-values for testing for order of antedependence of the Wheeze data 81

6.8 MLE of transition probabilities of the Wheeze data under AD(3) . . 81

6.9 Toenail data by treatment A 85

6.10 Toenail data by treatment B 86

6.11 Order selection by penalized likelihood criteria in the toenail data . 87

6.12 P-values for order selection by likelihood ratio test for the toenail data 87

6.13 MLE of transition probabilities of the toenail data under AD(1) . . 88

7.1 Comparison among triad for testing AD order 92

7.2 Comparison among triad for testing stationarity under AD(p) . . . 92

LIST OF FIGURES

Figure

4.1 Empirical rejection rate curves for (4.13), (4.14) and (4.15) 55

5.1 Empirical rejection rate curves for (5.2) 68

CHAPTER 1
INTRODUCTION

1.1 Antedependence (Markov) model

Longitudinal data are ubiquitous in applied scientific research, hence a huge statistical literature exists on models and methods for their analysis. Modern parametric models for longitudinal data are of three main types (Diggle et al., 2002): marginal, random-effects, and antedependence (also called Markov or transition) models. This article is concerned with models of the third type, by which the conditional distribution of the response variable at any time, given values of the response in the (recent) past and values of explanatory variables in the present and (recent) past, is modeled in terms of the quantities conditioned on. Specifically, index-ordered random variables $Y_1, \ldots, Y_n$ are said to be *antedependent of (variable) order* $(p_1, p_2, \ldots, p_n)$, or $\mathrm{AD}(p_1, p_2, \ldots, p_n)$, if Y_k, given at least p_k immediately preceding variables, is independent of all further preceding variables for $k = 1, 2, \ldots, n$ (Gabriel 1962, Macchiavelli and Arnold 1994). Note that $0 \leq p_k \leq k-1$ necessarily, and that $\mathrm{AD}(p_1, p_2, \ldots, p_n)$ variables are partially nested in the sense that

$$\mathrm{AD}(p_1, \ldots, p_n) \subset \mathrm{AD}(p_1 + q_1, \ldots, p_n + q_n)$$

if $q_k \geq 0$ for all k. The special case for which $p_k = \min(k-1, p)$ is known as pth-order antedependence and is denoted more concisely as $\mathrm{AD}(p)$. $\mathrm{AD}(p)$ variables are completely nested: that is,

$$\mathrm{AD}(0) \subset \mathrm{AD}(1) \subset \cdots \subset \mathrm{AD}(n-1),$$

with AD(0) being equivalent to mutual independence and AD($n-1$) being equivalent to completely general dependence (or a saturated model in the terminology of categorical data analysis).

1.2 Literature review

In this thesis, we consider likelihood-based inference procedures for antedependence models for categorical longitudinal data under multinomial sampling. Statistical methods for the analysis of antedependence models for continuous (primarily normal) longitudinal data are already well-developed; see Zimmerman and Núñez-Antón (2010) for a summary. Our main objective here is to develop categorical-data analogues for some of these methods, such as maximum likelihood estimation of transition probabilities under arbitrary order of antedependence and stationary transition probabilities under constant order of antedependence for complete and monotone missing data; penalized likelihood criteria to determine variable order of antedependence; hypothesis tests for determining constant order of antedependence; a modification to the likelihood ratio test that makes its empirical size agree more closely with its nominal size; parametric generalized linear model for autoregressive model of order p, AR(p) [transition probability stationary under nonsaturated model AD(p)] by maximum likelihood estimation; and an EM algorithm to deal with data with an arbitrary missing pattern. Moreover, we introduce some methods particular to categorical longitudinal data. For example, for continuous longitudinal data, constant variances and time-shift invariant correlations indicate weak stationarity, which implies strict stationarity for normal data. In contrast, Heagerty and Zeger (1998) pointed out the shortcomings of describing dependence in categorical data by correlations and recommended using log odds ratios for this purpose. Similarly,

we develop methods for describing dependence in categorical longitudinal data by conditional log-odds ratios instead of conditional correlations.

In recent years, considerable research has been devoted to the development of "structured" transition models for categorical longitudinal data, i.e. models that impose a parametric structure upon the transition probabilities or some transform of them. A general form for such a model is

$$g(\mu_{ik}) \equiv g(E(Y_{ik}|\mathcal{F}_{k-1})) = \boldsymbol{\beta}'\mathbf{Y}_{i,k-1} + \boldsymbol{\gamma}' f_{ik}(\mathbf{X}_{i,k-1}),\ k = p+1, \ldots, n, \tag{1.1}$$

g is link function, $\mathcal{F}_{k-1}$ represents all that is known to the observer up to and including time $k-1$ about the response and the covariate information, Y_{ik} is the k-th component of the i-th subject's categorical response vector $\mathbf{Y}_i$, $\mathbf{X}_i$ is the collection of all covariates, and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are column vectors of parameters. Note that (1.1) is given as the form of a generalized additive Markov model and it will turn out to be a generalized linear model when the f_{ik}'s are identity functions. If the Markov model is of order p, then

$$\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_p] \text{ and } \mathbf{Y}_i = [1, Y_{i,k-1}, \ldots, Y_{i,k-p}]'.$$

Cox and Snell (1989) introduced Markov models for binary time series data, where

$$E(Y_{ik}|\mathcal{F}_{k-1}) = P(Y_{ik} = 1|\mathcal{F}_{k-1})$$

and the link function g can be any of the following:

$$\text{logit} : g(z) = \log\left(\frac{z}{1-z}\right);$$

$$\text{probit} : g(z) = \boldsymbol{\Phi}^{-1}(z);$$

$$\text{log-log} : g(z) = -\log(-\log(z)); \text{ and}$$

$$\text{complementary log-log} : g(z) = \log(-\log(1-z));$$

where $\boldsymbol{\Phi}$ is the cumulative distribution function of the standard normal distribution. Denote $v_{ik} \equiv \text{var}(Y_{ik}|\mathcal{F}_{k-1})$. Zeger and Qaqish (1988) introduced a quasi-likelihood

(QL) approach to estimate parameter $\boldsymbol{\beta}$ by solving the estimating equation

$$U(\boldsymbol{\beta}) \equiv \sum_{k=1}^{n} \frac{\partial \mu_{ik}}{\partial \boldsymbol{\beta}} v_{ik}^{-1}(Y_{ik} - \mu_{ik}) = 0$$

using iteratively reweighted least squares. Heagerty and Zeger (2000) separated the Markov model into two parts, with the first part being a marginal mean model directly specifying the population-averaged effect of covariates on the responses and the second part being a conditional model describing serial dependence and identifying the joint distribution of the responses but specifying the dependence on covariates only implicitly and reparametrized the version of the model called "marginalized transition model (MTM)". In particular, for binary data, based on the early work by Azzalini (1994), Heagerty (2002) proposed the model labelled as MTM(p):

$$\begin{aligned} \log(\frac{\mu_{ik}^M}{1-\mu_{ik}^M}) &= \boldsymbol{\gamma}'\mathbf{X}_{ik},\ k = 1, \ldots, n \\ \log(\frac{\mu_{ik}^C}{1-\mu_{ik}^C}) &= \triangle_{ik} + \sum_{h=1}^{p} \phi_{ikh} Y_{i,k-h},\ k = p+1, \ldots, n \\ \phi_{ikh} &= \mathbf{z}_{ikh}'\eta_h\ ,\ k = p+1, \ldots, n. \end{aligned} \tag{1.2}$$

In model 1.2, superscript M in $\mu_{ik}^M \equiv E(Y_{ik}|X_{ik})$ refers to marginal, $\triangle_{ik}$ is an intercept parameter, ϕ_{ikh} is a subject-specific coefficient, $\mathbf{z}_{ikh}$ is a vector of covariates on subject i which are a subset of the covariates in X_{ik} and η_h is a parameter vector. Lee and Daniels (2007) extended the work by Heagerty (2002) to accommodate longitudinal ordinal data and developed Fisher-scoring algorithms for estimation. However, under all of these models the order of antedependence is time-invariant, as are the transition probabilities. In Chapter 5, we introduce how to fit the generalized linear model by maximum likelihood method for our special case of categorical longitudinal data without covariates when the assumption of stationary transition probability under constant order of antedependence is satisfied.

As for the determination of order of antedependence and testing for transition probability stationarity without covariates by likelihood-based methods, some relevant early work was performed by Anderson and Goodman (1957). By assuming complete data without empty cells, they derived maximum likelihood estimators (mles) for nonstationary transition probabilities of a first-order Markov process, for stationary transition probability first-order Markov process, for Markov process of higher constant order and for Markov process with bivariate response based on complete data with nonempty cells and considered some related testing problems by likelihood ratio test and score (Pearson's Chi-square) test. However, the fundamental assumption for order selection by hypothesis testing that the order of the Markov process is constant across time may not always be satisfied, since among all $n!$ possible variable-order models, one is not necessarily nested in another, which makes it inappropriate to do the initial order selection by hypothesis testing. In this thesis, we extend the methods to antedependence models of arbitrary variable order and to data that are incomplete or have empty cells, and we consider several additional inference problems for these models including parametric generalized linear model fitting for the stationary transition probability AD(p) model. The methods presented here may be useful at the initial stages of model formulation for categorical longitudinal data. In particular, we give methods for identifying the (variable) order of antedependence and, if the order is determined to be time-invariant, identifying various stationarity properties of the process for categorical longitudinal data without covariates, so that further inferences may be based on appropriate structured transition models.

1.3 Overview

The remainder of this thesis is organized as follows. In Chapter 2, we derive closed-form expressions for mles of multinomial transition probabilities under an antedependence model of arbitrary order, based on complete or monotone missing data. We also describe how the EM algorithm may be used to obtain mles from data with an arbitrary pattern of missingness, and we derive mles under constant-order antedependence models with two different stationarity properties. Chapters 3 and 4 describe model identification procedures for antedependence models: penalized likelihood criteria for model selection (Chapter 3) and likelihood-based (likelihood ratio, score, and Wald) tests for various hypotheses of interest (Chapter 4). Chapter 4 also includes a simulation study comparing the performance of the likelihood-based tests for pth-order antedependence against the saturated alternative. Chapter 5 gives likelihood-based tests for two stationarity properties under constant-order antedependence and discusses fitting a parametric generalized linear model for AR(p) by maximum likelihood estimation. Three examples are presented in Chapter 6. Chapter 7 contains a brief conclusion with a flowchart describing the methods introduced in this thesis and a discussion for open questions.

CHAPTER 2
MAXIMUM LIKELIHOOD ESTIMATION

2.1 Maximum likelihood estimation of transition probabilities under given AD order

Suppose that repeated observations of a categorical (nominal or ordinal) characteristic are taken over time on N subjects. Let $n \geq 2$ denote the number of measurement times and let $1, \ldots, c$ denote the categories of the characteristic (which are assumed not to change over time), where $c \geq 2$, although binary outcomes are commonly coded as 1 and 0, as is used in this thesis. Hence, if no observations are missing, the observational vector $\mathbf{Y}_i \equiv (Y_{i1}, \ldots, Y_{in})'$ for the ith subject has c^n possible outcomes. Let Y_k denote the observation at time point k for a generic subject. For each possible outcome $(y_1, \ldots, y_n)$, let

$$\pi_{y_1 \ldots y_n} \equiv P(Y_1 = y_1, \ldots, Y_n = y_n)$$

denote the true cell probability with corresponding observed cell count $N_{y_1 \cdots y_n}$, and put $\boldsymbol{\pi} = (\pi_{y_1 \ldots y_n})$. Accordingly,

$$N = \sum_{(y_1, \ldots, y_n) \in C_n} N_{y_1 \cdots y_n}$$

where $C_n \equiv \{1, \ldots, c\}^n$ is the set of all c^n possible outcomes. Unless noted otherwise, we assume that the $\mathbf{Y}_i$'s are independently and identically distributed as Multinomial$(N, \boldsymbol{\pi})$ and that covariates are either unavailable or not used in the analysis. To clarify the notation, an example of complete binary longitudinal data observed at three time points is depicted in Table 2.1, where $n = 3$ and $c = 2$.

Y_1	Y_2	Y_3	count	$\boldsymbol{\pi}$
1	1	1	N_{111}	π_{111}
1	1	0	N_{110}	π_{110}
1	0	1	N_{101}	π_{101}
1	0	0	N_{100}	π_{100}
0	1	1	N_{011}	π_{011}
0	1	0	N_{010}	π_{010}
0	0	1	N_{001}	π_{001}
0	0	0	N_{000}	π_{000}

Table 2.1: Complete binary longitudinal data observed at three time points

Since antedependence is defined in terms of certain conditional independencies, it is convenient to reparameterize in terms of certain conditional probabilities. Define

$$\pi_{y_k|y_1\cdots y_{k-1}} \equiv P(Y_k = y_k|Y_1 = y_1, \ldots, Y_{k-1} = y_{k-1})$$

for $k = 2, \ldots, n$ and $(y_1, \ldots, y_k) \in C_k$. It is easily verified that the mapping from the nonredundant cell-probability parameterization

$$\Theta_1 \equiv \left\{\pi_{y_1\cdots y_n} : (y_1, \ldots, y_n) \in C_n \setminus \{c, \ldots, c\}\right\}$$

to the nonredundant "sequential conditional probability" parameterization

$$\begin{aligned}\Theta_2 \equiv &\{\pi_{y_1+\cdots+} : y_1 = 1, \ldots, c-1\}\cup \\ &\{\pi_{y_k|y_1\cdots y_{k-1}} : k = 2, \ldots, n;\ y_k = 1, \ldots, c-1;\ (y_1, \ldots, y_{k-1}) \in C_{k-1}\}\end{aligned}$$

is one-to-one. (Here and subsequently, we indicate summation over a subscripted index by replacing that index with a "+.") For example,

$$\pi_{y_k|y_1\cdots y_{k-1}} = P(Y_k = y_k|Y_1 = y_1, \ldots, Y_{k-1} = y_{k-1}) = \frac{\pi_{y_1\cdots y_{k-1}y_k+\cdots+}}{\pi_{y_1\cdots y_{k-1}+\cdots+}} \tag{2.1}$$

(provided the denominator is positive) and

$$\pi_{y_1\cdots y_n} = \pi_{y_1+\cdots+}\prod_{k=2}^{n}\pi_{y_k|y_1\cdots y_{k-1}}.$$

Moreover, under an AD$(p_1,\ldots,p_n)$ model, for each k such that $p_k \geq 1$ and $k-p_k \geq 2$ and each fixed $(y_{k-p_k},\ldots,y_{k-1}) \in C_{p_k}$, the elements of

$$\{\pi_{y_k|y_1\cdots y_{k-p_k-1}y_{k-p_k}\cdots y_{k-1}} : (y_1,\ldots,y_{k-p_k-1}) \in C_{k-p_k-1}\} \text{ are equal;} \tag{2.2}$$

hence we may represent their common value by a transition probability parameter $\pi_{y_k|y_{k-p_k}\cdots y_{k-1}}$. Thus, the AD$(p_1,\ldots,p_n)$ model may be parameterized by the nonredundant set of parameters

$$\begin{aligned}\Theta^{(p_1\cdots p_n)} \equiv & \bigcup_{k\ni p_k=0}\{\pi_{+\cdots+y_k+\cdots+} : y_k = 1,\ldots,c-1\} \\ & \bigcup\bigcup_{k\ni p_k\geq 1}\{\pi_{y_k|y_{k-p_k}\cdots y_{k-1}} : y_k = 1,\ldots,c-1;\ (y_{k-p_k},\ldots,y_{k-1}) \in C_{p_k}\},\end{aligned}$$

which we call the *transition-probability parameterization*. It is easily verified that

$$dim(\Theta^{(p_1\cdots p_n)}) = (c-1)\sum_{k=1}^{n} c^{p_k}. \tag{2.3}$$

In what follows, we give several results pertaining to maximum likelihood estimation of the transition-probability parameterization of an AD$(p_1,\ldots,p_n)$ process.

Theorem 2.1.1. *Under AD$(p_1,p_2,\ldots,p_n)$, the complete-data mles of the parameters of $\Theta^{(p_1\cdots p_n)}$ are as follows: for k such that $p_k = 0$, $\hat{\pi}^{(p_1\cdots p_n)}_{+\cdots+y_k+\cdots+} = \dfrac{N_{+\cdots+y_k+\cdots+}}{N}$; for other k,*

$$\hat{\pi}^{(p_1\cdots p_n)}_{y_k|y_{k-p_k}\cdots y_{k-1}} = \begin{cases} 0 & \text{if } N_{+\cdots+y_{k-p_k}\cdots y_{k-1}+\cdots+} = 0, \\ \dfrac{N_{+\cdots+y_{k-p_k}\cdots y_k+\cdots+}}{N_{+\cdots+y_{k-p_k}\cdots y_{k-1}+\cdots+}} & \text{otherwise.}\end{cases}$$

Proof. We start the proof by parameterization Θ_1 and transform it to Θ_2. The

likelihood function is proportional to

$$\prod_{(y_1,\ldots,y_n)\in C_n} (\pi_{y_1\cdots y_n})^{N_{y_1\cdots y_n}}$$

$$= \prod_{(y_1,\ldots,y_n)\in C_n} \left(\pi_{y_1+\cdots+}\prod_{k=2}^{n}\pi_{y_k|y_1\cdots y_{k-1}}\right)^{N_{y_1\cdots y_n}} \tag{2.4}$$

$$= \prod_{(y_1,\ldots,y_n)\in C_n} \left(\prod_{k=1}^{n}\left[I(p_k=0)\pi_{+\cdots+y_k+\cdots+} + I(p_k\geq 1)\pi_{y_k|y_{k-p_k}\cdots y_{k-1}}\right]\right)^{N_{y_1\cdots y_n}} \tag{2.5}$$

$$= \prod_{k=1}^{n}\left[\left(I(p_k=0)\prod_{y_k=1}^{c}\pi_{+\cdots+y_k+\cdots+}^{N_{+\cdots+y_k+\cdots+}}\right) + \left(I(p_k\geq 1)\prod_{(y_{k-p_k},\ldots,y_k)\in C_{p_k+1}}\pi_{y_k|y_{k-p_k}\cdots y_{k-1}}^{N_{+\cdots+y_{k-p_k}\cdots y_k+\cdots+}}\right)\right]. \tag{2.6}$$

The equality between (2.5) and (2.6) holds because $I(p_k = 0)I(p_k \geq 1) = 0$ for all k. For k such that $p_k = 0$, the kth term of the outermost product in (2.6) is the kernel of the likelihood of a saturated c-nomial distribution with cell probabilities $\{\pi_{+\cdots+y_k+\cdots+} : y_k = 1,\ldots,c\}$; for other k, the kth term is the product of c^{p_k} independent likelihood kernels, each corresponding to a saturated c-nomial distribution with cell probabilities $\{\pi_{y_k|y_{k-p_k}\cdots y_{k-1}} : y_k = 1,\ldots,c\}$. The cell probabilities for each kernel sum to one and lie within $[0,1)$, but are not otherwise constrained under the AD$(p_1,\ldots,p_n)$ model. Thus for those k such that $p_k = 0$, $\hat{\pi}^{(p_1\cdots p_n)}_{+\cdots+y_k+\cdots+} = \dfrac{N_{+\cdots+y_k+\cdots+}}{N}$; for other k, if $N_{+\cdots+y_{k-p_k}\cdots y_{k-1}+\cdots+} = 0$, we have $N_{+\cdots+y_{k-p_k}\cdots y_k+\cdots+} = 0$, implying $\hat{\pi}^{(p_1\cdots p_n)}_{y_k|y_{k-p_k}\cdots y_{k-1}} = 0$, (for a saturated multinomial distribution the mle of cell probability for the event with empty cell is well known to be zero) and if $N_{+\cdots+y_{k-p_k}\cdots y_{k-1}+\cdots+} \neq 0$, we have

$$\hat{\pi}^{(p_1\cdots p_n)}_{y_k|y_{k-p_k}\cdots y_{k-1}} = \frac{N_{+\cdots+y_{k-p_k}\cdots y_k+\cdots+}}{N_{+\cdots+y_{k-p_k}\cdots y_{k-1}+\cdots+}}. \tag{2.7}$$

□

Upon substituting $\min(k-1,p)$ for p_k $(k = 1,\ldots,n)$ in Theorem 2.1.1, we

realize that the parameter space $\Theta^{(p_1\cdots p_n)}$ simplifies to

$$\Theta^{(p)} \equiv \bigcup \{\pi_{y_1\cdots y_p+\cdots+} : (y_1,\ldots,y_p) \in C_p\}$$
$$\bigcup \bigcup_{k=\{p+1,\ldots,n\}} \{\pi_{y_{p+1}^{(k)}|y_1^{(k-p)}\cdots y_p^{(k-1)}} : y_{p+1} = 1,\ldots,c-1;\ (y_1,\ldots,y_p) \in C_p^+\},$$

where $\pi_{y_{p+1}^{(k)}|y_1^{(k-p)}\cdots y_p^{(k-1)}} \equiv P(Y_k = y_{p+1}|Y_{k-p} = y_1, y_{k-p+1} = y_2,\ldots,Y_{k-1} = y_p)$ and we obtain the following corollary.

Corollary 2.1.2. *Under AD(p), the complete-data mles of parameters of* $\Theta^{(p)}$ *are as follows: if* $p = 0$, $\hat{\pi}^{(p)}_{+\cdots+y_k+\cdots+} = \dfrac{N_{+\cdots+y_k+\cdots+}}{N}$ *for* $k = 1,\ldots,n$; *if* $p \geq 1$, $\hat{\pi}^{(p)}_{y_1\cdots y_p+\cdots+} = \dfrac{N_{y_1\cdots y_p+\cdots+}}{N}$ *and for* $k \geq p+1$

$$\hat{\pi}^{(p)}_{y_k|y_{k-p}\cdots y_{k-1}} = \begin{cases} 0 & \text{if } N_{+\cdots+y_{k-p}\cdots y_{k-1}+\cdots+} = 0 \\ \dfrac{N_{+\cdots+y_{k-p}\cdots y_k+\cdots+}}{N_{+\cdots+y_{k-p}\cdots y_{k-1}+\cdots+}} & \text{otherwise.} \end{cases}$$

Theorem 2.1.1 and Corollary 2.1.2 can be extended easily to handle ignorable monotone missing data ("dropouts"), defined by the condition that $Y_{i,k+1}$ is missing whenever $Y_{i,k}$ is missing ($i = 1,\ldots,N;\ k = 2,\ldots,n-1$). Let $N^{\bullet(k)}$ be the number of subjects having complete observations between time points 1 and k (inclusive), and let $N^{\bullet(k)}_{+\cdots+y_{k-p_k}\cdots y_k+\cdots+}$ be the number of these subjects for which $Y_{k-p_k} = y_{k-p_k},\ldots,Y_k = y_k$, regardless of whether $Y_{k+1},\ldots,\ Y_n$ are observed or missing. Similarly, $N^{\bullet(k)}_{+\cdots+y_k+\cdots+}$ is that for which $Y_k = y_k$ and $N^{\bullet(k)}_{+\cdots+y_{k-p_k}\cdots y_{k-1}+\cdots+}$ is that for which $Y_{k-p_k} = y_{k-p_k},\ldots,Y_{k-1} = y_{k-1}$, regardless of whether the responses at all the other time points indicated by "+" are observed or missing.

Theorem 2.1.3. *Under AD($p_1, p_2,\ldots,p_n$), the monotone-missing-data mles of the parameters of* $\Theta^{(p_1\cdots p_n)}$ *(assuming ignorability), denoted by* $\hat{\pi}^{\bullet(p_1\cdots p_n)}_{+\cdots+y_k+\cdots+}$ *and* $\hat{\pi}^{\bullet(p_1\cdots p_n)}_{y_k|y_{k-p_k}\cdots y_{k-1}}$, *are given by expressions identical to those in Theorem 2.1.1 except that* $N^{\bullet(k)}$,

$N^{\bullet(k)}_{+\cdots+y_k+\cdots+}$, $N^{\bullet(k)}_{+\cdots+y_{k-p_k}\cdots y_k+\cdots+}$, *and* $N^{\bullet(k)}_{+\cdots+y_{k-p_k}\cdots y_{k-1}+\cdots+}$ *are substituted for the corresponding complete-data counts; thus*

$$\hat{\pi}^{\bullet(p_1\cdots p_n)}_{y_k|y_{k-p_k}\cdots y_{k-1}} = \begin{cases} 0 & \text{if } N^{\bullet(k)}_{+\cdots+y_{k-p_k}\cdots y_{k-1}+\cdots+} = 0, \\ \dfrac{N^{\bullet(k)}_{+\cdots+y_{k-p_k}\cdots y_k+\cdots+}}{N^{\bullet(k)}_{+\cdots+y_{k-p_k}\cdots y_{k-1}+\cdots+}} & \text{otherwise.} \end{cases} \tag{2.8}$$

Under AD(p), the monotone-missing-data mles of the parameters of $\Theta^{(p)}$ *are given by substituting the analogous quantities into Corollary 2.1.2; thus*

$$\hat{\pi}^{\bullet(p)}_{y_k|y_{k-p}\cdots y_{k-1}} = \begin{cases} 0 & \text{if } N^{\bullet(k)}_{+\cdots+y_{k-p}\cdots y_{k-1}+\cdots+} = 0, \\ \dfrac{N^{\bullet(k)}_{+\cdots+y_{k-p}\cdots y_k+\cdots+}}{N^{\bullet(k)}_{+\cdots+y_{k-p}\cdots y_{k-1}+\cdots+}} & \text{otherwise.} \end{cases} \tag{2.9}$$

Proof. For ignorable monotone missing data, it is easily verified that the kernel of the likelihood function is of exactly the same form as (2.6), except that $N^{\bullet(k)}_{+\cdots+y_k+\cdots+}$ and $N^{\bullet(k)}_{+\cdots+y_{k-p_k}\cdots y_k+\cdots+}$ appear in place of $N_{+\cdots+y_k+\cdots+}$ and $N_{+\cdots+y_{k-p_k}\cdots y_k+\cdots+}$, respectively. More specifically, a straightforward extension of (2.6) to monotone missing data is

$$\prod_{k=1}^{n}\left[\left(I(p_k=0)\prod_{y_k=1}^{c}\pi_{+\cdots+y_k+\cdots+}^{N^{\bullet(k)}_{+\cdots+y_k+\cdots+}}\right)+\left(I(p_k\geq 1)\prod_{(y_{k-p_k},\ldots,y_k)\in C_{p_k+1}}\pi_{y_k|y_{k-p_k}\cdots y_{k-1}}^{N^{\bullet(k)}_{+\cdots+y_{k-p_k}\cdots y_{k-1}y_k+\cdots+}}\right)\right]. \tag{2.10}$$

The result follows by the same arguments as those used in the proof of Theorem 2.1.1. □

Mles under AD(p) may also be obtained easily for ignorable missing data with monotone drop-ins (also known as delayed or staggered entry), defined by the condition that $Y_{i,k+1}$ is observed whenever $Y_{i,k}$ is observed ($i = 1, \ldots, N;\ k = 2, \ldots, n-1$). For such data, mles are as given by Theorem 2.1.3 but applied to the data in reverse time order. This follows from the fact that pth-order antedependent random variables are also pth-order antedependent when arranged in reverse time order (Zimmerman and Núñez-Antón 2010, p. 151). Mathematically, we can convert

monotone drop-in data into monotone missing data by premultiplying the matrix $\mathbf{Y}$ by the exchange matrix

$$\mathbf{E}_s \equiv \begin{bmatrix} 0 & \cdots & 0 & 1 \\ 0 & \cdots & 1 & 0 \\ \vdots & & \vdots & \vdots \\ 1 & \cdots & 0 & 0 \end{bmatrix}.$$

But there is not an analogous result for variable-order antedependent random variables.

Note that (2.6) is a product of kernels of saturated multinomial distributions. Thus for ignorable missing data with an arbitrary pattern of missingness, the EM algorithm (Dempster, Laird, and Rubin, 1977) may be used to obtain mles of cell probabilities under an AD$(p_1, \ldots, p_n)$ model. Schafer (1999, Sec. 7.3) described the use of the EM algorithm for estimation in the saturated multinomial model, while we apply the EM algorithm and do count completion alternately and chronologically. For this purpose, we define the following notations: for $k = 2, \ldots, n-1$, $\hat{N}^{\bullet(k-1)}_{+\cdots+y_k+\cdots+}$ and $\hat{N}^{\bullet(k-1)}_{+\cdots+y_{k-p_k}\cdots y_k+\cdots+}$ are the maximum likelihood estimated counts of subjects having realizations y_k at time point k and $y_{k-p_k} \cdots y_k$ from time point $k - p_k$ to time point k, respectively, regardless of the realizations (missing or observed) at all the other time points after count completion through time point $k - 1$; $\hat{N}^{\bullet(k-1)}_{+\cdots+\bullet+\cdots+}$ and $\hat{N}^{\bullet(k-1)}_{+\cdots+y_{k-p_k}\cdots y_{k-1}\bullet+\cdots+}$ are the maximum likelihood estimated counts of subjects having realizations missing at time point k and $y_{k-p_k} \cdots y_{k-1}$ from time point $k-p_k$ to time point $k - 1$ and missing at time point k, respectively, regardless of the realizations (missing or observed) at all the other time points after count completion through time point $k - 1$. When $k = 1$, $\hat{N}^{\bullet(0)}_{y_1+\cdots+} \equiv N_{y_1+\cdots+}$ and $\hat{N}^{\bullet(0)}_{\bullet+\cdots+} \equiv N_{\bullet+\cdots+}$. We describe the procedure in Theorem 2.1.4.

Theorem 2.1.4. *Under AD($p_1, \ldots, p_n$), for data with an arbitrary missingness pattern, for time points $k = 1, \cdots, n-1$, we apply the EM algorithm to obtain the mle of transition probability and complete the counts at this time point after the algorithm converges. More specifically, for $k = 1, \ldots, n-1$, the iteration of EM algorithm can be expressed as follows: if $p_k = 0$, then*

$$\hat{\pi}^{(p_1\cdots p_n)(j+1)}_{+\cdots+y_k+\cdots+} = \frac{\hat{N}^{\bullet(k-1)}_{+\cdots+y_k+\cdots+} + \hat{\pi}^{(p_1\cdots p_n)(j)}_{+\cdots+y_k+\cdots+}\hat{N}^{\bullet(k-1)}_{+\cdots+\bullet+\cdots+}}{N} \tag{2.11}$$

where j stands for the step of iteration;

if $p_k \geq 1$, then

$$\hat{\pi}^{(p_1\cdots p_n)(j+1)}_{y_k|y_{k-p_k}\cdots y_{k-1}} = \frac{\hat{N}^{\bullet(k-1)}_{+\cdots+y_{k-p_k}\cdots y_k+\cdots+} + \hat{\pi}^{(p_1\cdots p_n)(j)}_{y_k|y_{k-p_k}\cdots y_{k-1}}\hat{N}^{\bullet(k-1)}_{+\cdots+y_{k-p_k}\cdots y_{k-1}\bullet+\cdots+}}{\hat{N}^{\bullet(k-1)}_{+\cdots+y_{k-p_k}\cdots y_{k-1}+\cdots+}}, \tag{2.12}$$

when $\hat{N}^{\bullet(k-1)}_{+\cdots+y_{k-p_k}\cdots y_{k-1}+\cdots+} \neq 0$ and $\hat{\pi}^{(p_1\cdots p_n)(j+1)}_{y_k|y_{k-p_k}\cdots y_{k-1}} = 0$ when $\hat{N}^{\bullet(k-1)}_{+\cdots+y_{k-p_k}\cdots y_{k-1}+\cdots+} = 0$

When the EM algorithm converges, complete the counts at time k by

$$\begin{aligned}&\hat{N}^{\bullet(k)}_{+\cdots+y_{k-p_k}\cdots y_k+\cdots+}\\ =&\hat{N}^{\bullet(k-1)}_{+\cdots+y_{k-p_k}\cdots y_k+\cdots+} + \left(I(p_k=0)\hat{\pi}^{(p_1\cdots p_n)(\infty)}_{+\cdots+y_k+\cdots+} + I(p_k\geq 1)\hat{\pi}^{(p_1\cdots p_n)(\infty)}_{y_k|y_{k-p_k}\cdots y_{k-1}}\right)\hat{N}^{\bullet(k-1)}_{+\cdots+y_{k-p_k}\cdots y_{k-1}\bullet+\cdots+}\end{aligned} \tag{2.13}$$

Repeat the EM algorithm and count completion alternately for $k = 1, \ldots, n-1$ so that the counts are complete through time point $n-1$. Perform Theorem 2.1.1 if no data are missing at time point n and Theorem 2.1.3 if some data are missing at time point n to obtain $\hat{\pi}^{(p_1\cdots p_n)}_{y_n|y_{n-p_n}\cdots y_{n-1}}$ if $p_n \geq 1$ or $\hat{\pi}^{(p_1\cdots p_n)}_{+\cdots+y_n}$ if $p_n = 0$.

Proof. First we show the E-step of the EM algorithm. For $k = 1, \ldots, n-1$, after completing the counts at the first $k-1$ time points, if $p_k \geq 1$, for all the subjects whose observation at time point k, $\hat{N}^{\bullet(k-1)}_{+\cdots+y_{k-p_k}\cdots y_{k-1}\bullet+\cdots+}$, is missing, we proportionally assign $y_k = 1, \ldots, c$ according to

$$\text{Multinomial}\left(\hat{N}^{\bullet(k-1)}_{+\cdots+y_{k-p_k}\cdots y_{k-1}\bullet+\cdots+},\ \left(\pi^{(p_1\cdots p_n)}_{Y_k=1|y_{k-p_k}\cdots y_{k-1}}, \ldots, \pi^{(p_1\cdots p_n)}_{Y_k=c|y_{k-p_k}\cdots y_{k-1}}\right)\right).$$

Thus, by including $\hat{N}^{\bullet(k-1)}_{+\cdots+y_{k-p_k}\cdots y_k+\cdots+}$, the subjects whose realizations at time point

k are observed, we have

$$E(N_{+\cdots+y_{k-p_k}\cdots y_k+\cdots+}) = \hat{N}^{\bullet(k-1)}_{+\cdots+y_{k-p_k}\cdots y_k+\cdots+} + \hat{N}^{\bullet(k-1)}_{+\cdots+y_{k-p_k}\cdots y_{k-1}\bullet+\cdots+}\pi^{(p_1\cdots p_n)}_{y_k|y_{k-p_k}\cdots y_{k-1}}$$

By the invariance property of mle, the M-step is

$$\hat{\pi}^{(p_1\cdots p_n)}_{y_k|y_{k-p_k}\cdots y_{k-1}} = \frac{E(N_{+\cdots+y_{k-p_k}\cdots y_k+\cdots+})}{\hat{N}^{\bullet(k-1)}_{+\cdots+y_{k-p_k}\cdots y_{k-1}+\cdots+}}$$

By combining the two steps, we have the iteration (2.12). Similarly, when $p_k = 0$, we obtain the iteration (2.11). Also, by the invariance property of mle, we can complete the counts at time point k to yield (2.13). □

Next we show how to use Theorem 2.1.4 by a simple toy example. In Table 2.2, we created a toy example and for illustration purpose, we show the steps of obtaining mles of transition probabilities by the EM algorithm under an AD(1) model, which can be written as AD$(0, 1, 1)$. In this example, we observe binary longitudinal data at three time points. Part A stands for complete observations, while parts B, C, D, E, F and G stand for observations with missingness. Table 2.2 contains the complete data and data with all possible patterns of missingness.

Note that in this toy example, in order to distinguish different missing patterns, we use "$\bullet$" to denote missingness at that time point and "+" to denote summing over the index for the part of missing pattern indicated by the corresponding letter in the superscript. By (2.11), for the EM algorithm, we iterate

$$\hat{\pi}^{(0,1,1)(j+1)}_{Y_1=1} = \frac{\hat{N}^{\bullet(0)}_{1++} + \hat{\pi}^{(0,1,1)(j)}_{Y_1=1}\hat{N}^{\bullet(0)}_{\bullet++}}{N} = \frac{N_{1++} + \hat{\pi}^{(0,1,1)(j)}_{Y_1=1}N_{\bullet++}}{N};$$

until convergence. Let superscript (∞) denote the mle obtained when EM algorithm converges. Then

$$\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=0} = 1 - \hat{\pi}^{(0,1,1)(\infty)}_{Y_1=1}.$$

	Y_1	Y_2	Y_3	count
A (complete)	1	1	1	N^A_{111}
	1	1	0	N^A_{110}
	1	0	1	N^A_{101}
	1	0	0	N^A_{100}
	0	1	1	N^A_{011}
	0	1	0	N^A_{010}
	0	0	1	N^A_{001}
	0	0	0	N^A_{000}
B (only Y_1 missing)	$\bullet$	1	1	$N^B_{\bullet 11}$
	$\bullet$	1	0	$N^B_{\bullet 10}$
	$\bullet$	0	1	$N^B_{\bullet 01}$
	$\bullet$	0	0	$N^B_{\bullet 00}$
C (only Y_2 missing)	1	$\bullet$	1	$N^C_{1\bullet 1}$
	1	$\bullet$	0	$N^C_{1\bullet 0}$
	0	$\bullet$	1	$N^C_{0\bullet 1}$
	0	$\bullet$	0	$N^C_{0\bullet 0}$
D (only Y_3 missing)	1	1	$\bullet$	$N^D_{11\bullet}$
	1	0	$\bullet$	$N^D_{10\bullet}$
	0	1	$\bullet$	$N^D_{01\bullet}$
	0	0	$\bullet$	$N^D_{00\bullet}$
E (Y_1 and Y_2 missing)	$\bullet$	$\bullet$	1	$N^E_{\bullet\bullet 1}$
	$\bullet$	$\bullet$	0	$N^E_{\bullet\bullet 0}$
F (Y_1 and Y_3 missing)	$\bullet$	1	$\bullet$	$N^F_{\bullet 1\bullet}$
	$\bullet$	0	$\bullet$	$N^F_{\bullet 0\bullet}$
G (Y_2 and Y_3 missing)	1	$\bullet$	$\bullet$	$N^G_{1\bullet\bullet}$
	0	$\bullet$	$\bullet$	$N^G_{0\bullet\bullet}$

Table 2.2: Toy example for EM algorithm with missingness

Next we complete the counts for each data segment at time point $k = 1$ by

$$\hat{N}^{\bullet(1)}_{1++} = \hat{N}^{\bullet(0)}_{1++} + \hat{N}^{\bullet(0)}_{\bullet++}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=1} = N_{1++} + N_{\bullet++}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=1}$$

Similarly, we can obtain $\hat{N}^{\bullet(1)}_{0++}$. By partitioning the counts according to their patterns of missingness, we have the data with the first time point completed summarized in Table 2.3, where superscript 1 in A, C, D and G stands for completion of the first time point by EM algorithm. Now we use the EM algorithm to obtain $\hat{\pi}^{(0,1,1)}_{Y_2=1|Y_1=1}$ and $\hat{\pi}^{(0,1,1)}_{Y_2=1|Y_1=0}$. By (2.12), we have

$$\begin{aligned}
\hat{\pi}^{(0,1,1)(j+1)}_{Y_2=1|Y_1=1} &= \frac{\hat{N}^{\bullet(1)}_{11+} + \hat{\pi}^{(0,1,1)(j)}_{Y_2=1|Y_1=1}\hat{N}^{\bullet(1)}_{1\bullet+}}{\hat{N}^{\bullet(1)}_{1++}} \\
&= \Big[N^{A}_{11+} + N^{B}_{\bullet 1+}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=1} + (N^{C}_{1\bullet+} + N^{E}_{\bullet\bullet+}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=1})\hat{\pi}^{(0,1,1)(j)}_{Y_2=1|Y_1=1} + N^{D}_{11\bullet} \\
&\quad + N^{F}_{\bullet 1\bullet}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=1} + N^{G}_{1\bullet\bullet}\hat{\pi}^{(0,1,1)(j)}_{Y_2=1|Y_1=1}\Big] \Big/ \Big[N^{A}_{1++} + N^{B}_{\bullet++}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=1} + N^{C}_{1\bullet+} + N^{E}_{\bullet\bullet+}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=1} \\
&\quad + N^{D}_{1+\bullet} + N^{F}_{\bullet+\bullet}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=1} + N^{G}_{1\bullet\bullet}\Big],
\end{aligned}$$

and similarly we have

$$\begin{aligned}
\hat{\pi}^{(0,1,1)(j+1)}_{Y_2=1|Y_1=0} &= \frac{\hat{N}^{\bullet(1)}_{01+} + \hat{\pi}^{(0,1,1)(j)}_{Y_2=1|Y_1=0}\hat{N}^{\bullet(1)}_{0\bullet+}}{\hat{N}^{\bullet(1)}_{0++}} \\
&= \Big[N^{A}_{01+} + N^{B}_{\bullet 1+}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=0} + (N^{C}_{0\bullet+} + N^{E}_{\bullet\bullet+}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=0})\hat{\pi}^{(0,1,1)(j)}_{Y_2=1|Y_1=0} + N^{D}_{01\bullet} \\
&\quad + N^{F}_{\bullet 1\bullet}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=0} + N^{G}_{0\bullet\bullet}\hat{\pi}^{(0,1,1)(j)}_{Y_2=1|Y_1=0}\Big] \Big/ \Big[N^{A}_{0++} + N^{B}_{\bullet++}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=0} + N^{C}_{0\bullet+} + N^{E}_{\bullet\bullet+}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=0} \\
&\quad + N^{D}_{0+\bullet} + N^{F}_{\bullet+\bullet}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=0} + N^{G}_{0\bullet\bullet}\Big].
\end{aligned}$$

When the algorithm converges, we have

$$\hat{\pi}^{(0,1,1)(\infty)}_{Y_2=0|Y_1=1} = 1 - \hat{\pi}^{(0,1,1)(\infty)}_{Y_2=1|Y_1=1} \text{ and } \hat{\pi}^{(0,1,1)(\infty)}_{Y_2=0|Y_1=0} = 1 - \hat{\pi}^{(0,1,1)(\infty)}_{Y_2=1|Y_1=0}.$$

Now we complete the missingness for each data segment at time point $k = 2$ by

$$\hat{N}^{\bullet(2)}_{11+} = \hat{N}^{\bullet(1)}_{11+} + \hat{\pi}^{(0,1,1)(\infty)}_{Y_2=1|Y_1=1}\hat{N}^{\bullet(1)}_{1\bullet+}$$

$$\hat{N}^{\bullet(2)}_{10+} = \hat{N}^{\bullet(1)}_{10+} + \hat{\pi}^{(0,1,1)(\infty)}_{Y_2=0|Y_1=1}\hat{N}^{\bullet(1)}_{1\bullet+}$$

$$\hat{N}^{\bullet(2)}_{01+} = \hat{N}^{\bullet(1)}_{01+} + \hat{\pi}^{(0,1,1)(\infty)}_{Y_2=1|Y_1=0}\hat{N}^{\bullet(1)}_{0\bullet+} \text{ and}$$

$$\hat{N}^{\bullet(2)}_{00+} = \hat{N}^{\bullet(1)}_{00+} + \hat{\pi}^{(0,1,1)(\infty)}_{Y_2=0|Y_1=0}\hat{N}^{\bullet(1)}_{0\bullet+}$$

This way, we have the data with counts completed on the second time point, as is listed in Table 2.4, where superscript 2 in A and C stands for completion of the first two time points by the EM algorithm.

Note that Table 2.4 is actually an instance of monotone missing data. In general, for longitudinal data with n time points, after completing the counts through the first $n-1$ time points, the data will have a monotone missing pattern. Thus, by the invariance property of mle, for efficiency in computation, we may obtain the mles of the transition probabilities at time point n, using expressions exploiting the monotone missingness rather than by the EM algorithm. By Theorem 2.1.3, we have

$$\hat{\pi}_{Y_3=1|Y_2=1} = \frac{N^{A^2\bullet(3)}_{+11}}{N^{A^2\bullet(3)}_{+1+}}$$

$$= \frac{N^{A}_{+11} + N^{B}_{\bullet 11} + (N^{C}_{1\bullet 1} + N^{E}_{\bullet\bullet 1}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=1})\hat{\pi}^{(0,1,1)(\infty)}_{Y_2=1|Y_1=1} + (N^{C}_{0\bullet 1} + N^{E}_{\bullet\bullet 1}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=0})\hat{\pi}^{(0,1,1)(\infty)}_{Y_2=1|Y_1=0}}{N^{A}_{+1+} + N^{B}_{\bullet 1+} + (N^{C}_{1\bullet +} + N^{E}_{\bullet\bullet +}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=1})\hat{\pi}^{(0,1,1)(\infty)}_{Y_2=1|Y_1=1} + (N^{C}_{0\bullet +} + N^{E}_{\bullet\bullet +}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=0})\hat{\pi}^{(0,1,1)(\infty)}_{Y_2=1|Y_1=0}}$$

and

$$\hat{\pi}_{Y_3=1|Y_2=0} = \frac{N^{A^2\bullet(3)}_{+01}}{N^{A^2\bullet(3)}_{+0+}}$$

$$= \frac{N^{A}_{+01} + N^{B}_{\bullet 01} + (N^{C}_{1\bullet 1} + N^{E}_{\bullet\bullet 1}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=1})\hat{\pi}^{(0,1,1)(\infty)}_{Y_2=0|Y_1=1} + (N^{C}_{0\bullet 1} + N^{E}_{\bullet\bullet 1}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=0})\hat{\pi}^{(0,1,1)(\infty)}_{Y_2=0|Y_1=0}}{N^{A}_{+0+} + N^{B}_{\bullet 0+} + (N^{C}_{1\bullet +} + N^{E}_{\bullet\bullet +}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=1})\hat{\pi}^{(0,1,1)(\infty)}_{Y_2=0|Y_1=1} + (N^{C}_{0\bullet +} + N^{E}_{\bullet\bullet +}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=0})\hat{\pi}^{(0,1,1)(\infty)}_{Y_2=0|Y_1=0}}.$$

	Y_1	Y_2	Y_3	estimated count
A^1 (complete)	1	1	1	$N^A_{111} + N^B_{\bullet 11}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=1}$
	1	1	0	$N^A_{110} + N^B_{\bullet 10}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=1}$
	1	0	1	$N^A_{101} + N^B_{\bullet 01}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=1}$
	1	0	0	$N^A_{100} + N^B_{\bullet 00}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=1}$
	0	1	1	$N^A_{011} + N^B_{\bullet 11}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=0}$
	0	1	0	$N^A_{010} + N^B_{\bullet 10}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=0}$
	0	0	1	$N^A_{001} + N^B_{\bullet 01}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=0}$
	0	0	0	$N^A_{000} + N^B_{\bullet 00}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=0}$
C^1 (only Y_2 missing)	1	$\bullet$	1	$N^C_{1\bullet 1} + N^E_{\bullet\bullet 1}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=1}$
	1	$\bullet$	0	$N^C_{1\bullet 0} + N^E_{\bullet\bullet 0}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=1}$
	0	$\bullet$	1	$N^C_{0\bullet 1} + N^E_{\bullet\bullet 1}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=0}$
	0	$\bullet$	0	$N^C_{0\bullet 0} + N^E_{\bullet\bullet 0}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=0}$
D^1 (only Y_3 missing)	1	1	$\bullet$	$N^D_{11\bullet} + N^F_{\bullet 1\bullet}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=1}$
	1	0	$\bullet$	$N^D_{10\bullet} + N^F_{\bullet 0\bullet}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=1}$
	0	1	$\bullet$	$N^D_{01\bullet} + N^F_{\bullet 1\bullet}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=0}$
	0	0	$\bullet$	$N^D_{00\bullet} + N^F_{\bullet 0\bullet}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=0}$
G^1 (Y_2 and Y_3 missing)	1	$\bullet$	$\bullet$	$N^G_{1\bullet\bullet}$
	0	$\bullet$	$\bullet$	$N^G_{0\bullet\bullet}$

Table 2.3: Toy example for EM algorithm with Y_1 completed

	Y_1	Y_2	Y_3	estimated count
A^2 (complete)	1	1	1	$N^A_{111} + N^B_{\bullet 11}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=1} + (N^C_{1\bullet 1} + N^E_{\bullet\bullet 1}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=1})\hat{\pi}^{(0,1,1)(\infty)}_{Y_2=1\|Y_1=1}$
	1	1	0	$N^A_{110} + N^B_{\bullet 10}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=1} + (N^C_{1\bullet 0} + N^E_{\bullet\bullet 0}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=1})\hat{\pi}^{(0,1,1)(\infty)}_{Y_2=1\|Y_1=1}$
	1	0	1	$N^A_{101} + N^B_{\bullet 01}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=1} + (N^C_{1\bullet 1} + N^E_{\bullet\bullet 1}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=1})\hat{\pi}^{(0,1,1)(\infty)}_{Y_2=0\|Y_1=1}$
	1	0	0	$N^A_{100} + N^B_{\bullet 00}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=1} + (N^C_{1\bullet 0} + N^E_{\bullet\bullet 0}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=1})\hat{\pi}^{(0,1,1)(\infty)}_{Y_2=0\|Y_1=1}$
	0	1	1	$N^A_{011} + N^B_{\bullet 11}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=0} + (N^C_{0\bullet 1} + N^E_{\bullet\bullet 1}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=0})\hat{\pi}^{(0,1,1)(\infty)}_{Y_2=1\|Y_1=0}$
	0	1	0	$N^A_{010} + N^B_{\bullet 10}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=0} + (N^C_{0\bullet 0} + N^E_{\bullet\bullet 0}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=0})\hat{\pi}^{(0,1,1)(\infty)}_{Y_2=1\|Y_1=0}$
	0	0	1	$N^A_{001} + N^B_{\bullet 01}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=0} + (N^C_{0\bullet 1} + N^E_{\bullet\bullet 1}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=0})\hat{\pi}^{(0,1,1)(\infty)}_{Y_2=0\|Y_1=0}$
	0	0	0	$N^A_{000} + N^B_{\bullet 00}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=0} + (N^C_{0\bullet 0} + N^E_{\bullet\bullet 0}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=0})\hat{\pi}^{(0,1,1)(\infty)}_{Y_2=0\|Y_1=0}$
D^2 (Y_3 missing)	1	1	$\bullet$	$N^D_{11\bullet} + N^F_{\bullet 1\bullet}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=1} + N^G_{1\bullet\bullet}\hat{\pi}^{(0,1,1)(\infty)}_{Y_2=1\|Y_1=1}$
	1	0	$\bullet$	$N^D_{10\bullet} + N^F_{\bullet 0\bullet}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=1} + N^G_{1\bullet\bullet}\hat{\pi}^{(0,1,1)(\infty)}_{Y_2=0\|Y_1=1}$
	0	1	$\bullet$	$N^D_{01\bullet} + N^F_{\bullet 1\bullet}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=0} + N^G_{0\bullet\bullet}\hat{\pi}^{(0,1,1)(\infty)}_{Y_2=1\|Y_1=0}$
	0	0	$\bullet$	$N^D_{00\bullet} + N^F_{\bullet 0\bullet}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=0} + N^G_{0\bullet\bullet}\hat{\pi}^{(0,1,1)(\infty)}_{Y_2=0\|Y_1=0}$

Table 2.4: Toy example for EM algorithm with Y_1 and Y_2 completed

2.2 Maximum likelihood estimation of transition probabilities under two types of stationarity given AD order

If measurement times are equally spaced, it may be of interest to estimate parameters under an AD(p) model with a stationarity property imposed. Two such properties may be of interest: time-invariant transition probabilities, and strict stationarity. If $p \geq 1$, for $k = p+1, \ldots, n$, and we let

$$\pi_{y^{(k)}_{p+1}|y^{(k-p)}_1 \ldots y^{(k-1)}_p} \equiv P(Y_k = y_{p+1}|Y_{k-p} = y_1, y_{k-p+1} = y_2, \ldots, Y_{k-1} = y_p),$$

the property of time-invariant pth-order transition probabilities imposes the constraint

$$\pi_{y_{p+1}^{(p+1)}|y_1^{(1)}\cdots y_p^{(p)}} = \pi_{y_{p+1}^{(p+2)}|y_1^{(2)}\cdots y_p^{(p+1)}} = \cdots = \pi_{y_{p+1}^{(n)}|y_1^{(n-p)}\cdots y_p^{(n-1)}} \tag{2.14}$$

$$\text{with } 1 \le p \le n-2 \text{ for all } (y_1, \ldots, y_p) \in C_p^+ \text{ and } y_{p+1} = 1, \ldots, c-1,$$

where a superscript "+" in C_{p+1}^+ means that the relative positions in time of $y_1, \ldots, y_{p+1}$ are taken into consideration while their absolute positions in time are ignored. Note that (2.14) implies

$$\pi_{c^{(p+1)}|y_1^{(1)}\cdots y_p^{(p)}} = \pi_{c^{(p+2)}|y_1^{(2)}\cdots y_p^{(p+1)}} = \cdots = \pi_{c^{(n)}|y_1^{(n-p)}\cdots y_p^{(n-1)}}.$$

Strict stationarity, which is stronger, imposes the constraint that joint probabilities of all events are invariant to time shifts.

We now give some results relevant to maximum likelihood estimation of an AD(p) model under each stationarity property.

Theorem 2.2.1. *Under AD(p) with $1 \le p \le n-2$ and time-invariant pth-order transition probabilities, $\hat{\pi}^{(p)}_{y_1\cdots y_p+\cdots+} = \dfrac{N_{y_1\cdots y_p+\cdots+}}{N}$; the complete-data mle of the common pth-order transition probability, denoted by $\hat{\pi}^{(p)}_{y_{p+1}^+|y_1^+\cdots y_p^+}$, is as follows: if $\sum_{k=p+1}^{n} N_{+\cdots+y_1^{(k-p)}\cdots y_p^{(k-1)}+\cdots+} = 0$, then $\hat{\pi}^{(p)}_{y_{p+1}^+|y_1^+\cdots y_p^+} = 0$; otherwise,*

$$\hat{\pi}^{(p)}_{y_{p+1}^+|y_1^+\cdots y_p^+} = \frac{\sum_{k=p+1}^{n} N_{+\cdots+y_1^{(k-p)}\cdots y_{p+1}^{(k)}+\cdots+}}{\sum_{k=p+1}^{n} N_{+\cdots+y_1^{(k-p)}\cdots y_p^{(k-1)}+\cdots+}}. \tag{2.15}$$

The theorem says essentially that the mle of the pth-order transition probabilities may be pooled when they are time-invariant to yield the mle of the common pth-order transition probability. The special case of Theorem 2.2.1 in which $p = 1$ and all cells are non-empty was proved by Anderson and Goodman (1957); our proof of the more general result here is very similar.

Proof. Under AD(p), the likelihood (2.4) simplifies to

$$\prod_{(y_1,\ldots,y_p)\in C_p} \pi_{y_1\cdots y_p+\cdots+}^{N_{y_1\cdots y_p+\cdots+}} \prod_{(y_1,\ldots,y_{p+1})\in C_{p+1}^+} \prod_{k=p+1}^{n} \pi_{y_{p+1}^{(k)}|y_1^{(k-p)}\cdots y_p^{(k-1)}}^{N_{+\cdots+y_1^{(k-p)}\cdots y_{p+1}^{(k)}+\cdots+}}. \tag{2.16}$$

In (2.16), $\prod_{(y_1,\ldots,y_p)\in C_p} \pi_{y_1\cdots y_p+\cdots+}^{N_{y_1\cdots y_p+\cdots+}}$ is the product of kernels of multinomial distributions. Thus for each combination of $y_1,\ldots,y_p$, $\hat{\pi}_{y_1\cdots y_p+\cdots+}^{(p)} = \dfrac{N_{y_1\cdots y_p+\cdots+}}{N}$. Now suppose that the transition probabilities are stationary. Then for each given combination of $(y_1,\ldots,y_p)$, the likelihood function of the distribution of $N_{+\cdots+y_1^{(k-p)}\cdots y_{p+1}^{(k)}+\cdots+}$ is proportional to

$$\prod_{y_{p+1}=1}^{c} \prod_{k=p+1}^{n} \pi_{y_{p+1}^+|y_1^+\cdots y_p^+}^{N_{+\cdots+y_1^{(k-p)}\cdots y_{p+1}^{(k)}+\cdots+}} = \prod_{y_{p+1}=1}^{c} \pi_{y_{p+1}^+|y_1^+\cdots y_p^+}^{\sum_{k=p+1}^{n} N_{+\cdots+y_1^{(k-p)}\cdots y_{p+1}^{(k)}+\cdots+}} \tag{2.17}$$

with cell probabilities $\pi_{y_{p+1}^+|y_1^+\cdots y_p^+}$.

Thus, if $\sum_{k=p+1}^{n} N_{+\cdots+y_1^{(k-p)}\cdots y_{p+1}^{(k)}+\cdots+} = 0$, $\hat{\pi}_{y_{p+1}^+|y_1^+\cdots y_p^+}^{(p)} = 0$. Otherwise, $\sum_{k=p+1}^{n} N_{+\cdots+y_1^{(k-p)}\cdots y_{p+1}^{(k)}+\cdots+} \neq 0$ implies $\sum_{y_{p+1}=1}^{c}\sum_{k=p+1}^{n} N_{+\cdots+y_1^{(k-p)}\cdots y_{p+1}^{(k)}+\cdots+} \neq 0$ and

$$\hat{\pi}_{y_{p+1}^+|y_1^+\cdots y_p^+}^{(p)} = \frac{\sum_{k=p+1}^{n} N_{+\cdots+y_1^{(k-p)}\cdots y_{p+1}^{(k)}+\cdots+}}{\sum_{y_{p+1}=1}^{c}\sum_{k=p+1}^{n} N_{+\cdots+y_1^{(k-p)}\cdots y_{p+1}^{(k)}+\cdots+}},$$

yielding (2.15). □

Similarly to the extension from Theorem 2.1.2 to Theorem 2.1.3, we can derive the mle of stationary transition probability under AD(p) when the data are monotone missing.

Theorem 2.2.2. *Under AD(p) for $1 \le p \le n-2$, if the transition probabilities are stationary and the data are monotone missing, the mle of the stationary transition probabilities is given by $\hat{\pi}_{y_1+\cdots+}^{\bullet(p)} = \dfrac{N_{y_1+\cdots+}^{\bullet(1)}}{N^{\bullet(1)}}$, $\hat{\pi}_{y_k|y_1\cdots y_{k-1}}^{\bullet(p)} = \dfrac{N_{y_1\cdots y_k+\cdots+}^{\bullet(k)}}{N_{y_1\cdots y_{k-1}+\cdots+}^{\bullet(k)}}$ for $k =$*

$2, \ldots, p$ *and*

$$\hat{\pi}^{\bullet(p)}_{y^+_{p+1}|y^+_1\cdots y^+_p} = \frac{\sum_{k=p+1}^{n} N^{\bullet(k)}_{+\cdots+y_1^{(k-p)}\cdots y_{p+1}^{(k)}+\cdots+}}{\sum_{k=p+1}^{n} N^{\bullet(k)}_{+\cdots+y_1^{(k-p)}\cdots y_{p}^{(k-1)}+\cdots+}}. \tag{2.18}$$

Proof. Note that for monotone missing data under stationary transition probability AD(p), (2.16) simplifies to

$$\begin{aligned}
&\pi_{y_1}^{N^{\bullet(1)}_{y_1+\cdots+}} \pi_{y_2|y_1}^{N^{\bullet(2)}_{y_1y_2+\cdots+}} \cdots \pi_{y_p|y_1\cdots y_{p-1}}^{N^{\bullet(p)}_{y_1\cdots y_p+\cdots+}} \prod_{(y_1,\ldots,y_{p+1})\in C^+_{p+1}} \prod_{k=p+1}^{n} \pi_{y^+_{p+1}|y^+_1\cdots y^+_p}^{N^{\bullet(k)}_{+\cdots+y_1^{(k-p)}\cdots y_{p+1}^{(k)}+\cdots+}} \\
=&\pi_{y_1}^{N^{\bullet(1)}_{y_1+\cdots+}} \pi_{y_2|y_1}^{N^{\bullet(2)}_{y_1y_2+\cdots+}} \cdots \pi_{y_p|y_1\cdots y_{p-1}}^{N^{\bullet(p)}_{y_1\cdots y_p+\cdots+}} \prod_{(y_1,\ldots,y_{p+1})\in C^+_{p+1}} \pi_{y^+_{p+1}|y^+_1\cdots y^+_p}^{\sum_{k=p+1}^{n} N^{\bullet(k)}_{+\cdots+y_1^{(k-p)}\cdots y_{p+1}^{(k)}+\cdots+}}.
\end{aligned}$$

Thus, (2.18) can be obtained by following the procedure in the proof of Theorem 2.2.1. □

In case of data with arbitrary missing pattern, we have to use the EM algorithm to obtain the mles of stationary transition probabilities under AD(p). In this situation, in contrast to that of the previous section, it is extremely cumbersome to present the EM algorithm in complete generality. Instead, we merely illustrate its application to the toy example of the previous section, for which $n = 3$ and the process is AD(1), but with the added assumption that the transition probabilities are time-invariant. For the first time point, the procedure is the same as that which goes from Table 2.2 to Table 2.3. So we start from Table 2.3. To move forward for stationary transition probabilities under AD(1) from Table 2.3, we have

$$\begin{aligned}
\left(N^{C^1}_{11+}, N^{C^1}_{10+}\right) &\sim \text{Multinomial}\left(N^{C^1}_{1\bullet+}, (\pi_{1^+|1^+},\ \pi_{0^+|1^+})\right), \\
\left(N^{G^1}_{11\bullet}, N^{G^1}_{10\bullet}\right) &\sim \text{Multinomial}\left(N^{G^1}_{1\bullet\bullet}, (\pi_{1^+|1^+},\ \pi_{0^+|1^+})\right).
\end{aligned}$$

The E-step for N_{11+} is

$$E(N_{11+}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+})$$
$$=E(N_{11+}^{A^1} + N_{11+}^{D^1} + N_{11+}^{C^1} + N_{11+}^{G^1}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+})$$
$$=N_{11+}^{A} + N_{\bullet 1+}^{B}\hat{\pi}_{Y_1=1}^{(0,1,1)(\infty)} + N_{11\bullet}^{D} + N_{\bullet 1\bullet}^{F}\hat{\pi}_{Y_1=1}^{(0,1,1)(\infty)} + (N_{1\bullet+}^{C} + N_{\bullet\bullet+}^{E}\hat{\pi}_{Y_1=1}^{(0,1,1)(\infty)})\pi_{1^+|1^+} + N_{1\bullet\bullet}^{G}\pi_{1^+|1^+};$$

the E-step for N_{1++} is

$$E(N_{1++}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+})$$
$$=E(N_{1++}^{A^1} + N_{1++}^{D^1} + N_{1++}^{C^1} + N_{1++}^{G^1}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+})$$
$$=N_{1++}^{A} + N_{\bullet++}^{B}\hat{\pi}_{Y_1=1}^{(0,1,1)(\infty)} + N_{1+\bullet}^{D} + N_{\bullet+\bullet}^{F}\hat{\pi}_{Y_1=1}^{(0,1,1)(\infty)} + N_{1\bullet+}^{C} + N_{\bullet\bullet+}^{E}\hat{\pi}_{Y_1=1}^{(0,1,1)(\infty)} + N_{1\bullet\bullet}^{G}.$$

The E-steps for N_{11+} and N_{1++} are straightforward, while the E-steps for N_{+11} and N_{+1+} are based on the E-steps for N_{11+} and N_{1++}. To move further forward, the likelihood (2.17) also indicates that

$$\left(N_{+11}^{C^1}, N_{+10}^{C^1}\right) \sim \text{Multinomial}\left(N_{+1+}^{C^1}, (\pi_{1^+|1^+},\ \pi_{0^+|1^+})\right),$$
$$\text{where } N_{+1+}^{C^1} = N_{1\bullet+}^{C^1}\pi_{1^+|1^+} + N_{0\bullet+}^{C^1}\pi_{1^+|0^+}$$
$$\left(N_{+11}^{D^1}, N_{+10}^{D^1}\right) \sim \text{Multinomial}\left(N_{+1\bullet}^{D^1}, (\pi_{1^+|1^+},\ \pi_{0^+|1^+})\right),$$
$$\left(N_{+11}^{G^1}, N_{+10}^{G^1}\right) \sim \text{Multinomial}\left(N_{+1\bullet}^{G^1}, (\pi_{1^+|1^+},\ \pi_{0^+|1^+})\right),$$
$$\text{where } N_{+1\bullet}^{G^1} = N_{1\bullet\bullet}^{G^1}\pi_{1^+|1^+} + N_{0\bullet\bullet}^{G^1}\pi_{1^+|0^+}.$$

So clearly

$$E(N_{+11}^{A^1}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+}) = N_{+11}^{A^1} = N_{+11}^{A} + N_{\bullet 11}^{B}$$

and

$$E(N_{+11}^{D^1}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+}) = N_{+1\bullet}^{D^1}\pi_{1^+|1^+} = (N_{+1\bullet}^{D} + N_{\bullet 1\bullet}^{F})\pi_{1^+|1^+}.$$

But

$$
\begin{aligned}
&E(N^{C^1}_{+11}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+})\\
=&E(N^{C^1}_{111}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+}) + E(N^{C^1}_{011}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+})\\
=&E(N^{C^1}_{1\bullet 1}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+})\pi_{1^+|1^+} + E(N^{C^1}_{0\bullet 1}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+})\pi_{1^+|0^+}\\
=&(N^{C}_{1\bullet 1} + N^{E}_{\bullet\bullet 1}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=1})\pi_{1^+|1^+} + (N^{C}_{0\bullet 1} + N^{E}_{\bullet\bullet 1}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=0})\pi_{1^+|0^+}
\end{aligned}
$$

and

$$
\begin{aligned}
&E(N^{G^1}_{+11}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+})\\
=&E(N^{G^1}_{111}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+}) + E(N^{G^1}_{011}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+})\\
=&E(N^{G^1}_{1\bullet\bullet}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+})\pi_{1^+|1^+}\pi_{1^+|1^+} + E(N^{G^1}_{0\bullet\bullet}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+})\pi_{1^+|0^+}\pi_{1^+|1^+}\\
=&N^{G}_{1\bullet\bullet}\pi_{1^+|1^+}\pi_{1^+|1^+} + N^{G}_{0\bullet\bullet}\pi_{1^+|0^+}\pi_{1^+|1^+}.
\end{aligned}
$$

Thus we have the E-step for N_{+11}:

$$
\begin{aligned}
&E(N_{+11}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+})\\
=&E(N^{A^1}_{+11} + N^{D^1}_{+11} + N^{C^1}_{+11} + N^{G^1}_{+11}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+})\\
=&N^{A}_{+11} + N^{B}_{\bullet 11} + (N^{D}_{+1\bullet} + N^{F}_{\bullet 1\bullet})\pi_{1^+|1^+}\\
&+ (N^{C}_{1\bullet 1} + N^{E}_{\bullet\bullet 1}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=1})\pi_{1^+|1^+} + (N^{C}_{0\bullet 1} + N^{E}_{\bullet\bullet 1}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=0})\pi_{1^+|0^+}\\
&+ (N^{G}_{1\bullet\bullet}\pi_{1^+|1^+} + N^{G}_{0\bullet\bullet}\pi_{1^+|0^+})\pi_{1^+|1^+}.
\end{aligned}
$$

Similarly to the procedure of obtaining $N^{A^1}_{+11},\ N^{D^1}_{+11},\ N^{C^1}_{+11},\ N^{G^1}_{+11}$, we have

$$E(N^{A^1}_{+1+}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+}) = N^{A^1}_{+1+} = N^{A}_{+1+} + N^{B}_{\bullet 1+}$$

and

$$E(N^{D^1}_{+1+}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+}) = N^{D^1}_{+1\bullet} = N^{D}_{+1\bullet} + N^{F}_{\bullet 1\bullet},$$

but

$$
\begin{aligned}
&E(N^{C^1}_{+1+}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+})\\
=&E(N^{C^1}_{11+}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+}) + E(N^{C^1}_{01+}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+})\\
=&E(N^{C^1}_{1\bullet+}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+})\pi_{1^+|1^+} + E(N^{C^1}_{0\bullet+}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+})\pi_{1^+|0^+}\\
=&(N^{C}_{1\bullet+} + N^{E}_{\bullet\bullet+}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=1})\pi_{1^+|1^+} + (N^{C}_{0\bullet+} + N^{E}_{\bullet\bullet+}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=0})\pi_{1^+|0^+}
\end{aligned}
$$

and

$$
\begin{aligned}
&E(N^{G^1}_{+1+}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+})\\
=&E(N^{G^1}_{11+}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+}) + E(N^{G^1}_{01+}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+})\\
=&E(N^{G^1}_{1\bullet\bullet}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+})\pi_{1^+|1^+} + E(N^{G^1}_{0\bullet\bullet}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+})\pi_{1^+|0^+}\\
=&N^{G}_{1\bullet\bullet}\pi_{1^+|1^+} + N^{G}_{0\bullet\bullet}\pi_{1^+|0^+}.
\end{aligned}
$$

Thus we have the E-step for N_{+1+}:

$$
\begin{aligned}
&E(N_{+1+}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+})\\
=&E(N^{A^1}_{+1+} + N^{D^1}_{+1+} + N^{C^1}_{+1+} + N^{G^1}_{+1+}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+})\\
=&N^{A}_{+1+} + N^{B}_{\bullet 1+} + (N^{D}_{+1\bullet} + N^{F}_{\bullet 1\bullet})\\
&+\left((N^{C}_{1\bullet+} + N^{E}_{\bullet\bullet+}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=1})\pi_{1^+|1^+} + (N^{C}_{0\bullet+} + N^{E}_{\bullet\bullet+}\hat{\pi}^{(0,1,1)(\infty)}_{Y_1=0})\pi_{1^+|0^+}\right)\\
&+N^{G}_{1\bullet\bullet}\pi_{1^+|1^+} + N^{G}_{0\bullet\bullet}\pi_{1^+|0^+}.
\end{aligned}
$$

The M-step is

$$
\hat{\pi}_{1^+|1^+} = \frac{E(N_{11+}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+}) + E(N_{+11}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+})}{E(N_{1++}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+}) + E(N_{+1+}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+})}.
$$

Combining the two steps yields a single iteration of EM,

$$
\hat{\pi}^{(0,1,1)(j+1)}_{1^+|1^+} = \frac{E(N_{11+}|data, \hat{\pi}^{(0,1,1)(j)}_{1^+|1^+},\ \hat{\pi}^{(0,1,1)(j)}_{1^+|0^+}) + E(N_{+11}|data, \hat{\pi}^{(0,1,1)(j)}_{1^+|1^+},\ \hat{\pi}^{(0,1,1)(j)}_{1^+|0^+})}{E(N_{1++}|data, \hat{\pi}^{(0,1,1)(j)}_{1^+|1^+},\ \hat{\pi}^{(0,1,1)(j)}_{1^+|0^+}) + E(N_{+1+}|data, \hat{\pi}^{(0,1,1)(j)}_{1^+|1^+},\ \hat{\pi}^{(0,1,1)(j)}_{1^+|0^+})}
$$

where j stands for the step of iteration.

Similarly,

$$
\hat{\pi}^{(0,1,1)(j+1)}_{1^+|0^+} = \frac{E(N_{01+}|data, \hat{\pi}^{(0,1,1)(j)}_{1^+|1^+},\ \hat{\pi}^{(0,1,1)(j)}_{1^+|0^+}) + E(N_{+01}|data, \hat{\pi}^{(0,1,1)(j)}_{1^+|1^+},\ \hat{\pi}^{(0,1,1)(j)}_{1^+|0^+})}{E(N_{0++}|data, \hat{\pi}^{(0,1,1)(j)}_{1^+|1^+},\ \hat{\pi}^{(0,1,1)(j)}_{1^+|0^+}) + E(N_{+0+}|data, \hat{\pi}^{(0,1,1)(j)}_{1^+|1^+},\ \hat{\pi}^{(0,1,1)(j)}_{1^+|0^+})},
$$

where $E(N_{01+}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+})$, $E(N_{+01}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+})$, $E(N_{0++}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+})$ and $E(N_{+0+}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+})$ can be obtained in a similar way to $E(N_{11+}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+})$, $E(N_{+11}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+})$, $E(N_{1++}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+})$ and $E(N_{+1+}|data, \pi_{1^+|1^+},\ \pi_{1^+|0^+})$ respectively. In each iteration, we do the E-step chronologically but we do the M-step one at a time. Thus we see that the $\hat{\pi}^{(0,1,1)(j+1)}_{1^+|1^+}$ does not only depend on $\hat{\pi}^{(0,1,1)(j)}_{1^+|1^+}$, but also depends on $\hat{\pi}^{(0,1,1)(j)}_{1^+|0^+}$. Because of such relationships between $\hat{\pi}^{(0,1,1)}_{1^+|1^+}$ and $\hat{\pi}^{(0,1,1)}_{1^+|0^+}$, we see that it is impossible to express $\hat{\pi}^{(0,1,1)}_{1^+|1^+}$ and $\hat{\pi}^{(0,1,1)}_{1^+|0^+}$ in terms of observations explicitly like (2.11) and (2.12), and thus there is no explicit closed form of mles of $\hat{\pi}^{(0,1,1)}_{1^+|1^+}$ and $\hat{\pi}^{(0,1,1)}_{1^+|0^+}$ for the iterations in EM algorithm when the data have an arbitrary missing pattern.

After completing missingness and obtaining mle for stationary transition probabilities for this toy example, we generalize the procedure for an arbitrary number of time points n, arbitrary constant order of antedependence ($1 \leq p \leq n-2$) and arbitrary number of categories of the outcomes c. Based on likelihood (2.16), we can complete the missingness at the first p time points by following the same procedure shown in Section 2.1. To obtain $\hat{\pi}^{(p)}_{y^+_{p+1}|y^+_1\cdots y^+_p}$, in each iteration of the EM algorithm, we perform the E-step chronologically: for each $k = p+1, \ldots, n$, we not only consider the subjects with observed realizations of $y_1^{(k-p)}, \ldots, y_p^{(k-1)}, y_{p+1}^{(k)}$ but also consider the subjects with incomplete realizations at time point k, $N_{+\cdots+y_1^{(k-p)}\cdots y_p^{(k-1)}\bullet+\cdots+}$, having probability $\hat{\pi}^{(p)}_{y^+_{p+1}|y^+_1\cdots y^+_p}$ to have realization y_{p+1} at time point k; and we perform the M-step one at a time by

$$\hat{\pi}^{(p)(j+1)}_{y^+_{p+1}|y^+_1\cdots y^+_p} = \frac{\sum_{k=p+1}^{n} \hat{N}^{(j)}_{+\cdots+y_1^{(k-p)}\cdots y_{p+1}^{(k)}+\cdots+}}{\sum_{k=p+1}^{n} \hat{N}^{(j)}_{+\cdots+y_1^{(k-p)}\cdots y_p^{(k-1)}+\cdots+}}$$

where the estimated counts at the j-th iteration $\hat{N}^{(j)}_{+\cdots+y_1^{(k-p)}\cdots y_{p+1}^{(k)}+\cdots+}$ and $\hat{N}^{(j)}_{+\cdots+y_1^{(k-p)}\cdots y_p^{(k-1)}+\cdots+}$

are functions of $\hat{\pi}^{(p)(j)}_{y^+_{p+1}|y^+_1\cdots y^+_p}$.

We noted previously that strict stationarity imposes time-shift invariance upon the joint probabilities of all events. The following lemma gives an alternative set of necessary and sufficient conditions for strict stationarity under the AD(p) model, which makes explicit the additional restrictions, beyond those required for time-invariance of pth-order transition probabilities, imposed by strict stationarity.

Lemma 2.2.3. *Under AD(p) with $p \geq 1$, the variables $Y_1, \ldots, Y_n$ are strictly stationary if and only if (2.14) holds and for $1 \leq p \leq n-1$,*

$$\pi_{y_1\cdots y_p+\cdots+} = \sum_{y_0=1}^{c} \pi_{y_p|y_0\cdots y_{p-1}} \pi_{y_0\cdots y_{p-1}+\cdots+} \text{ for all } (y_1, \ldots, y_p) \in C_p \setminus \{c, \ldots, c\}. \tag{2.19}$$

Note that when $p = n-1$, the model is saturated where transition probability stationarity makes no sense, while the sufficient and necessary condition of strict stationarity under AD($n-1$) is well known to simplify to (2.19) only.

Proof. Note that (2.19) is algebraically equivalent to

$$\pi_{y_1\cdots y_p+\cdots+} = \pi_{+y_1\cdots y_p+\cdots+} \text{ for all } (y_1, \ldots, y_p) \in C_p \setminus \{c, \ldots, c\}. \tag{2.20}$$

Define the *length* of a joint probability of an event involving only consecutively-indexed random variables as the number of those random variables. Then, strict stationarity implies the time-shift invariance of all joint probabilities of length $1, 2, \ldots, n-1$. Since each conditional probability in (2.14) may be written as the ratio of two joint probabilities involving only consecutively-indexed variables, strict stationarity also implies time-shift invariance of each such conditional probability. Thus, necessity is established.

Next, we prove sufficiency. Equations (2.20) and the first equality in (2.14) imply that for all $(y_1, \ldots, y_{p+1}) \in C_{p+1}$,

$$P(Y_1 = y_1, \ldots, Y_p = y_p, Y_{p+1} = y_{p+1}) = P(Y_2 = y_1, \ldots, Y_{p+1} = y_p, Y_{p+2} = y_{p+1}).$$

Summing over y_1 yields, for all $(y_2, \ldots, y_{p+1}) \in C_p$,

$$P(Y_2 = y_2, \ldots, Y_p = y_p, Y_{p+1} = y_{p+1}) = P(Y_3 = y_2, \ldots, Y_{p+1} = y_p, Y_{p+2} = y_{p+1}). \tag{2.21}$$

Similarly, pairing equation (2.21) with the second equality in (2.14) and summing over y_2 yields, for all $(y_3, \ldots, y_{p+2}) \in C_p$,

$$P(Y_3 = y_3, \ldots, Y_{p+2} = y_{p+2}) = P(Y_4 = y_3, \ldots, Y_{p+3} = y_{p+2}).$$

Successive such pairings and summation establish the time-shift invariance of joint probabilities of all events of length p. By summing over the possible outcomes for variables on either end of the sequence, it is easily seen that the same invariance also holds for all joint probabilities of length less than p. Thus, since any joint probability of length $p+1$ or more can be expressed as the product of a joint probability of length p and conditional probabilities for which the conditioning event involves only the p immediately preceding variables, more specifically, for $l = p+1, \ldots, n-1$,

$$\begin{aligned} &P(Y_1 = y_1, ..., Y_l = y_l) \\ &= \left[\prod_{k=p+1}^{l} P(Y_k = y_k | Y_{k-p} = y_{k-p}, \ldots, Y_{k-1} = y_{k-1}) \right] \times P(Y_1 = y_1, \ldots, Y_p = y_p), \end{aligned}$$

all joint probabilities of lengths 1 to $n-1$ are time-shift invariant. Invariance of joint probabilities of events involving non-consecutive variables can be derived from that of joint probabilities of events involving consecutively-indexed variables by summing over the possible values of the out-of-sequence variables. Thus, sufficiency is proved. □

Unfortunately, mles of neither transition nor cell probabilities of an AD(p) model under strict stationarity can be expressed in closed form. However, since the constraints imposed by (2.2), (2.14), and (2.20) may be written as nonredundant homogeneous smooth functions of the cell probabilities $\pi_{y_1 \cdots y_n}$, the mles can be

obtained numerically using the algorithm of Lang (2004), which can maximize the multinomial likelihood function subject to arbitrary homogeneous constraints on cell probabilities, i.e $\mathbf{h}(\boldsymbol{\pi}) = \mathbf{0}$. Here,

$$\mathbf{h}(\boldsymbol{\pi}) = \left[\begin{array}{c} \mathbf{M}_p\boldsymbol{\Psi} \\ \hline \log\left(\dfrac{\pi_{y_{p+1}^{(p+1)}|y_1^{(1)}\cdots y_p^{(p)}}}{\pi_{y_{p+1}^{(p+2)}|y_1^{(2)}\cdots y_p^{(p+1)}}}\right) \\ \vdots \\ \log\left(\dfrac{\pi_{y_{p+1}^{(n-1)}|y_1^{(n-p-1)}\cdots y_p^{(n-2)}}}{\pi_{y_{p+1}^{(n)}|y_1^{(n-p)}\cdots y_p^{(n-1)}}}\right) \\ \hline \log\left(\dfrac{\pi_{y_1\cdots y_p+\cdots+}}{\pi_{+y_1\cdots y_p+\cdots+}}\right) \end{array}\right]. \tag{2.22}$$

In (2.22),

$$\mathbf{M}_p \equiv \left[\begin{array}{c|c} \mathbf{0}_{d_p\times(d_0-d_p)} & \mathbf{I}_{d_p\times d_p} \end{array}\right]$$

and $\boldsymbol{\Psi}$ is a vector containing all conditional log odds ratios, which are functions of cell probabilities $\pi_{y_1\cdots y_n}$, defined in Section 4.1.3. In that same section it will be shown that $\mathbf{M}_p\boldsymbol{\Psi} = \mathbf{0}$ is equivalent to (2.2), all conditional independences implied by AD(p).

$$\left[\begin{array}{c} \log\left(\dfrac{\pi_{y_{p+1}^{(p+1)}|y_1^{(1)}\cdots y_p^{(p)}}}{\pi_{y_{p+1}^{(p+2)}|y_1^{(2)}\cdots y_p^{(p+1)}}}\right) \\ \vdots \\ \log\left(\dfrac{\pi_{y_{p+1}^{(n-1)}|y_1^{(n-p-1)}\cdots y_p^{(n-2)}}}{\pi_{y_{p+1}^{(n)}|y_1^{(n-p)}\cdots y_p^{(n-1)}}}\right) \end{array}\right] = \mathbf{0}$$

are equivalent to (2.14) and they are functions of cell probabilities $\pi_{y_1\cdots y_n}$ by (2.1).

Note that since the likelihood function for strict stationarity under AD(p) cannot be expressed explicitly as functions of saturated multinomial likelihoods, our EM algorithm would not work for incomplete data since E-step requires the

kernel of saturated multinomial likelihood.

CHAPTER 3
MODEL SELECTION USING PENALIZED LOG-LIKELIHOOD

3.1 Order selection

In general, variable order models are more parsimonious than constant order models. For example AD(0,1,2,3,2,1,0,1) is nested and more parsimonious than AD(3). Thus for parsimony, we start the order selection from variable order. However, likelihood-based hypothesis testing is not amenable to selection among all possible variable-order antedependence models because the models are not completely nested. Consequently, for this purpose we propose the use of penalized likelihood criteria. For AD$(p_1, \ldots, p_n)$ models, we consider information criteria of the form

$$IC(p_1, \ldots, p_n) = a(N)(c-1)\sum_{k=1}^{n} c^{p_k} - 2\log L(\hat{\boldsymbol{\pi}}^{(p_1 \cdots p_n)}),$$

where $a(N)$ is a specified function of N. Many penalized likelihood criteria are cases of this general form, including AIC, BIC, corrected AIC, quasi-AIC, and quasi-BIC. In our examples we will feature AIC and BIC, for which $a(N) = 2$ and $a(N) = \log N$, respectively. Note that the criteria are expressed in "smaller is better" form.

Since p_k can be any nonnegative integer less than or equal to $k-1$, the number of variable-order AD models to compare is $n!$. Computing the information criterion for all $n!$ models can be burdensome or even impractical when n is not small; when $n = 7$ and the data are complete, about 24 hours of computing time on an Intel(R) Xeon(R) with W3520 processor (speed 2.67 GHz, memory 8158412 kB) are required to obtain $AIC(p_1, \ldots, p_n)$ for all 5040 models. It turns out, however, that (provided the data are complete or monotone missing) the variable-order AD

model that minimizes the information criterion can be determined by optimizing p_k separately for each k, so that only $[n(n+1)/2]-1$ models must be fitted and much computational time can be saved. This is a consequence of the following theorem.

Theorem 3.1.1. *(a) Suppose that the data are complete. Then the minimizer of $IC(p_1,\ldots,p_n)$ is given by*

$$\begin{aligned}\hat{p}_k &= \operatorname*{argmin}_{p_k=0,\ldots,k-1}\Bigg(\frac{a(N)}{2}(c-1)c^{p_k} - \Bigg[I(p_k=0)\sum_{y_k\in C_1} N_{+\cdots+y_k+\cdots+}\log\hat{\pi}^{(p_1\cdots p_n)}_{+\cdots+y_k+\cdots+} \\ &\quad +I(p_k\geq 1)\sum_{(y_{k-p_k},\ldots,y_k)\in C_{p_k+1}} N_{+\cdots+y_{k-p_k}\cdots y_k+\cdots+}\log\hat{\pi}^{(p_1\cdots p_n)}_{y_k|y_{k-p_k}\cdots y_{k-1}}\Bigg]\Bigg) \\ &\equiv \operatorname*{argmin}_{p_k=0,\ldots,k-1} AIC_k \qquad (3.1)\end{aligned}$$

for $k=1,\ldots,n$.

(b) Suppose that the data are ignorably monotone missing. Then the minimizer of $IC(p_1,\ldots,p_n)$ is given by an expression for $\hat{p}_k$ identical to (3.1) except that $N^{\bullet(k)}$, $N^{\bullet(k)}_{+\cdots+y_k+\cdots+}$, $N^{\bullet(k)}_{+\cdots+y_{k-p_k}\cdots y_k+\cdots+}$, $\hat{\pi}^{\bullet(p_1\cdots p_n)}_{+\cdots+y_k+\cdots+}$ and $\hat{\pi}^{\bullet(p_1\cdots p_n)}_{y_k|y_{k-p_k}\cdots y_{k-1}}$ are substituted for N, $N_{+\cdots+y_k+\cdots+}$, and $N_{+\cdots+y_{k-p_k}\cdots y_k+\cdots+}$, $\hat{\pi}^{(p_1\cdots p_n)}_{+\cdots+y_k+\cdots+}$ and $\hat{\pi}^{(p_1\cdots p_n)}_{y_k|y_{k-p_k}\cdots y_{k-1}}$, respectively.

(c) Suppose that the data are of arbitrary missing pattern. Then the minimizer of $IC(p_1,\ldots,p_n)$ is given by an expression for $\hat{p}_k$ identical to (3.1) except that $\hat{N}^{\bullet(k)}$, $\hat{N}^{\bullet(k)}_{+\cdots+y_k+\cdots+}$, $\hat{N}^{\bullet(k)}_{+\cdots+y_{k-p_k}\cdots y_k+\cdots+}$, $\hat{\pi}^{\star(p_1\cdots p_n)}_{+\cdots+y_k+\cdots+}$ and $\hat{\pi}^{\star(p_1\cdots p_n)}_{y_k|y_{k-p_k}\cdots y_{k-1}}$ are substituted for N, $N_{+\cdots+y_k+\cdots+}$, $N_{+\cdots+y_{k-p_k}\cdots y_k+\cdots+}$, $\hat{\pi}^{(p_1\cdots p_n)}_{+\cdots+y_k+\cdots+}$ and $\hat{\pi}^{(p_1\cdots p_n)}_{y_k|y_{k-p_k}\cdots y_{k-1}}$, where the mle of transition probabilities $\hat{\pi}^{\star(p_1\cdots p_n)}_{+\cdots+y_k+\cdots+}$ and $\hat{\pi}^{\star(p_1\cdots p_n)}_{y_k|y_{k-p_k}\cdots y_{k-1}}$ can be obtained by Theorem 2.1.4 when EM algorithm converges.

Proof. Consider part (a) first. Using (2.6), we have

$$\begin{aligned}&IC(p_1,\ldots,p_n)\\ =&a(N)(c-1)\sum_{k=1}^{n}c^{p_k} - 2\log L(\hat{\boldsymbol{\pi}}^{(p_1\cdots p_n)})\end{aligned}$$

$$
\begin{aligned}
&= a(N)(c-1)\sum_{k=1}^{n} c^{p_k} - 2\log(N!) + 2 \sum_{(y_1\cdots y_n)\in\{1,\ldots,c\}^n} \log(N_{y_1\cdots y_n}!) \\
&\quad - 2\log\left(\prod_{(y_1\cdots y_n)\in\{1,\ldots,c\}^n} \left(\hat{\pi}^{(p_1\cdots p_n)}_{y_1\cdots y_n}\right)^{N_{y_1\cdots y_n}}\right) \qquad (3.2) \\
&= constant + 2\sum_{k=1}^{n}\left(\frac{a(N)}{2}(c-1)c^{p_k} - \left[I(p_k=0)\sum_{y_k\in C_1} N_{+\cdots+y_k+\cdots+}\log\hat{\pi}^{(p_1\cdots p_n)}_{+\cdots+y_k+\cdots+}\right.\right. \\
&\quad \left.\left. + I(p_k\geq 1)\sum_{(y_{k-p_k},\ldots,y_k)\in C_{p_k+1}} N_{+\cdots+y_{k-p_k}\cdots y_k+\cdots+}\log\hat{\pi}^{(p_1\cdots p_n)}_{y_k|y_{k-p_k}\cdots y_{k-1}}\right]\right). \qquad (3.3)
\end{aligned}
$$

Only the kth summand in (3.3) depends on p_k, yielding part (a). The proof of part (b) is identical, apart from using the likelihood (2.10) rather than (2.6). For part (c), after the EM algorithm converges, we have the kernel of the maximized likelihood function:

$$
\begin{aligned}
&\prod_{k=1}^{n}\left[\left(I(p_k=0)\prod_{y_k\in C_1}\left(\hat{\pi}^{\star(p_1\cdots p_n)}_{+\cdots+y_k+\cdots+}\right)^{\hat{N}^{\bullet(k)}_{+\cdots+y_k+\cdots+}}\right)\right. \\
&\quad \left. + \left(I(p_k\geq 1)\prod_{(y_{k-p_k},\ldots,y_k)\in C_{p_k+1}}\left(\hat{\pi}^{\star(p_1\cdots p_n)}_{y_k|y_{k-p_k}\cdots y_{k-1}}\right)^{\hat{N}^{\bullet(k)}_{+\cdots+y_{k-p_k}\cdots y_{k-1}y_k+\cdots+}}\right)\right]. \qquad (3.4)
\end{aligned}
$$

Thus, in case of data with arbitrary missing pattern, by using (3.4) as the maximized likelihood, we have

$$
\begin{aligned}
&IC(p_1,\ldots,p_n) \\
&= constant + 2\sum_{k=1}^{n}\left(\frac{a(N)}{2}(c-1)c^{p_k} - \left[I(p_k=0)\sum_{y_k\in C_1}\hat{N}^{\bullet(k)}_{+\cdots+y_k+\cdots+}\log\hat{\pi}^{\star(p_1\cdots p_n)}_{+\cdots+y_k+\cdots+}\right.\right. \\
&\quad \left.\left. + I(p_k\geq 1)\sum_{(y_{k-p_k},\ldots,y_k)\in C_{p_k+1}}\hat{N}^{\bullet(k)}_{+\cdots+y_{k-p_k}\cdots y_k+\cdots+}\log\hat{\pi}^{\star(p_1\cdots p_n)}_{y_k|y_{k-p_k}\cdots y_{k-1}}\right]\right). \qquad (3.5)
\end{aligned}
$$

Only the kth summand in (3.5) depends on p_k, yielding part (c).

□

Of course, an advantage of determining the values of an information criterion for all $n!$ possible models is that doing so indicates which other variable-order AD

models besides the one with minimum IC might also be worthy of consideration. Burnham and Anderson (2002, p. 70) suggest that models with $AIC \leq AIC_{min} + 2$ should be considered (with a similar rule if BIC is used rather than AIC). A way to implement the efficient algorithm of Theorem 3.1.1 to determine additional variable-order AD models that adhere to Burnham and Anderson's suggestion is to retain for further consideration any p_k for which

$$AIC_k \leq \min(AIC_k) + (2/n),$$

where AIC_k is defined in (3.1).

CHAPTER 4
HYPOTHESIS TESTS FOR THE ORDER OF ANTEDEPENDENCE

In this chapter, we consider likelihood-based tests of hypotheses concerning antedependence, under the same sampling framework of the previous two chapters. Hypothesis tests of interest may include those of pth-order antedependence versus a saturated model (antedependence of order $n-1$), pth order antedependence versus qth-order antedependence (where $p < q < n-1$), and most generally, variable-order antedependence of given order versus variable-order antedependence of a higher order. For comparison purposes, we also introduce how to determine the order of antedependence by an adaptation of Freeman and Halton's (1951) exact test. For simplicity of presentation, we assume in this chapter that the data are complete. Modifications to deal with missing data are similar to those introduced in Chapter 2.

4.1 AD(p) versus AD($n-1$)

4.1.1 Score test

To derive the score test for antedependence of constant order p against saturated alternative, we write the likelihood as

$$\prod_{(y_1,\ldots,y_n)\in C_n} (\pi_{y_1\cdots y_n})^{N_{y_1\cdots y_n}}$$
$$= \prod_{(y_1,\ldots,y_n)\in C_n} \left(\pi_{y_1\cdots y_p+\cdots+} \prod_{k=p+1}^{n} \pi_{y_k|y_1\cdots y_{k-1}}\right)^{N_{y_1\cdots y_n}}$$

$$= \left(\prod_{(y_1,\ldots,y_p)\in C_p} \pi_{y_1\cdots y_p+\cdots+}^{N_{y_1\cdots y_p+\cdots+}} \right) \left(\prod_{k=p+1}^{n} \prod_{(y_1,\ldots,y_{k-1})\in C_{k-1}} \prod_{y_k=1}^{c} \pi_{y_k|y_1\cdots y_{k-1}}^{N_{y_1\cdots y_k+\cdots+}} \right). \tag{4.1}$$

In (4.1), $\prod_{(y_1,\ldots,y_p)\in C_p} \pi_{y_1\cdots y_p+\cdots+}^{N_{y_1\cdots y_p+\cdots+}}$ is the kernel of a saturated multinomial distribution of sample size N and c^p cells with cell probability $\pi_{y_1\cdots y_p+\cdots+}$, and for each given k and combination of $y_1,\ldots,y_{k-1}$, $\prod_{y_k=1}^{c} \pi_{y_k|y_1\cdots y_{k-1}}^{N_{y_1\cdots y_k+\cdots+}}$ is the kernel of a saturated multinomial distribution of sample size $N_{y_1\cdots y_{k-1}+\cdots+}$ and c cells with cell probability $\pi_{y_k|y_1\cdots y_{k-1}}$. Note that likelihood (4.1) implies mutual independence among all multinomial distributions mentioned above. In $\prod_{(y_1,\ldots,y_p)\in C_p} \pi_{y_1\cdots y_p+\cdots+}^{N_{y_1\cdots y_p+\cdots+}}$, for each cell, the observed count and expected count under the null model are $N_{y_1\cdots y_p+\cdots+}$ and $N\hat{\pi}_{y_1\cdots y_p+\cdots+}^{(p)}$, respectively, and the score (Pearson chi-square) statistic simplifies to

$$\sum_{(y_1\cdots y_p)\in C_p} \frac{\left(N_{y_1\cdots y_p+\cdots+} - N\hat{\pi}_{y_1\cdots y_p+\cdots+}^{(p)}\right)^2}{N\hat{\pi}_{y_1\cdots y_p+\cdots+}^{(p)}} = 0.$$

Similarly, in $\prod_{y_k=1}^{c} \pi_{y_k|y_1\cdots y_{k-1}}^{N_{y_1\cdots y_k+\cdots+}}$, the observed count and expected count under the null model are $N_{y_1\cdots y_k+\cdots+}$ and $N_{y_1\cdots y_{k-1}+\cdots+}\hat{\pi}_{y_k|y_1\cdots y_{k-1}}^{(p)}$, respectively, and the score statistic simplifies to

$$\sum_{y_k=1}^{c} \frac{\left(N_{y_1\cdots y_k+\cdots+} - N_{y_1\cdots y_{k-1}+\cdots+}\hat{\pi}_{y_k|y_1\cdots y_{k-1}}^{(p)}\right)^2}{N_{y_1\cdots y_{k-1}+\cdots+}\hat{\pi}_{y_k|y_1\cdots y_{k-1}}^{(p)}}.$$

Based on the independence among all the multinomial distributions mentioned above, the score statistic for testing H_0: AD(p) versus H_1: AD($n-1$) simplifies to

$$X_p^2 \equiv \sum_{k=p+1}^{n} \sum_{(y_1\cdots y_{k-1})\in C_{k-1}} \sum_{y_k=1}^{c} \frac{\left(N_{y_1\cdots y_k+\cdots+} - N_{y_1\cdots y_{k-1}+\cdots+}\hat{\pi}_{y_k|y_1\cdots y_{k-1}}^{(p)}\right)^2}{N_{y_1\cdots y_{k-1}+\cdots+}\hat{\pi}_{y_k|y_1\cdots y_{k-1}}^{(p)}}, \tag{4.2}$$

where by the AD(p) property,

$$\begin{aligned}\hat{\pi}_{y_k|y_1\cdots y_{k-1}}^{(p)} =& I(p=0)\hat{\pi}_{+\cdots+y_k+\cdots+}^{(0)} + I(p\geq 1)\hat{\pi}_{y_k|y_{k-p}\cdots y_{k-1}}^{(p)} \\ =& I(p=0)\frac{N_{+\cdots+y_k+\cdots+}}{N} + I(p\geq 1)\frac{N_{+\cdots+y_{k-p}\cdots y_{k-1}y_k+\cdots+}}{N_{+\cdots+y_{k-p}\cdots y_{k-1}+\cdots+}}\end{aligned} \tag{4.3}$$

as shown in Chapter 2 and we define C_{k-1} as the set of outcomes $(y_1,\ldots,y_{k-1}) \in$

$\{1,\ldots,c\}^{k-1}$ for which $N_{y_1\cdots y_{k-1}+\cdots+} > 0$. (Note thus that the score test is not affected by zero cell counts.)

Write $\Theta^{(p)}$ for $\Theta^{(p_1\cdots p_n)}$ when $p_k = min(k-1,p)$. Thus by (2.3), we have

$$\begin{aligned} dim(\Theta^{(p)}) =& (c-1)\left\{c^0 + c^1 + \cdots + c^p + [n-(p+2)+1]c^p\right\} \\ =& (n-p)c^{p+1} - (n-p-1)c^p - 1 \text{ for } p \in \{1,\ldots,n-2\} \end{aligned}$$

and $dim(\Theta^{(0)}) = n(c-1)$.

By the additivity property of independent chi-square random variables, the limiting (as $N \to \infty$) null distribution of X_p^2 is chi-square with degrees of freedom, $d_p = dim(\Theta) - dim(\Theta^{(p)})$. Thus, for arbitrary p,

$$d_p = c^n - (n-p)c^{p+1} + (n-p-1)c^p.$$

Note that the score test may also be applied to monotone missing data by simply replacing $N_{y_1\cdots y_k+\cdots+}$, $N_{y_1\cdots y_{k-1}+\cdots+}$, $N_{+\cdots+y_k+\cdots+}$, N and $\hat{\pi}^{(p)}_{y_k|y_1\cdots y_{k-1}}$ with $N^{\bullet(k)}_{y_1\cdots y_k+\cdots+}$, $N^{\bullet(k)}_{y_1\cdots y_{k-1}+\cdots+}$, $N^{\bullet(k)}_{+\cdots+y_k+\cdots+}$, $N^{\bullet(k)}$ and $\hat{\pi}^{\bullet(p)}_{y_k|y_1\cdots y_{k-1}}$ respectively in (4.2) and (4.3), where an event count with superscript "$\bullet(k)$" denotes the number of subjects having event listed in the subscript with complete data up to time point k.

However, the score test may not be applied to data with an arbitrary pattern of missingness, since the sample size of the multinomial distribution determined by $y_1,\ldots,y_k$ is not observed. In contrast, the likelihood ratio test can be applied to complete data, monotone missing data and data with arbitrary missing pattern.

4.1.2 Likelihood ratio test and its modification

For testing the null hypothesis of pth-order antedependence against the saturated alternative, the likelihood ratio test statistic, G_p^2, simplifies to twice the log of the ratio of multinomial likelihood kernels evaluated at $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\pi}}^{(p)}$, respectively.

Thus we have

$$G_p^2 = 2 \sum_{k=p+1}^{n} \sum_{(y_1,\ldots,y_k)\in C_k} N_{y_1\cdots y_k+\cdots+} \log\left(\frac{\hat{\pi}_{y_k|y_1\cdots y_{k-1}}}{\hat{\pi}^{(p)}_{y_k|y_1\cdots y_{k-1}}}\right)$$

where $\hat{\pi}^{(p)}_{y_k|y_1\cdots y_{k-1}}$ can be obtained by (4.3) and $\hat{\pi}_{y_k|y_1\cdots y_{k-1}} = \dfrac{N_{y_1\cdots y_{k-1}y_k+\cdots+}}{N_{y_1\cdots y_{k-1}+\cdots+}}$ is the mle of transition probability under the saturated model. Note thus that similarly to the score test, the likelihood ratio test is not affected by zero cell counts. The limiting null distribution of G_p^2 is chi-square with degrees of freedom d_p.

For monotone missing data, the likelihood ratio test statistic is obtained by doing replacements in the test statistic similarly to that in score test. Moreover, the likelihood ratio test can handle data with an arbitrary pattern of missingness, where we replace $N_{y_1\cdots y_k+\cdots+}$, $\hat{\pi}_{y_k|y_1\cdots y_{k-1}}$ and $\hat{\pi}^{(p)}_{y_k|y_1\cdots y_{k-1}}$ with $\hat{N}^{\bullet(k)}_{y_1\cdots y_k+\cdots+}$, $\hat{\pi}^{\star}_{y_k|y_1\cdots y_{k-1}}$ and $\hat{\pi}^{\star(p)}_{y_k|y_1\cdots y_{k-1}}$, which can be obtained numerically by applying Theorem 2.1.4.

As will be shown by simulation in Section 4.1.5, the likelihood ratio test is oversensitive under the null hypothesis. Williams (1976) showed that modifying the likelihood ratio test by multiplying G_p^2 by a multiplier γ_p in order to equate $E(\gamma_p G_p^2)$ to d_p results in a test statistic that is better approximated by its asymptotic χ^2-distribution. Under the null, for each multinomial distribution determined by $y_1, \ldots, y_{k-1}$ with sample size $N_{y_1\cdots y_{k-1}+\cdots+}$ and cell probability $\pi^{(p)}_{y_k|y_1\cdots y_{k-1}}$, the expected cell count is $N_{y_1\cdots y_{k-1}+\cdots+}\pi^{(p)}_{y_k|y_1\cdots y_{k-1}}$ and the marginal distribution of each cell count is binomially distributed with sample size $N_{y_1\cdots y_{k-1}+\cdots+}$ and probability of success $\pi^{(p)}_{y_k|y_1\cdots y_{k-1}}$; similarly, for each multinomial distribution determined by $y_{k-p}, \ldots, y_{k-1}$ with sample size $N_{+\cdots+y_{k-p}\cdots y_{k-1}+\cdots+}$ and cell probability $\pi^{(p)}_{y_k|y_{k-p}\cdots y_{k-1}}$, the expected cell count is $N_{+\cdots+y_{k-p}\cdots y_{k-1}+\cdots+}\pi^{(p)}_{y_k|y_{k-p}\cdots y_{k-1}}$, and the marginal distribution of each cell count is binomially distributed with sample size $N_{+\cdots+y_{k-p}\cdots y_{k-1}+\cdots+}$ and probability of success $\pi^{(p)}_{y_k|y_{k-p}\cdots y_{k-1}}$. We can obtain the multiplier γ_p by following the result derived by Williams (1976, sec. 2.1). For a binomial distribution, let

X, N and π denote the number of successes, sample size and probability of success, respectively; accordingly the expected number of successes is $\mu = N\pi$. Thus, a Taylor expansion of $X \log X$ about $X = \mu$ gives:

$$X \log X = \mu \log \mu + (X - \mu)(1 + \log \mu) + \sum_{r=2}^{\infty} \frac{(-1)^r (X-\mu)^r}{r(r-1)\mu^{r-1}}.$$

The first four central moments of X are

$$\begin{aligned}
\mu_1 &= N\pi = \mu, \\
\mu_2 &= N\pi(1-\pi) = \mu(1-\pi), \\
\mu_3 &= N\pi(1-\pi)(1-2\pi) = \mu(1-\pi)(1-2\pi) \\
\text{and } \mu_4 &= N(-1+\pi)\pi(3N\pi^2 - 6\pi^2 - 3N\pi + 6\pi - 1) \\
&= (-1+\pi)(3\pi\mu^2 - 6\mu\pi^2 - 3\mu^2 + 6\pi\mu - \mu)
\end{aligned}$$

and for $r \geq 2$, μ_{2r} and μ_{2r+1} are $O(\mu^r)$. Thus we have

$$\begin{aligned}
E(X \log X) =& \mu \log \mu + \frac{1-\pi}{2} + \frac{1-\pi^2}{12\mu} + O(\mu^{-2}) \\
\doteq& N\pi \log(N\pi) + \frac{1-\pi}{2} + \frac{1-\pi^2}{12N\pi} \equiv f(N, \pi).
\end{aligned} \tag{4.4}$$

By applying (4.4) into $E(G_p^2)$, we have

$$\begin{aligned}
E(G_p^2) = \sum_{k=p+1}^{n} \Bigg[& \sum_{(y_1,\ldots,y_{k-1}) \in C_{k-1}} \sum_{y_k=1}^{c} 2E\left(N_{y_1 \cdots y_k + \cdots +} \log\left(\hat{\pi}_{y_k | y_1 \cdots y_{k-1}}\right) \right) \\
& - \sum_{(y_{k-p},\ldots,y_{k-1}) \in C_p} \sum_{y_k=1}^{c} 2E\left(N_{+\cdots+y_{k-p}\cdots y_k+\cdots+} \log\left(\hat{\pi}^{(p)}_{y_k | y_{k-p}\cdots y_{k-1}}\right) \right) \Bigg] \\
= \sum_{k=p+1}^{n} \Bigg[& \sum_{(y_1,\ldots,y_{k-1}) \in C_{k-1}} \sum_{y_k=1}^{c} 2E\left(N_{y_1 \cdots y_k + \cdots +} \log\left(\frac{N_{y_1 \cdots y_k + \cdots +}}{N_{y_1 \cdots y_{k-1} + \cdots +}} \right) \right) \\
& - \sum_{(y_{k-p},\ldots,y_{k-1}) \in C_p} \sum_{y_k=1}^{c} 2E\left(N_{+\cdots+y_{k-p}\cdots y_k+\cdots+} \log\left(\frac{N_{+\cdots+y_{k-p}\cdots y_k+\cdots+}}{N_{+\cdots+y_{k-p}\cdots y_{k-1}+\cdots+}} \right) \right) \Bigg]
\end{aligned}$$

where by (4.4), under AD(p) we have

$$
\begin{aligned}
&E\left[N_{y_1\cdots y_k+\cdots+}\log\left(\frac{N_{y_1\cdots y_k+\cdots+}}{N_{y_1\cdots y_{k-1}+\cdots+}}\right)\right]\\
=&E\left(N_{y_1\cdots y_k+\cdots+}\log\left(N_{y_1\cdots y_k+\cdots+}\right)\right)-E\left(N_{y_1\cdots y_k+\cdots+}\log\left(N_{y_1\cdots y_{k-1}+\cdots+}\right)\right)\\
\doteq&f(N,\pi^{(p)}_{y_1\cdots y_k+\cdots+})-E\left[E\left(N_{y_1\cdots y_k+\cdots+}\log\left(N_{y_1\cdots y_{k-1}+\cdots+}\right)\middle|N_{y_1\cdots y_{k-1}+\cdots+}\right)\right]\\
=&f(N,\pi^{(p)}_{y_1\cdots y_k+\cdots+})-\pi^{(p)}_{y_k|y_1\cdots y_{k-1}}E\left(N_{y_1\cdots y_{k-1}+\cdots+}\log\left(N_{y_1\cdots y_{k-1}+\cdots+}\right)\right)\\
=&f(N,\pi^{(p)}_{y_1\cdots y_k+\cdots+})-\pi^{(p)}_{y_k|y_1\cdots y_{k-1}}f(N,\pi^{(p)}_{y_1\cdots y_{k-1}+\cdots+})
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
&E\left[N_{+\cdots+y_{k-p}\cdots y_k+\cdots+}\log\left(\frac{N_{+\cdots+y_{k-p}\cdots y_k+\cdots+}}{N_{+\cdots+y_{k-p}\cdots y_{k-1}+\cdots+}}\right)\right]\\
\doteq&f(N,\pi^{(p)}_{+\cdots+y_{k-p}\cdots y_k+\cdots+})-\pi^{(p)}_{y_k|y_{k-p}\cdots y_{k-1}}f(N,\pi^{(p)}_{+\cdots+y_{k-p}\cdots y_{k-1}+\cdots+}).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
E(G_p^2)\doteq&\sum_{k=p+1}^{n}\sum_{(y_1,\ldots,y_k)\in C_k}\Bigg[f(N,\pi^{(p)}_{y_1\cdots y_k+\cdots+})-\pi^{(p)}_{y_k|y_1\cdots y_{k-1}}f(N,\pi^{(p)}_{y_1\cdots y_{k-1}+\cdots+})\\
&-\left(f(N,\pi^{(p)}_{+\cdots+y_{k-p}\cdots y_k+\cdots+})-\pi^{(p)}_{y_k|y_{k-p}\cdots y_{k-1}}f(N,\pi^{(p)}_{+\cdots+y_{k-p}\cdots y_{k-1}+\cdots+})\right)\Bigg]\\
\equiv&\frac{d_p}{\gamma_p}
\end{aligned}
\tag{4.5}
$$

But since the true probabilities are unknown, we obtain $\hat{\gamma}_p$ in practice by replacing the true probabilities in (4.5) with their mles.

4.1.3 Wald test

Now we consider a Wald test. In contrast to the likelihood ratio and score tests, the Wald test is not invariant to transformations of the parameter space. It is inconvenient to express all nonredundant equations implied by (2.2) in a well organized way to perform a Wald test. Alternatively, we consider another commonly used way to express conditional independence. We parameterize in terms of conditional log odds ratios, or COLORs, where the conditioning is on different

realizations of all intervening variables. More specifically, we define COLORs of lag two and higher as follows:

$$\begin{aligned}\psi_{y_{k-h},y_k|y_{k-h+1}\cdots y_{k-1}} \equiv \quad & \log\left(\frac{P(Y_{k-h}=c,Y_{k-h+1}=y_{k-h+1},\ldots,Y_{k-1}=y_{k-1},Y_k=c)}{P(Y_{k-h}=c,Y_{k-h+1}=y_{k-h+1},\ldots,Y_{k-1}=y_{k-1},Y_k=y_k)}\right.\\ & \left.\times\frac{P(Y_{k-h}=y_{k-h},Y_{k-h+1}=y_{k-h+1},\ldots,Y_{k-1}=y_{k-1},Y_k=y_k)}{P(Y_{k-h}=y_{k-h},Y_{k-h+1}=y_{k-h+1},\ldots,Y_{k-1}=y_{k-1},Y_k=c)}\right)\\ = \quad & \log\pi_{+\cdots+cy_{k-h+1}\cdots y_{k-1}c+\cdots+}+\log\pi_{+\cdots+y_{k-h}y_{k-h+1}\cdots y_{k-1}y_k+\cdots+}\\ & -\log\pi_{+\cdots+cy_{k-h+1}\cdots y_{k-1}y_k+\cdots+}-\log\pi_{+\cdots+y_{k-h}y_{k-h+1}\cdots y_{k-1}c+\cdots+},\end{aligned} \tag{4.6}$$

where $(y_{k-h},y_k)\in\{1,\ldots,c-1\}^2$, $(y_{k-h+1},\ldots,y_{k-1})\in C_{h-1}$, $k=3,4,\ldots,n$, and $h=2,3,\ldots,k-1$. In addition, we define lag-one COLORs (for which there are no intervening variables) by an expression identical to (4.6) but with $h=1$; thus, lag-one COLORs coincide with lag-one log odds ratios. The following lemma helps us to show the equivalence between the total number of conditional log odds ratios and conditional independence under AD(p) implied by (2.2).

Lemma 4.1.1. *For real value $c>1$, we have*

$$(c-1)^2\sum_{k=p}^{n-2}(n-1-k)c^k = c^n-(n-p)c^{p+1}+(n-1-p)c^p$$

Proof.

$$\begin{aligned}&(c-1)^2\sum_{k=p}^{n-2}(n-1-k)c^k\\ &=(c-1)^2\Big[(n-1)\left(c^p+c^{p+1}+\cdots+c^{n-2}\right)-\left(pc^p+(p+1)c^{p+1}+\cdots+(n-2)c^{n-2}\right)\Big]\\ &=(c-1)^2\left[(n-1)\frac{c^{n-1}-c^p}{c-1}-\sum_{k=p}^{n-2}kc^k\right].\end{aligned} \tag{4.7}$$

Now,

$$\sum_{k=p}^{n-2}c^k=\frac{c^{n-1}-c^p}{c-1}. \tag{4.8}$$

By taking derivatives on both sides of (4.8) with respect to c, we have

$$\sum_{k=p}^{n-2} kc^{k-1} = \frac{\left(pc^{p-1} - (n-1)c^{n-2}\right)(1-c) + (c^p - c^{n-1})}{(1-c)^2}. \tag{4.9}$$

Multiplying by c on both sides of (4.9), we have

$$\sum_{k=p}^{n-2} kc^{k} = \frac{\left(pc^{p-1} - (n-1)c^{n-2}\right)(1-c)c + (c^p - c^{n-1})c}{(1-c)^2}. \tag{4.10}$$

By combining (4.7) and (4.10), we have

$$\begin{aligned}
&(c-1)^2 \sum_{k=p}^{n-2} (n-1-k)c^k \\
&= (c-1)^2 \left[(n-1)\frac{c^{n-1} - c^p}{c-1} - \frac{\left(pc^{p-1} - (n-1)c^{n-2}\right)(1-c)c + (c^p - c^{n-1})c}{(1-c)^2} \right] \\
&= c^n - (n-p)c^{p+1} + (n-1-p)c^p \\
&= d_p.
\end{aligned}$$

□

By directly counting the COLORs, we find that their number is $(c-1)^2 \sum_{k=0}^{n-2} (n-1-k)c^k$; algebraic manipulations by applying Lemma 4.1.1 when $p = 0$ yield the alternative expression $c^n - nc + n - 1 = d_0$. Let $\mathbf{\Psi}$ denote the vector of COLORs listed in order from no intervenors (lag one) to $n-2$ intervenors (lag $n-1$).

Under the saturated model, the mle of $\psi_{y_{k-h}, y_k | y_{k-h+1} \cdots y_{k-1}}$ is given by

$$\begin{aligned}
\hat{\psi}_{y_{k-h}, y_k | y_{k-h+1} \cdots y_{k-1}} &= \log(N_{+\cdots+cy_{k-h+1}\cdots y_{k-1}c+\cdots+}) + \log(N_{+\cdots+y_{k-h}y_{k-h+1}\cdots y_{k-1}y_k+\cdots+}) \\
&\quad - \log(N_{+\cdots+cy_{k-h+1}\cdots y_{k-1}y_k+\cdots+}) - \log(N_{+\cdots+y_{k-h}y_{k-h+1}\cdots y_{k-1}c+\cdots+}).
\end{aligned}$$

Let $\hat{\mathbf{\Psi}} \equiv (\hat{\psi}_{y_{k-h}, y_k | y_{k-h+1} \cdots y_{k-1}})$, with ordering of elements identical to that of $\mathbf{\Psi}$.

Lemma 4.1.2. *As $N \to \infty$, $\sqrt{N}(\hat{\mathbf{\Psi}} - \mathbf{\Psi})$ converges in distribution to $N(\mathbf{0}, \mathbf{\Sigma})$, where*

$$\mathbf{\Sigma} = \mathbf{B}[diag(\mathbf{A}\boldsymbol{\pi})]^{-1}\mathbf{A}[diag(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}']\mathbf{A}'[diag(\mathbf{A}\boldsymbol{\pi})]^{-1}\mathbf{B}',$$

where matrices $\mathbf{A}$, $\mathbf{B}$ are defined in the proof.

Proof. Since each COLOR is a function of the cell probabilities, there exists a function $g : R^{c^n} \to R^{d_0}$ mapping $\boldsymbol{\pi}$ to $\boldsymbol{\Psi}$. Specifically, $g \equiv f_1 \circ f_2 \circ f_3$ where:

(a) $f_3 : R^{c^n} \to R^{4d_0}$ maps $\boldsymbol{\pi}$ to a vector whose elements are of the form

$$\pi_{+\cdots+y_{k_1}y_{k_1+1}\cdots y_{k_2-1}y_{k_2}+\cdots+} \tag{4.11}$$

via premultiplication by a matrix $\mathbf{A}_{4d_0\times c^n}$, i.e. $f_3(\boldsymbol{\pi}) \equiv \mathbf{A}\boldsymbol{\pi}$. The $4d_0$ elements in $f_3(\boldsymbol{\pi})$ can be classified into d_0 groups so that the four terms in each COLOR have their own corresponding summation terms of the form of equation (4.11) in vector $f_3(\boldsymbol{\pi})$ with $(y_{k-h}, y_k) \in \{(c,c),(c,j),(i,c),(i,j) : \ i = 1,\ldots,c-1; j = 1,\ldots,c-1\}$. Accordingly, $\mathbf{A}$ is a matrix containing only 0's and 1's in each row, with 1's appearing only when the realized values of the variables between time points $k-h$ and k are of the form $(y_{k-h+1},\ldots,y_{k-1})$. For convenience and consistency, the sums in $f_3(\boldsymbol{\pi})$ are listed in the same order as the COLORs in $\boldsymbol{\Psi}$.

(b) $f_2 : R^{4d_0} \to R^{4d_0}$ is the elementwise log function. For a vector $\mathbf{u}$ with positive elements, $f_2(\mathbf{u}) \equiv [log(\mathbf{u})]$, thus $\dfrac{\partial f_2}{\partial \mathbf{u}} = [diag(\mathbf{u})]^{-1}$, which is a $4d_0 \times 4d_0$ matrix. Here, $\mathbf{u} = \mathbf{A}\boldsymbol{\pi}$ and accordingly $\dfrac{\partial f_2}{\partial f_3} = [diag(\mathbf{A}\boldsymbol{\pi})]^{-1}$.

(c) $f_1 : R^{4d_0} \to R^{d_0}$ maps to $\boldsymbol{\Psi}$, i.e. $f_1(\mathbf{v}) = \mathbf{B}\mathbf{v}$ for some matrix $\mathbf{B}_{d_0\times 4d_0}$ with each row consisting of all zeros except for two $+1$ elements and two -1 elements in the positions occupied by relevant elements of vector $f_2(f_3(\boldsymbol{\pi}))$ to form the given COLOR. More specifically, the two $+1$ and two -1 elements are in the positions where the log of sums have realizations of the form $(y_{k-h+1},\ldots,y_{k-1})$ between time points $k-h$ and k but $(y_{k-h}, y_k) = (c,c),(i,j)$, and $(c,j),(i,c)$ for $i = 1,\ldots,c-1; j = 1,\ldots,c-1$ respectively.

Under AD$(n-1)$, all elements of $\hat{\boldsymbol{\Psi}}$ can be computed by replacing each element in $\boldsymbol{\pi}$ with the corresponding element in $\hat{\boldsymbol{\pi}}$: $\hat{\boldsymbol{\Psi}} = g(\hat{\boldsymbol{\pi}}) = f_1 \circ f_2 \circ f_3(\hat{\boldsymbol{\pi}})$. It is well

Y_1	Y_2	Y_3	$\boldsymbol{\pi}_{toy}$	count
1	1	1	π_{111}	N_{111}
1	1	0	π_{110}	N_{110}
1	0	1	π_{101}	N_{101}
1	0	0	π_{100}	N_{100}
0	1	1	π_{011}	N_{011}
0	1	0	π_{010}	N_{010}
0	0	1	π_{001}	N_{001}
0	0	0	π_{000}	N_{000}

Table 4.1: Toy example for Wald test

known (Agresti 2002, p. 580) that

$$\sqrt{N}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) \rightsquigarrow \mathrm{N}_{c^n}(\mathbf{0}, diag(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'),$$

where $\rightsquigarrow$ denotes convergence in distribution and N_{c^n} denotes multivariate normality of dimension c^n. Then by the Delta method,

$$\begin{aligned}\sqrt{N}(\hat{\boldsymbol{\Psi}} - \boldsymbol{\Psi}) = \sqrt{N}[g(\hat{\boldsymbol{\pi}}) - g(\boldsymbol{\pi})] &\rightsquigarrow \mathrm{N}_{d_0}(\mathbf{0}, (\frac{\partial g}{\partial \boldsymbol{\pi}})(diag(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}')(\frac{\partial g}{\partial \boldsymbol{\pi}})') \\ &\equiv \mathrm{N}_{d_0}(\mathbf{0}, \boldsymbol{\Sigma}),\end{aligned}$$

where $\dfrac{\partial g}{\partial \boldsymbol{\pi}} = \dfrac{\partial f_1}{\partial f_2}\dfrac{\partial f_2}{\partial f_3}\dfrac{\partial f_3}{\partial \boldsymbol{\pi}} = \mathbf{B}[diag(\mathbf{A}\boldsymbol{\pi})]^{-1}\mathbf{A}$, i.e.

$$\sqrt{N}(\hat{\boldsymbol{\Psi}} - \boldsymbol{\Psi}) \rightsquigarrow \mathrm{N}_{d_0}(\mathbf{0}, \boldsymbol{\Sigma}),$$

$$\text{with } \boldsymbol{\Sigma} = \mathbf{B}[diag(\mathbf{A}\boldsymbol{\pi})]^{-1}\mathbf{A}[diag(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}']\mathbf{A}'[diag(\mathbf{A}\boldsymbol{\pi})]^{-1}\mathbf{B}'.$$

□

Now consider a simple toy example shown in Table 4.1 to apply Lemma 4.1.2. This toy example has a binary outcome so there are four conditional log odds ratios

in $\boldsymbol{\Psi}_{toy} = [\psi_{y_1=0,y_2=0|},\ \psi_{y_2=0,y_3=0|},\ \psi_{y_1=0,y_3=0|y_2=0},\ \psi_{y_1=0,y_3=0|y_2=0}]'$ where

$$\begin{aligned}
\psi_{y_1=0,y_2=0|} \equiv & COLOR(Y_1=0,Y_2=0) = \log\left(\frac{(\pi_{111}+\pi_{110})\times(\pi_{001}+\pi_{000})}{(\pi_{101}+\pi_{100})\times(\pi_{011}+\pi_{010})}\right) \\
= & \log(\pi_{111}+\pi_{110}) + \log(\pi_{001}+\pi_{000}) - \log(\pi_{101}+\pi_{100}) - \log(\pi_{011}+\pi_{010}); \\
\psi_{y_2=0,y_3=0|} \equiv & COLOR(Y_2=0,Y_3=0) = \log\left(\frac{(\pi_{111}+\pi_{011})\times(\pi_{100}+\pi_{000})}{(\pi_{110}+\pi_{010})\times(\pi_{101}+\pi_{001})}\right) \\
= & \log(\pi_{111}+\pi_{011}) + \log(\pi_{100}+\pi_{000}) - \log(\pi_{110}+\pi_{010}) - \log(\pi_{101}+\pi_{001}); \\
\psi_{y_1=0,y_3=0|y_2=0} \equiv & COLOR(Y_1=0,Y_3=0|Y_2=0) = \log\left(\frac{\pi_{101}\pi_{000}}{\pi_{100}\pi_{001}}\right) \\
= & \log(\pi_{101}) + \log(\pi_{000}) - \log(\pi_{100}) - \log(\pi_{001}); \\
\psi_{y_1=0,y_3=0|y_2=0} \equiv & COLOR(Y_1=0,Y_3=0|Y_2=1) = \log\left(\frac{\pi_{111}\pi_{010}}{\pi_{110}\pi_{011}}\right) \\
= & \log(\pi_{111}) + \log(\pi_{010}) - \log(\pi_{110}) - \log(\pi_{011})
\end{aligned}$$

For this toy example,

$$\mathbf{A}_{toy} = \begin{bmatrix}
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\
1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\
 & & & \mathbf{I}_{8\times 8} & & & &
\end{bmatrix}$$

so that

$$\begin{aligned}
& f_3(\boldsymbol{\pi}_{toy}) = \mathbf{A}_{toy}\boldsymbol{\pi}_{toy} \\
= & \Big[\pi_{111}+\pi_{110}, \pi_{101}+\pi_{100}, \pi_{011}+\pi_{010}, \pi_{001}+\pi_{000}, \pi_{111}+\pi_{011}, \pi_{110}+\pi_{010}, \pi_{101}+\pi_{001}, \\
& \pi_{100}+\pi_{000}, \pi_{111}, \pi_{110}, \pi_{101}, \pi_{100}, \pi_{011}, \pi_{010}, \pi_{001}, \pi_{000}\Big]'.
\end{aligned}$$

Next we have

$$
\begin{aligned}
&f_2(f_3(\boldsymbol{\pi}_{toy})) = \log(f_3(\boldsymbol{\pi}_{toy})) \\
=&\Big[\log(\pi_{111}+\pi_{110}), \log(\pi_{101}+\pi_{100}), \log(\pi_{011}+\pi_{010}), \log(\pi_{001}+\pi_{000}), \log(\pi_{111}+\pi_{011}), \\
&\log(\pi_{110}+\pi_{010}), \log(\pi_{101}+\pi_{001}), \log(\pi_{100}+\pi_{000}), \log(\pi_{111}), \log(\pi_{110}), \log(\pi_{101}), \\
&\log(\pi_{100}), \log(\pi_{011}), \log(\pi_{010}), \log(\pi_{001}), \log(\pi_{000})\Big]'.
\end{aligned}
$$

Finally, let

$$
\mathbf{B}_{toy} = \begin{bmatrix}
1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0
\end{bmatrix}
$$

and we have

$$
f_1(f_2(f_3(\boldsymbol{\pi}_{toy}))) = \mathbf{B}_{toy} f_2(f_3(\boldsymbol{\pi}_{toy})) = \mathbf{B}_{toy}\log(f_3(\boldsymbol{\pi}_{toy})) = \mathbf{B}_{toy}\log(\mathbf{A}_{toy}\boldsymbol{\pi}_{toy}) = \boldsymbol{\Psi}_{toy}.
$$

It is well known that

$$
\sqrt{N}(\hat{\boldsymbol{\pi}}_{toy} - \boldsymbol{\pi}_{toy}) \rightsquigarrow \mathrm{N}_8(\mathbf{0}, diag(\boldsymbol{\pi}_{toy}) - \boldsymbol{\pi}_{toy}\boldsymbol{\pi}'_{toy}),
$$

where we apply the Delta method to conclude the result indicated by Lemma 4.1.2 for this toy example.

By Lemma 4.1.2, we obtain the limiting distribution of all conditional log odds ratios. Under an AD(p) model (with $p < n-1$), all COLORs of lag $p+1$ and higher are equal to zero, as is easily verified. Consequently, pth-order antedependence is equivalent to the condition $\mathbf{M}_p\boldsymbol{\Psi} = \mathbf{0}$, where

$$
\mathbf{M}_p \equiv \left[\ \mathbf{0}_{d_p\times(d_0-d_p)} \ \middle|\ \mathbf{I}_{d_p\times d_p}\ \right]
$$

and $\mathbf{0}$ and $\mathbf{I}$ are null and identity matrices respectively. Then the Wald test statistic is given by

$$
T_p^2 \equiv N(\mathbf{M}_p\hat{\boldsymbol{\Psi}})'(\mathbf{M}_p\hat{\boldsymbol{\Sigma}}\mathbf{M}_p')^{-1}(\mathbf{M}_p\hat{\boldsymbol{\Psi}}). \tag{4.12}
$$

Here, $\hat{\mathbf{\Sigma}}$ is the mle of $\mathbf{\Sigma}$, which is given by

$$\hat{\mathbf{\Sigma}} = \mathbf{B}[diag(\mathbf{A}\hat{\boldsymbol{\pi}})]^{-1}\mathbf{A} \times [diag(\hat{\boldsymbol{\pi}}) - \hat{\boldsymbol{\pi}}\hat{\boldsymbol{\pi}}'] \times \mathbf{A}'[diag(\mathbf{A}\hat{\boldsymbol{\pi}})]^{-1}\mathbf{B}',$$

where all the diagonal elements of $\hat{\mathbf{\Sigma}}$ are of the well known form of the asymptotic variance of the log odds ratio. For example, the diagonal elements of $\hat{\mathbf{\Sigma}}_{toy}$ are

$$\begin{aligned}
\hat{\mathbf{\Sigma}}_{toy11} &= \frac{1}{\hat{\pi}_{111}+\hat{\pi}_{110}} + \frac{1}{\hat{\pi}_{001}+\hat{\pi}_{000}} + \frac{1}{\hat{\pi}_{101}+\hat{\pi}_{100}} + \frac{1}{\hat{\pi}_{011}+\hat{\pi}_{010}};\\
\hat{\mathbf{\Sigma}}_{toy22} &= \frac{1}{\hat{\pi}_{111}+\hat{\pi}_{011}} + \frac{1}{\hat{\pi}_{100}+\hat{\pi}_{000}} + \frac{1}{\hat{\pi}_{110}+\hat{\pi}_{010}} + \frac{1}{\hat{\pi}_{101}+\hat{\pi}_{001}},\\
\hat{\mathbf{\Sigma}}_{toy33} &= \frac{1}{\hat{\pi}_{101}} + \frac{1}{\hat{\pi}_{000}} + \frac{1}{\hat{\pi}_{100}} + \frac{1}{\hat{\pi}_{001}} \text{ and}\\
\hat{\mathbf{\Sigma}}_{toy44} &= \frac{1}{\hat{\pi}_{111}} + \frac{1}{\hat{\pi}_{010}} + \frac{1}{\hat{\pi}_{110}} + \frac{1}{\hat{\pi}_{011}}.
\end{aligned}$$

It follows from Lemma 4.1.2 and Slutsky's Theorem that the limiting null distribution of T_p^2 is chi-square with d_p degrees of freedom. However, the Wald test is affected by empty cells. If any cell counts are zero, 0.5 may be added to each cell count to avoid zero denominators in the mles of COLORs.

4.1.4 Adaptation of Freeman and Halton's exact test

In addition to the likelihood ratio, score, and Wald tests, we included a testing procedure adapted from Freeman and Halton's (1951) exact test for independence between row and column factors in a two-way table (Agresti 2002 p. 97). The adaptation tests for *conditional* independence of row and column factors in the $c \times c$ table corresponding to each pair of variables lagged $p+1$ or more times apart and each realization of their intervenors. There are $\sum_{k=p}^{n-2}(n-1-k)c^k$ such exact tests for conditional independence. The Bonferroni inequality ensures that performing each of these tests at a level equal to the nominal size divided by $\sum_{k=p}^{n-2}(n-1-k)c^k$ yields an overall Type I error probability no larger than the specified nominal size. We apply Bonferroni's inequality to Freeman and Halton's exact test rather than other

tests because Freeman and Halton's exact test gives the exact P-value for each $c \times c$ table.

For illustration, consider the toy example shown in Table 4.1 again. Suppose we wish to test AD(1) against AD(2). Then we would test for $Y_3 \perp Y_1 | Y_2$ using the 2×2 tables corresponding to both $Y_2 = 0$ and $Y_2 = 1$. Table 4.2 displays these tables and their cell counts. If we fail to reject both $Y_3 \perp Y_1 | (Y_2 = 1)$ and $Y_3 \perp Y_1 | (Y_2 = 0)$ by Fisher's exact tests, each at a "Bonferronized" significance level of 0.025, then we conclude, with type I error probability at most 0.05, that the variables are AD(1) at most. Otherwise, the variables are concluded to be AD(2). Because this approach to testing for p^{th}-order antedependence requires that Fisher's exact test be performed on each individual 2×2 table with a Bonferroni adjustment for multiplicity, it is conservative, especially when n is large. Consequently, it is expected that this approach will not be as powerful as the likelihood-based approaches described previously.

$Y_2 = 1$	$Y_1 = 1$	$Y_1 = 0$
$Y_3 = 1$	N_{111}	N_{011}
$Y_3 = 0$	N_{110}	N_{010}
$Y_2 = 0$	$Y_1 = 1$	$Y_1 = 0$
$Y_3 = 1$	N_{101}	N_{001}
$Y_3 = 0$	N_{100}	N_{000}

Table 4.2: Table 4.1 partitioned into two 2×2 tables for different values of Y_2

4.1.5 Simulation study

A simulation study was performed to examine and compare the performance of the proposed tests for pth-order antedependence versus the saturated model. Simulations were generated from a "benchmark" binary linear process and two variations of it. The binary linear process is as follows:

$$Y_1 = \begin{cases} 1 & \text{w.p. } \frac{1}{2} \\ 0 & \text{w.p. } \frac{1}{2} \end{cases} ; \; Y_2|Y_1 = \begin{cases} Y_1 & \text{w.p. } \frac{1}{3} \\ 1 & \text{w.p. } \frac{1}{3} \\ 0 & \text{w.p. } \frac{1}{3} \end{cases} ;$$

$$Y_3|(Y_1,Y_2) = \begin{cases} Y_1 & \text{w.p. } \frac{\theta_1}{4} \\ Y_2 & \text{w.p. } \frac{2-\theta_1}{4} \\ 1 & \text{w.p. } \frac{1}{4} \\ 0 & \text{w.p. } \frac{1}{4} \end{cases} ; \; Y_4|(Y_1,Y_2,Y_3) = \begin{cases} Y_1 & \text{w.p. } \frac{\theta_1}{8} \\ Y_2 & \text{w.p. } \frac{\theta_1}{8} \\ Y_3 & \text{w.p. } \frac{2-\theta_1}{4} \\ 1 & \text{w.p. } \frac{1}{4} \\ 0 & \text{w.p. } \frac{1}{4} \end{cases} \tag{4.13}$$

Here $\theta_1 \in [0,1]$ controls the degree of departure from first-order antedependence: when $\theta_1 = 0$ the process is AD(1), when $\theta_1 > 0$ the process is AD(3), and as θ_1 increases to 1 the departure from AD(1) is larger in the sense, for example, that the COLORs of lags two and three increase. Note that the marginal distribution of Y_k is Bernoulli($\frac{1}{2}$) for all k, regardless of θ_1.

The first variation is a nonlinear binary process, which differs from the benchmark only by redefining Y_3 and Y_4 as follows:

$$Y_3|(Y_1,Y_2) = \begin{cases} Y_1Y_2 & \text{w.p. } \frac{\theta_1}{4} \\ Y_2 & \text{w.p. } \frac{2-\theta_1}{4} \\ 1 & \text{w.p. } \frac{1}{4} \\ 0 & \text{w.p. } \frac{1}{4} \end{cases} , \; Y_4|(Y_1,Y_2,Y_3) = \begin{cases} Y_1Y_2Y_3 & \text{w.p. } \frac{\theta_1}{8} \\ Y_2Y_3 & \text{w.p. } \frac{\theta_1}{8} \\ Y_3 & \text{w.p. } \frac{2-\theta_1}{4} \\ 1 & \text{w.p. } \frac{1}{4} \\ 0 & \text{w.p. } \frac{1}{4} \end{cases} \tag{4.14}$$

This process likewise is AD(1) when $\theta_1 = 0$, and is AD(3) when $\theta_1 > 0$.

The second variation is a binary linear process that is AD(2) [rather than AD(1)] when $\theta_1 = 0$. It differs from the benchmark only by redefining Y_3 and Y_4 as follows:

$$Y_3|(Y_1, Y_2) = \begin{cases} Y_1 & \text{w.p. } \frac{1}{4} \\ Y_2 & \text{w.p. } \frac{1}{4} \\ 1 & \text{w.p. } \frac{1}{4} \\ 0 & \text{w.p. } \frac{1}{4} \end{cases}, \quad Y_4|(Y_1, Y_2, Y_3) = \begin{cases} Y_1 & \text{w.p. } \frac{\theta_1}{4} \\ Y_2 & \text{w.p. } \frac{2-\theta_1}{8} \\ Y_3 & \text{w.p. } \frac{2-\theta_1}{8} \\ 1 & \text{w.p. } \frac{1}{4} \\ 0 & \text{w.p. } \frac{1}{4} \end{cases} \tag{4.15}$$

For each process, each of three sample sizes — $N = 50$ ("small"), $N = 200$ ("medium"), and $N = 1000$ ("large") — and each θ_1 over a range of values, 10,000 samples were generated. Rejection rates for selected values of θ_1 from applying the three likelihood-based tests of AD(1) versus AD(3) for (4.13) and (4.14) and AD(2) versus AD(3) for (4.15), and 95% Wald-based confidence limits of sizes (rejection rate when $\theta_1 = 0$), are displayed in Table 4.3. The adaptation of Freeman and Halton's exact test has also been applied in the simulation but its rejection rates are not listed in Table 4.3. As expected, it was not as powerful as the score and likelihood ratio tests, though it was occasionally more powerful than the Wald test. Among all the likelihood based tests, the Wald test is weakest and its test statistic converges to its limiting distribution more slowly than the likelihood ratio and score test statistics. Thus, we do not consider the Wald test any further in this thesis. The score test performs well for complete data: its empirical sizes match the nominal sizes, it is powerful with test statistics converging to their limiting distribution fast and it requires no adjustment for zero cell counts.

However, for data with arbitrary missing pattern, the score test cannot be used and the likelihood ratio test is a good remedy. Since the likelihood ratio test is oversensitive under the null hypothesis, as was shown in Table 4.3, we use the

modified likelihood ratio test described in Section 4.1.2 and compare its rejection rates by simulation on processes (4.13), (4.14) and (4.15) with those of the score and unmodified likelihood ratio tests. Empirical rejection rates of the score, likelihood ratio and modified likelihood ratio tests, are listed in Table 4.4, where the modified likelihood ratio test is denoted as "LRT1". Also listed are the 95% Wald-based confidence limits of the sizes (rejection rate when $\theta_1 = 0$) of the tests. For the modified likelihood ratio test, we used both γ_p and $\hat{\gamma}_p$ as multipliers; the results shown in Table 4.4 correspond to the use of $\hat{\gamma}_p$, but using γ_p did not yield substantially different results.

By comparison of the rejection rates of the three tests on processes (4.13), (4.14) and (4.15), we see that the score test is the best for moderate and large sample sizes while the modified likelihood ratio test can be considered an improvement when the sample size is small ($N = 50$) as its empirical size is less than 0.05, while it is more powerful than score test. To compare the rejection rates among Wald, score, likelihood ratio and modified likelihood ratio tests visually, we plot the empirical rejection rates against θ_1 in Figure 4.1, where the ones for (4.13) (first), (4.14)(second) and (4.15)(third) are listed in the top, middle and bottom rows and those for small ($N = 50$), medium ($N = 200$) and large ($N = 1000$) sample sizes are listed in the left, middle and right columns.

N	θ_1	(4.13)			(4.14)			(4.15)		
		Wald	LRT	Score	Wald	LRT	Score	Wald	LRT	Score
50	0	0.002	0.073	0.019	0.002	0.073	0.019	0.003	0.097	0.017
	0.2	0.004	0.09	0.026	0.006	0.1	0.029	0.003	0.111	0.023
	0.4	0.005	0.108	0.031	0.013	0.138	0.044	0.009	0.14	0.040
	0.6	0.01	0.141	0.047	0.029	0.202	0.076	0.014	0.183	0.063
	0.8	0.016	0.184	0.07	0.064	0.285	0.119	0.023	0.24	0.094
	1	0.026	0.248	0.093	0.119	0.396	0.187	0.047	0.322	0.147
	LB	0.001	0.068	0.016	0.001	0.068	0.016	0.002	0.092	0.015
	UB	0.003	0.078	0.022	0.003	0.078	0.022	0.004	0.103	0.020
200	0	0.02	0.068	0.037	0.02	0.068	0.037	0.028	0.063	0.047
	0.2	0.045	0.087	0.067	0.065	0.102	0.087	0.058	0.091	0.087
	0.4	0.1	0.149	0.135	0.199	0.231	0.232	0.149	0.186	0.193
	0.6	0.208	0.275	0.263	0.444	0.47	0.476	0.304	0.351	0.360
	0.8	0.372	0.444	0.443	0.729	0.745	0.75	0.523	0.565	0.577
	1	0.586	0.649	0.656	0.917	0.924	0.924	0.741	0.772	0.780
	LB	0.017	0.063	0.033	0.017	0.063	0.033	0.025	0.059	0.043
	UB	0.022	0.073	0.041	0.022	0.073	0.041	0.031	0.068	0.051
1000	0	0.044	0.053	0.051	0.044	0.053	0.051	0.047	0.055	0.053
	0.2	0.161	0.159	0.17	0.286	0.277	0.293	0.205	0.2	0.209
	0.4	0.571	0.563	0.581	0.878	0.865	0.871	0.689	0.682	0.692
	0.6	0.927	0.926	0.931	0.999	0.999	0.999	0.973	0.97	0.973
	0.8	0.997	0.998	0.998	1	1	1	0.999	0.999	0.999
	1	1	1	1	1	1	1	1	1	1.000
	LB	0.04	0.049	0.046	0.04	0.049	0.046	0.043	0.051	0.049
	UB	0.048	0.058	0.055	0.048	0.058	0.055	0.051	0.059	0.058

Table 4.3: Rejection rates by Triad (Wald, likelihood ratio and score tests)

N	θ_1	(4.13)			(4.14)			(4.15)		
		Score	LRT	LRT1	Score	LRT	LRT1	Score	LRT	LRT1
50	0	0.019	0.073	0.029	0.019	0.073	0.029	0.017	0.097	0.038
	0.2	0.026	0.090	0.042	0.029	0.100	0.047	0.023	0.111	0.044
	0.4	0.031	0.108	0.051	0.044	0.138	0.068	0.040	0.140	0.062
	0.6	0.047	0.141	0.076	0.076	0.202	0.115	0.063	0.183	0.087
	0.8	0.070	0.184	0.102	0.119	0.285	0.181	0.094	0.240	0.123
	1	0.093	0.248	0.149	0.187	0.396	0.268	0.147	0.322	0.185
	LB	0.016	0.068	0.026	0.016	0.068	0.026	0.015	0.092	0.034
	UB	0.022	0.078	0.032	0.022	0.078	0.032	0.020	0.103	0.041
200	0	0.037	0.068	0.044	0.037	0.068	0.044	0.047	0.063	0.049
	0.2	0.067	0.087	0.061	0.087	0.102	0.073	0.087	0.091	0.073
	0.4	0.135	0.149	0.115	0.232	0.231	0.184	0.193	0.186	0.160
	0.6	0.263	0.275	0.219	0.476	0.470	0.406	0.360	0.351	0.310
	0.8	0.443	0.444	0.378	0.750	0.745	0.687	0.577	0.565	0.524
	1	0.656	0.649	0.588	0.924	0.924	0.895	0.780	0.772	0.738
	LB	0.033	0.063	0.040	0.033	0.063	0.040	0.043	0.059	0.044
	UB	0.041	0.073	0.048	0.041	0.073	0.048	0.051	0.068	0.053
1000	0	0.051	0.053	0.051	0.051	0.053	0.051	0.053	0.055	0.051
	0.2	0.170	0.159	0.156	0.293	0.277	0.272	0.209	0.200	0.196
	0.4	0.581	0.563	0.557	0.871	0.865	0.862	0.692	0.682	0.675
	0.6	0.931	0.926	0.923	0.999	0.999	0.999	0.973	0.970	0.968
	0.8	0.998	0.998	0.997	1	1	1	0.999	0.999	0.999
	1	1	1	1.000	1	1	1	1	1	1
	LB	0.046	0.049	0.047	0.046	0.049	0.047	0.049	0.051	0.046
	UB	0.055	0.058	0.055	0.055	0.058	0.055	0.058	0.059	0.055

Table 4.4: Rejection rates by modified likelihood ratio test (LRT1)

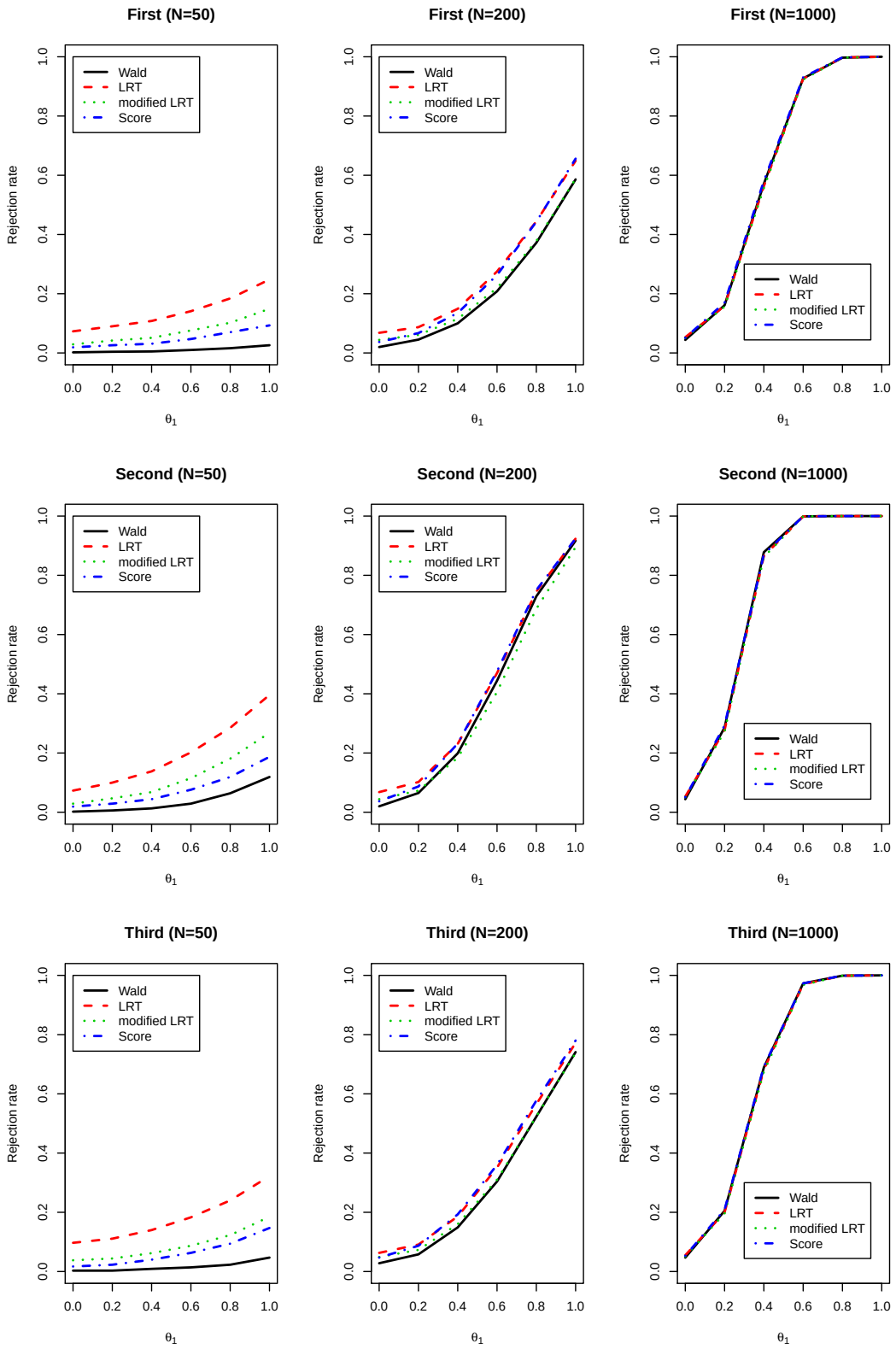


Figure 4.1: Empirical rejection rate curves for (4.13), (4.14) and (4.15)

4.2 AD(p) versus AD(q) for $0 \leq p < q \leq n-2$

The likelihood function under AD(q) is given by

$$\prod_{(y_1,\ldots,y_n)\in C_n} \left[\pi_{y_1\cdots y_p+\cdots+} \left(\prod_{j=p+1}^{q} \pi_{y_j|y_1\cdots y_{j-1}}\right)\left(\prod_{k=q+1}^{n} \pi_{y_k|y_1\cdots y_{k-1}}\right)\right]^{N_{y_1,\ldots,y_n}}$$

$$= \prod_{(y_1,\ldots,y_p)\in C_p} \pi_{y_1\cdots y_p+\cdots+}^{N_{y_1\cdots y_p+\cdots+}} \prod_{j=p+1}^{q} \prod_{(y_1,\ldots,y_{j-1})\in C_{j-1}} \prod_{y_j=1}^{c} \pi_{y_j|y_1\cdots y_{j-1}}^{N_{y_1\cdots y_{j-1}y_j+\cdots+}}$$

$$\times \prod_{k=q+1}^{n} \prod_{(y_{k-q},\ldots,y_{k-1})\in C_q} \prod_{y_k=1}^{c} \pi_{y_k|y_{k-q}\cdots y_{k-1}}^{N_{+\cdots+y_{k-q}\cdots y_{k-1}y_k+\cdots+}} \tag{4.16}$$

Note that (4.16) is the product of the kernels of mutually independent saturated multinomial distribution at different time points. Each individual multinomial kernel is parametrized by transition probabilities at that time point conditioning on different combinations of the realizations from lag-1 to lag-q, which are subject to only the constraint that they sum up to one. The vectors of transition probabilities at different time points conditioning on different realizations of the past p variables are distinct from each other. This makes the information matrix corresponding to (4.16) be diagonal block, resulting that the corresponding score test statistic is just the summation of Pearson's chi-square statistics on each of those multinomial distribution parametrized by transition probabilities. In the remainder of this thesis, all the score test statistics under nonsaturated alternative, such as AD($q_1, \ldots, q_n$) in Sections 4.3 and 4.4 and AD(p) in Chapter 5, are derived based on this idea.

Thus, the score (Pearson chi-square) statistic for testing the null hypothesis of pth-order antedependence against the alternative hypothesis of qth-order antedependence is

$$X^2_{p,q} = \sum_{j=p+1}^{q} \sum_{(y_1\cdots y_{j-1})\in C_{j-1}} \sum_{y_j=1}^{c} \frac{\left(N_{y_1\cdots y_{j-1}y_j+\cdots+} - N_{y_1\cdots y_{j-1}+\cdots+}\hat{\pi}^{(p)}_{y_j|y_1\cdots y_{j-1}}\right)^2}{N_{y_1\cdots y_{j-1}+\cdots+}\hat{\pi}^{(p)}_{y_j|y_1\cdots y_{j-1}}} +$$
$$\sum_{k=q+1}^{n} \sum_{(y_{k-q}\cdots y_{k-1})\in C_q} \sum_{y_k=1}^{c} \frac{\left(N_{+\cdots+y_{k-q}\cdots y_{k-1}y_k+\cdots+} - N_{+\cdots+y_{k-q}\cdots y_{k-1}+\cdots+}\hat{\pi}^{(p)}_{y_k|y_{k-q}\cdots y_{k-1}}\right)^2}{N_{+\cdots+y_{k-q}\cdots y_{k-1}+\cdots+}\hat{\pi}^{(p)}_{y_k|y_{k-q}\cdots y_{k-1}}}, \tag{4.17}$$

where by the AD(p) property,

$$\hat{\pi}^{(p)}_{y_j|y_1\cdots y_{j-1}} = I(p=0)\hat{\pi}^{(0)}_{+\cdots+y_j+\cdots+} + I(p\geq 1)\hat{\pi}^{(p)}_{y_j|y_{j-p}\cdots y_{j-1}}$$
$$\text{and } \hat{\pi}^{(p)}_{y_k|y_{k-q}\cdots y_{k-1}} = I(p=0)\hat{\pi}^{(0)}_{+\cdots+y_k+\cdots+} + I(p\geq 1)\hat{\pi}^{(p)}_{y_k|y_{k-p}\cdots y_{k-1}}$$

and they can both be obtained by (4.3). Note that $X^2_{p,n-1} = X^2_p$ where $q = n-1$ and $k = n$ in (4.17). By the additivity property of independent chi-square random variables, the limiting null distribution of $X^2_{p,q}$ is chi-square with $d_q - d_p$ degrees of freedom.

The likelihood ratio statistic for testing the same hypotheses is given by

$$G^2_{p,q} \equiv 2\left[\sum_{j=p+1}^{q} \sum_{(y_1,\ldots,y_{j-1})\in C_{j-1}} \sum_{y_j=1}^{c} N_{y_1\cdots y_{j-1}y_j+\cdots+} \log\left(\frac{\hat{\pi}^{(q)}_{y_j|y_1\cdots y_{j-1}}}{\hat{\pi}^{(p)}_{y_j|y_1\cdots y_{j-1}}}\right) + \right.$$
$$\left. \sum_{k=q+1}^{n} \sum_{(y_{k-q},\ldots,y_k)\in C_{q+1}} N_{+\cdots+y_{k-q}\cdots y_k+\cdots+} \log\left(\frac{\hat{\pi}^{(q)}_{y_k|y_{k-q}\cdots y_{k-1}}}{\hat{\pi}^{(p)}_{y_k|y_{k-q}\cdots y_{k-1}}}\right)\right]$$

and the limiting null distribution of $G^2_{p,q}$ is also chi-square with $d_q - d_p$ degrees of freedom. A Wald test for this purpose is tedious and we do not consider it. Similarly to the modification of the likelihood ratio test developed for AD(p) versus AD($n-1$) in Section 4.1.2, we can find the multiplier $\gamma_{p,q}$ so that $\gamma_{p,q}G^2_{p,q}$ is better approximated by its asymptotic χ^2-distribution. Note that since algebraically $G^2_{p,q} = G^2_p - G^2_q$, by

(4.5) we have

$$
\begin{aligned}
E(G^2_{p,q}) =& E(G^2_p) - E(G^2_q) \\
=& \Bigg[\sum_{k=p+1}^{n} \sum_{(y_1,\ldots,y_k)\in C_k} \Bigg[f(N, \pi^{(p)}_{y_1\cdots y_k+\cdots+}) - \pi^{(p)}_{y_k|y_1\cdots y_{k-1}} f(N, \pi^{(p)}_{y_1\cdots y_{k-1}+\cdots+}) \\
& - \Big(f(N, \pi^{(p)}_{+\cdots+y_{k-p}\cdots y_k+\cdots+}) - \pi^{(p)}_{y_k|y_{k-p}\cdots y_{k-1}} f(N, \pi^{(p)}_{+\cdots+y_{k-p}\cdots y_{k-1}+\cdots+}) \Big) \Bigg] \Bigg] \\
& - \Bigg[\sum_{k=q+1}^{n} \sum_{(y_1,\ldots,y_k)\in C_k} \Bigg[f(N, \pi^{(q)}_{y_1\cdots y_k+\cdots+}) - \pi^{(q)}_{y_k|y_1\cdots y_{k-1}} f(N, \pi^{(q)}_{y_1\cdots y_{k-1}+\cdots+}) \\
& - \Big(f(N, \pi^{(q)}_{+\cdots+y_{k-q}\cdots y_k+\cdots+}) - \pi^{(q)}_{y_k|y_{k-q}\cdots y_{k-1}} f(N, \pi^{(q)}_{+\cdots+y_{k-q}\cdots y_{k-1}+\cdots+}) \Big) \Bigg] \Bigg] \\
\equiv& \frac{d_q - d_p}{\gamma_{p,q}},
\end{aligned}
\tag{4.18}
$$

where the estimated multiplier $\hat{\gamma}_{p,q}$ can be obtained by replacing the conditional probabilities in (4.18) with their mles.

The case $q = p+1$ is worthy of special mention. It is clear that the likelihood ratio and score test criteria for the sequence of tests of AD(p) versus AD($p+1$) ($p = 0, 1, \ldots, n-2$) sum to the likelihood ratio and score test criteria for AD(0) versus AD($n-1$). Thus, this sequence can be viewed as a decomposition of the likelihood ratio and score tests for complete independence [AD(0)] versus arbitrary dependence [AD($n-1$)] into steps of degree one. Furthermore, the sequence suggests two practical strategies for selecting the order of antedependence for a set of data. In practice, to determine the constant order of antedependence, we may mimic the selection of order of a polynomial model for the mean structure of a regression model: one may use either a forward selection strategy [starting by testing AD(0) versus AD($n-1$) (or AD(0) versus AD(1)), and if AD(0) is rejected then testing AD(1) versus AD($n-1$) (or AD(1) versus AD(2)), etc.] or a backward elimination strategy [starting with a test of AD($n-2$) versus AD($n-1$), and if AD($n-2$) is not

rejected then testing AD($n-3$) versus AD($n-1$) (or AD($n-3$) versus AD($n-2$)), etc.].

From an overall model selection perspective, how should the hypothesis tests for constant order of antedependence be combined with the use of penalized likelihood criteria? Since a constant-order AD model is just a special case of a variable-order AD model, it seems sensible to first select the best variable-order AD model using a penalized likelihood criterion. If this model is of constant order or nearly so, it may then be reasonable to compare it to selected constant-order AD models via hypothesis tests. Especially when different penalized likelihood criteria conclude different AD model, as is shown in Sections 6.1 and 6.2, hypothesis tests can further help to determine the order of antedependence.

4.3 Nested variable-order AD models

Consider testing AD($p_1, \ldots, p_n$) versus AD($q_1, \ldots, q_n$) where $p_k \leq q_k$ for all k and the inequality is strict for at least one k. Score and likelihood ratio for these hypotheses are straightforward extensions of those for AD(p) versus AD(q) given previously.

For example, based on likelihood (2.6), since

$$\sum_{y_1=1}^{c} \frac{\left(N_{y_1+\cdots+} - N\hat{\pi}^{(p_1,\ldots,p_n)}_{y_1+\cdots+}\right)^2}{N\hat{\pi}^{(p_1,\ldots,p_n)}_{y_1+\cdots+}} = 0,$$

$$\sum_{y_k=1}^{c} \frac{\left(N_{+\cdots+y_k+\cdots+} - N\hat{\pi}^{(p_1,\ldots,p_n)}_{+\cdots+y_k+\cdots+}\right)^2}{N\hat{\pi}^{(p_1,\ldots,p_n)}_{+\cdots+y_k+\cdots+}} = 0 \text{ when } q_k = 0 \text{ (which implies } p_k = 0),$$

and

$$\sum_{y_k=1}^{c} \frac{\left(N_{+\cdots+y_{k-q_k}\cdots y_{k-1}y_k+\cdots+} - N_{+\cdots+y_{k-q_k}\cdots y_{k-1}+\cdots+}\hat{\pi}^{(p_1,\ldots,p_n)}_{y_k|y_{k-q_k}\cdots y_{k-1}}\right)^2}{N_{+\cdots+y_{k-q_k}\cdots y_{k-1}+\cdots+}\hat{\pi}^{(p_1,\ldots,p_n)}_{y_k|y_{k-q_k}\cdots y_{k-1}}} = 0 \text{ when } q_k = p_k > 0,$$

the score test (Pearson chi-square) statistic simplifies to

$$X^2_{(p_1,\ldots,p_n),(q_1,\ldots,q_n)} = \sum_{k=2}^{n}\Bigg(I(q_k > p_k) \sum_{(y_{k-q_k},\ldots,y_{k-1})\in C_{q_k}} \sum_{y_k=1}^{c} \frac{\left(N_{+\cdots+y_{k-q_k}\cdots y_{k-1}y_k+\cdots+} - N_{+\cdots+y_{k-q_k}\cdots y_{k-1}+\cdots+}\hat{\pi}^{(p_1,\ldots,p_n)}_{y_k|y_{k-q_k}\cdots y_{k-1}}\right)^2}{N_{+\cdots+y_{k-q_k}\cdots y_{k-1}+\cdots+}\hat{\pi}^{(p_1,\ldots,p_n)}_{y_k|y_{k-q_k}\cdots y_{k-1}}}\Bigg),$$

and the likelihood ratio test is given by

$$G^2_{(p_1,\ldots,p_n),(q_1,\ldots,q_n)} \equiv 2\sum_{k=2}^{n} I(q_k > p_k) \sum_{(y_{k-q_k},\ldots,y_k)\in C_{q_k+1}} N_{+\cdots+y_{k-q_k}\cdots y_k+\cdots+} \log\left(\frac{\hat{\pi}^{(q_1,\ldots,q_n)}_{y_k|y_{k-q_k}\cdots y_{k-1}}}{\hat{\pi}^{(p_1,\ldots,p_n)}_{y_k|y_{k-q_k}\cdots y_{k-1}}}\right)$$

where $\hat{\pi}^{(p_1,\ldots,p_n)}_{y_k|y_{k-q_k}\cdots y_{k-1}} = I(p_k = 0)\hat{\pi}^{(p_1,\ldots,p_n)}_{+\cdots+y_k+\cdots+} + I(p_k \geq 1)\hat{\pi}^{(p_1,\ldots,p_n)}_{y_k|y_{k-p_k}\cdots y_{k-1}}$ can be obtained by Theorem 2.1.1. The limiting null distribution of each of these test statistics is chi-square with degrees of freedom equal to

$$dim(\Theta^{(q_1\cdots q_n)}) - dim(\Theta^{(p_1\cdots p_n)}) = (c-1)\sum_{k=1}^{n}(c^{q_k} - c^{p_k}).$$

Similarly to the modification of likelihood ratio test in Section 4.1.2, the following calculations yield the multiplier $\gamma_{(p_1,\ldots,p_n),(q_1,\ldots,q_n)}$ so that $\gamma_{(p_1,\ldots,p_n),(q_1,\ldots,q_n)}G^2_{(p_1,\ldots,p_n),(q_1,\ldots,q_n)}$ can be better approximated by its asymptotic χ^2-distribution:

$$\begin{aligned}
&E(G^2_{(p_1,\ldots,p_n),(q_1,\ldots,q_n)})\\
=&2\sum_{k=2}^{n} I(q_k > p_k) \sum_{(y_{k-q_k},\ldots,y_k)\in C_{q_k+1}} E\left[N_{+\cdots+y_{k-q_k}\cdots y_{k-1}y_k+\cdots+} \log\left(\frac{\hat{\pi}^{(q_1,\ldots,q_n)}_{y_k|y_{k-q_k}\cdots y_{k-1}}}{\hat{\pi}^{(p_1,\ldots,p_n)}_{y_k|y_{k-q_k}\cdots y_{k-1}}}\right)\right]\\
=&2\sum_{k=2}^{n} I(q_k > p_k) \sum_{(y_{k-q_k},\ldots,y_k)\in C_{q_k+1}} E\left[N_{+\cdots+y_{k-q_k}\cdots y_{k-1}y_k+\cdots+} \log\left(\frac{N_{y_{k-q_k}\cdots y_k}}{N_{y_{k-q_k}\cdots y_{k-1}}}\right)\right]\\
&-2\sum_{k=2}^{n} I(q_k > p_k) \sum_{(y_{k-p_k},\ldots,y_k)\in C_{p_k+1}} E\left[N_{+\cdots+y_{k-p_k}\cdots y_{k-1}y_k+\cdots+} \log\left(\frac{N_{y_{k-p_k}\cdots y_k}}{N_{y_{k-p_k}\cdots y_{k-1}}}\right)\right]
\end{aligned}$$

$$
\begin{aligned}
&=2\sum_{k=2}^{n} I(q_k > p_k) \sum_{(y_{k-q_k},\ldots,y_k)\in C_{q_k+1}} E\Bigg[\Bigg(f(N, \pi^{(q_1\cdots q_n)}_{+\cdots+y_{k-q_k}\cdots y_k+\cdots+}) \\
&\quad - \pi^{(q_1\cdots q_n)}_{y_k|y_{k-q_k}\cdots y_{k-1}} f(N, \pi^{(q_1\cdots q_n)}_{+\cdots+y_{k-q_k}\cdots y_{k-1}+\cdots+})\Bigg)\Bigg] \\
&\quad - 2\sum_{k=2}^{n} I(q_k > p_k) \sum_{(y_{k-p_k},\ldots,y_k)\in C_{p_k+1}} E\Bigg[\Bigg(f(N, \pi^{(p_1\cdots p_n)}_{+\cdots+y_{k-p_k}\cdots y_k+\cdots+}) \\
&\quad - \pi^{(p_1\cdots p_n)}_{y_k|y_{k-p_k}\cdots y_{k-1}} f(N, \pi^{(p_1\cdots p_n)}_{+\cdots+y_{k-p_k}\cdots y_{k-1}+\cdots+})\Bigg)\Bigg] \\
&\equiv \frac{(c-1)\sum_{k=1}^{n}(c^{q_k} - c^{p_k})}{\gamma_{(p_1,\ldots,p_n),(q_1,\ldots,q_n)}} \qquad (4.19)
\end{aligned}
$$

Since the true probabilities in (4.19) are unknown, in practice we obtain $\hat{\gamma}_{(p_1,\ldots,p_n),(q_1,\ldots,q_n)}$ by replacing the true probabilities with their mles.

4.4 Homogeneity in distribution of several groups

Suppose that the observational units of the study can be formed meaningfully into s groups. If the concluded orders of antedependence are similar among the groups, we may be interested in testing whether all the s groups are homogeneous in terms of antedependence for some given order $(p_1, \ldots, p_n)$. Equivalently, we are testing

$$
\begin{aligned}
H_0 : &\text{ for } k = 1, \ldots, n, \text{ if } p_k = 0, \pi^{1(p_1,\ldots,p_n)}_{+\cdots+y_k+\cdots+} = \cdots = \pi^{s(p_1,\ldots,p_n)}_{+\cdots+y_k+\cdots+} \equiv \pi^{pool(p_1,\ldots,p_n)}_{+\cdots+y_k+\cdots+}; \\
&\text{and if } p_k \neq 0,\ \pi^{1(p_1,\ldots,p_n)}_{y_k|y_{k-p_k}\cdots y_{k-1}} = \cdots = \pi^{s(p_1,\ldots,p_n)}_{y_k|y_{k-p_k}\cdots y_{k-1}} \equiv \pi^{pool(p_1,\ldots,p_n)}_{y_k|y_{k-p_k}\cdots y_{k-1}} \text{for} \\
&y_k = 1, \ldots, c-1, \text{ and } (y_{k-p_k}, \ldots, y_{k-1}) \in C_{p_k} \text{ under AD}(p_1, \ldots, p_n) \text{ model}
\end{aligned}
$$

against H_1 : inhomogeneity across groups under AD$(p_1, \ldots, p_n)$ model. By putting superscript g in event counts and event probabilities, we denote the observed count and true probability of the realization indicated in the subscript of the gth group.

Thus under H_0: homogeneity across groups, the kernel of the likelihood function is

$$\prod_{g=1}^{s}\left[\prod_{k=1}^{n}\left(I(p_k=0)\prod_{y_k=1}^{c}\left(\pi^{g(p_1,\ldots,p_n)}_{+\cdots+y_k+\cdots+}\right)^{N^g_{+\cdots+y_k+\cdots+}}\right.\right.$$

$$\left.\left.+\,I(p_k\geq 1)\prod_{(y_{k-p_k},\ldots,y_k)\in C_{p_k+1}}\left(\pi^{g(p_1,\ldots,p_n)}_{y_k|y_{k-p_k}\cdots y_{k-1}}\right)^{N^g_{+\cdots+y_{k-p_k}\cdots y_{k-1}y_k+\cdots+}}\right)\right]$$

$$=\prod_{k=1}^{n}\left(I(p_k=0)\prod_{y_k=1}^{c}\left(\pi^{pool(p_1,\ldots,p_n)}_{+\cdots+y_k+\cdots+}\right)^{\sum_{g=1}^{s}N^g_{+\cdots+y_k+\cdots+}}\right.$$

$$\left.+\,I(p_k\geq 1)\prod_{(y_{k-p_k},\ldots,y_k)\in C_{p_k+1}}\left(\pi^{pool(p_1,\ldots,p_n)}_{y_k|y_{k-p_k}\cdots y_{k-1}}\right)^{\sum_{g=1}^{s}N^g_{+\cdots+y_{k-p_k}\cdots y_{k-1}y_k+\cdots+}}\right).$$

Accordingly, we obtain the mle of homogeneous conditional probabilities by pooling across the s groups, yielding

$$\hat{\pi}^{pool(p_1\cdots p_n)}_{+\cdots+y_k+\cdots+}=\frac{\sum_{g=1}^{s}N^g_{+\cdots+y_k+\cdots+}}{\sum_{g=1}^{s}N^g}\text{ for } p_k=0$$

$$\text{and }\hat{\pi}^{pool(p_1\cdots p_n)}_{y_k|y_{k-p_k}\cdots y_{k-1}}=\frac{\sum_{g=1}^{s}N^g_{+\cdots+y_{k-p_k}\cdots y_{k-1}y_k+\cdots+}}{\sum_{g=1}^{s}N^g_{+\cdots+y_{k-p_k}\cdots y_{k-1}+\cdots+}}\text{ for } p_k\geq 1.$$

Under H_1, the likelihood is obtained by taking the product of each of the s multinomial likelihoods under $\mathrm{AD}(p_1,\ldots,p_n)$:

$$\prod_{g=1}^{s}\prod_{k=1}^{n}\left(I(p_k=0)\prod_{y_k=1}^{c}\left(\pi^{g(p_1\cdots p_n)}_{+\cdots+y_k+\cdots+}\right)^{N^g_{+\cdots+y_k+\cdots+}}\right.$$

$$\left.+\,I(p_k\geq 1)\prod_{(y_{k-p_k},\ldots,y_k)\in C_{p_k+1}}\left(\pi^{g(p_1\cdots p_n)}_{y_k|y_{k-p_k}\cdots y_{k-1}}\right)^{N^g_{+\cdots+y_{k-p_k}\cdots y_k+\cdots+}}\right),$$

whose maximum can be obtained by replacing $\pi^{g(p_1\cdots p_n)}_{+\cdots+y_k+\cdots+}$ and $\pi^{g(p_1\cdots p_n)}_{y_k|y_{k-p_k}\cdots y_{k-1}}$ with

$$\hat{\pi}^{g(p_1\cdots p_n)}_{+\cdots+y_k+\cdots+} = \frac{N^g_{+\cdots+y_k+\cdots+}}{N^g}$$

$$\text{and } \hat{\pi}^{g(p_1\cdots p_n)}_{y_k|y_{k-p_k}\cdots y_{k-1}} = \frac{N^g_{+\cdots+y_{k-p_k}\cdots y_{k-1}y_k+\cdots+}}{N^g_{+\cdots+y_{k-p_k}\cdots y_{k-1}+\cdots+}}$$

respectively for $g = 1, \ldots, s$. Thus the score statistic is

$$X^2_{homo(p_1\cdots p_n)} = \sum_{g=1}^{s}\sum_{k=1}^{n}\left[I(p_k=0)\sum_{y_k=1}^{c}\frac{\left(N^g_{+\cdots+y_k+\cdots+} - N^g\hat{\pi}^{pool(p_1\cdots p_n)}_{+\cdots+y_k+\cdots+}\right)^2}{N^g\hat{\pi}^{pool(p_1\cdots p_n)}_{+\cdots+y_k+\cdots+}}\right.$$

$$\left. +I(p_k\geq 1)\sum_{(y_{k-p_k},\ldots,y_k)\in C_{p_k+1}}\frac{\left(N^g_{+\cdots+y_{k-p_k}\cdots y_k+\cdots+} - N^g_{+\cdots+y_{k-p_k}\cdots y_{k-1}+\cdots+}\hat{\pi}^{pool(p_1\cdots p_n)}_{y_k|y_{k-p_k}\cdots y_{k-1}}\right)^2}{N^g_{+\cdots+y_{k-p_k}\cdots y_{k-1}+\cdots+}\hat{\pi}^{pool(p_1\cdots p_n)}_{y_k|y_{k-p_k}\cdots y_{k-1}}}\right],$$

and the likelihood ratio statistic, twice the difference of maximized log-likelihoods under H_1 and H_0, is

$$G^2_{homo(p_1\cdots p_n)} = 2\log\left(\frac{\hat{L}_{H_1}}{\hat{L}_{H_0}}\right)$$

$$=2\sum_{g=1}^{s}\sum_{k=1}^{n}\left[I(p_k=0)\sum_{y_k=1}^{c}N^g_{+\cdots+y_k+\cdots+}\log\left(\frac{\hat{\pi}^{g(p_1\cdots p_n)}_{+\cdots+y_k+\cdots+}}{\hat{\pi}^{pool(p_1\cdots p_n)}_{+\cdots+y_k+\cdots+}}\right)\right.$$

$$\left. + I(p_k\geq 1)\sum_{(y_{k-p_k},\ldots,y_k)\in C_{p_k+1}}N^g_{+\cdots+y_{k-p_k}\cdots y_k+\cdots+}\log\left(\frac{\hat{\pi}^{g(p_1\cdots p_n)}_{y_k|y_{k-p_k}\cdots y_{k-1}}}{\hat{\pi}^{pool(p_1\cdots p_n)}_{y_k|y_{k-p_k}\cdots y_{k-1}}}\right)\right].$$

The limiting null distribution of $X^2_{homo(p_1\cdots p_n)}$ and $G^2_{homo(p_1\cdots p_n)}$ is chi-square with degrees of freedom

$$c^n - 1 - (c-1)\sum_{k=1}^{n}c^{p_k} - \left[c^n - 1 - s(c-1)\sum_{k=1}^{n}c^{p_k}\right]$$

$$=(s-1)(c-1)\sum_{k=1}^{n}c^{p_k}.$$

Unfortunately, in this context the likelihood ratio test is not amenable to modification to improve its performance for small samples, since the mles contain summations of cell counts in both their numerators and denominators, which makes the Taylor expansion a lot more difficult than those described in Sections 4.1, 4.2 and 4.3.

Similarly to the extensions developed for monotone missing data in Section

4.1, $G^2_{p,q}$ and $X^2_{p,q}$ in Section 4.2, $G^2_{(p_1,\ldots,p_n),(q_1,\ldots,q_n)}$ and $X^2_{(p_1,\ldots,p_n),(q_1,\ldots,q_n)}$ in Section 4.3, $X^2_{homo(p_1\cdots p_n)}$ and $G^2_{homo(p_1\cdots p_n)}$ can also be extended to handle monotone missing data. Furthermore, for data with an arbitrary pattern of missingness, the EM algorithm may be applied to obtain mles of transition probabilities, so that the likelihood ratio test may be performed.

CHAPTER 5
STATIONARITY UNDER AD(P) MODEL

Variable-order AD models are inherently nonstationary; constant-order AD models, however, may be stationary or nonstationary. Thus, if the order of antedependence of a model is assumed or determined to be of constant order $p \leq n-1$, we may wish to test for stationarity under the AD(p) model. Two stationarity hypotheses may be of interest: time-invariant pth-order transition probabilities when $p \leq n-2$, and strict stationarity. We consider each hypothesis in turn, assuming again for simplicity that the data are complete.

5.1 Time-invariant transition probabilities under AD(p) for $1 \leq p \leq n-2$

5.1.1 Likelihood ratio and score tests

Note that under AD(p) the likelihood can be written as

$$\left(\prod_{(y_1,\ldots,y_p)\in C_p} \pi_{y_1\cdots y_p+\cdots+}^{N_{y_1\cdots y_p+\cdots+}}\right)\left(\prod_{k=p+1}^{n}\prod_{(y_1,\ldots,y_{p+1})\in C_{p+1}^{+}} \pi_{y_{p+1}^{(k)}|y_1^{(k-p)}\cdots y_p^{(k-1)}}^{N_{+\cdots+y_1^{(k-p)}\cdots y_{p+1}^{(k)}+\cdots+}}\right). \tag{5.1}$$

By Corollary 2.1.2 and the invariance of maximum likelihood estimation, the unrestricted mle of the pth-order transition probability at time k under AD(p) is

$$\hat{\pi}^{(p)}_{y_{p+1}^{(k)}|y_1^{(k-p)}\cdots y_p^{(k-1)}} = \frac{N_{+\cdots+y_1^{(k-p)}\cdots y_p^{(k-1)}y_{p+1}^{(k)}+\cdots+}}{N_{+\cdots+y_1^{(k-p)}\cdots y_p^{(k-1)}+\cdots+}}.$$

Thus, based on likelihood (5.1), the likelihood ratio test statistic for testing stationary transition probabilities under AD(p) is

$$G_t^2 = -2\left[\sum_{(y_1,\ldots,y_p,y_{p+1})\in C_{p+1}^{+}}\sum_{k=p+1}^{n} N_{+\cdots+y_1^{(k-p)}\cdots y_{p+1}^{(k)}+\cdots+}\log\left(\frac{\hat{\pi}^{(p)}_{y_{p+1}^{+}|y_1^{+}\cdots y_p^{+}}}{\hat{\pi}^{(p)}_{y_{p+1}^{(k)}|y_1^{(k-p)}\cdots y_p^{(k-1)}}}\right)\right].$$

The score test statistic for this purpose, on the other hand, is given by

$$X_t^2 = \sum_{k=p+1}^{n} \sum_{(y_1 \cdots y_{p+1}) \in C_{p+1}^+} \frac{\left(N_{+\cdots+y_1^{(k-p)}\cdots y_{p+1}^{(k)}+\cdots+} - N_{+\cdots+y_1^{(k-p)}\cdots y_p^{(k-1)}+\cdots+} \hat{\pi}^{(p)}_{y_{p+1}^+|y_1^+\cdots y_p^+} \right)^2}{N_{+\cdots+y_1^{(k-p)}\cdots y_p^{(k-1)}+\cdots+} \hat{\pi}^{(p)}_{y_{p+1}^+|y_1^+\cdots y_p^+}}.$$

The limiting null distribution of each test statistic is chi-square with degrees of freedom $(c-1)[n-(p+1)]c^p$, equal to the number of equations in condition (2.14).

5.1.2 Simulation

Define binary AD(1) random variables Y_1, Y_2, Y_3 and Y_4 as follows:

$$Y_1 = \begin{cases} 1 & \text{w.p. } \frac{1}{2} \\ 0 & \text{w.p. } \frac{1}{2} \end{cases}, \begin{cases} P(Y_t = 1|Y_{t-1} = 1) = \frac{1}{2} + \frac{(t-1)\theta_2}{10} \\ P(Y_t = 1|Y_{t-1} = 0) = \frac{1}{3} + \frac{(t-1)\theta_2}{10} \end{cases} \text{for } t = 2, 3, 4 \text{ and } 0 \le \theta_2 \le 1 \tag{5.2}$$

The transition probabilities $P(Y_t = 1|Y_{t-1} = 1)$ and $P(Y_t = 1|Y_{t-1} = 0)$ are designed in such a way that they are always unequal for arbitrary θ_2 and t. When $\theta_2 = 0$, process (5.2) is transition probability stationary; otherwise, it is not. We simulated the process 10000 times for each of several values of θ_2 ranging between 0 and 1, and performed the likelihood ratio and score tests for time-invariant transition probabilities. The empirical rejection rates are listed in Table 5.1. Rejection rates for small sample size ($N = 50$), moderate size ($N = 200$) and large size ($N = 1000$) are displayed in separate boxes. Lower bounds (LB) and upper bounds (UB) of 95% Wald-based confidence limits of empirical sizes (rejection rate when $\theta_2 = 0$) are given at the bottom of each box. We can see that for large and moderate size data, the performance of likelihood ratio test and that of score test are very close to each other: for data of large or moderate size, their empirical sizes are both insignificantly different from 0.05 and they are both powerful; for small size data, likelihood ratio test has uniformly higher empirical rejection rates than that of score

test for all values of θ_2. In particular, likelihood ratio test is a bit over sensitive under null hypothesis.

Since the likelihood ratio test performs well for data large sample size, we do not develop on any modification to it, as we did for determining the order of antedependence.

	$N = 50$		$N = 200$		$N = 1000$	
θ_2	LRT	score	LRT	score	LRT	score
0	0.0577	0.0534	0.0531	0.0498	0.0477	0.0473
0.1	0.0623	0.0563	0.0606	0.0615	0.0927	0.0914
0.2	0.0673	0.0606	0.0904	0.086	0.2597	0.2581
0.3	0.0794	0.074	0.1402	0.137	0.5604	0.5571
0.4	0.0952	0.0887	0.2197	0.2138	0.8386	0.8391
0.5	0.1176	0.1115	0.3329	0.3288	0.9688	0.9679
0.6	0.1491	0.1406	0.4703	0.4636	0.997	0.9976
0.7	0.1913	0.1796	0.6197	0.6141	1	1
0.8	0.2379	0.221	0.7619	0.7507	1	1
0.9	0.2994	0.2797	0.8692	0.8586	1	1
1	0.3745	0.3499	0.9414	0.9354	1	1
LB size	0.0531	0.0490	0.0487	0.0455	0.0435	0.0431
UB size	0.0623	0.0578	0.0575	0.0541	0.0519	0.0515

Table 5.1: Empirical rejection rates for tests of transition stationarity for (5.2)

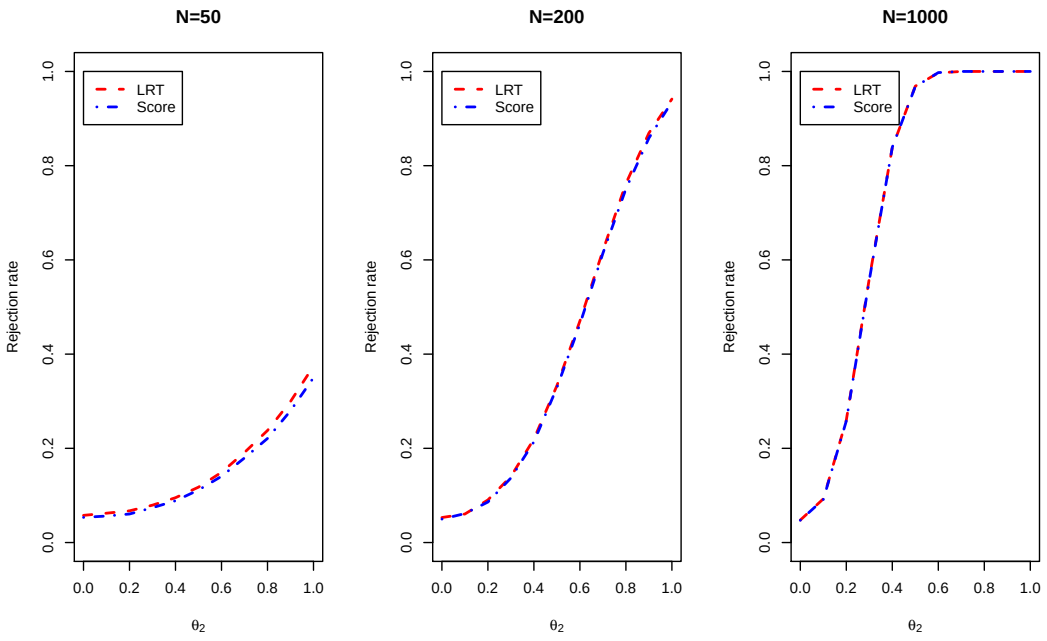


Figure 5.1: Empirical rejection rate curves for (5.2)

5.1.3 Parametric generalized linear model stationary AD(p) structure

Chapter 4 focused primarily on methods related to model selection, including checking for time-invariance of the order of antedependence and time-invariance of the transition probabilities. But if the best model is AD(p) and transition probability stationarity, and if the response is ordinal, we may wish to impose some generalized linear model structure. We do not consider imposing such a structure if the response is nominal since the coefficients in that case have no meaning. Diggle et al. (2002, ch. 10) and Zimmerman and Núñez-Antón (2010, sec. 9.6) discussed

possibilities for imposing such model structure in general. In the special case of binary time series without covariates, Kedem and Fokianos (2002, sec. 2.2) discussed maximum partial likelihood estimation for coefficients $\beta_0,\ \beta_1, \ldots, \beta_p$ in the binary autoregressive structure (5.3)

$$g(P(Y_t = 1|Y_{t-p} = y_p, \ldots, Y_{t-1} = y_1)) = \beta_0 + \beta_1 y_1 + \cdots + \beta_p y_p, \tag{5.3}$$

where g is a link function. Here, in our special case of categorical longitudinal data without covariates, we similarly introduce maximum likelihood estimation for the coefficients in Section 5.1.3.1 and link selection in Section 5.1.3.2 in a generalized linear model structure by maximizing the multinomial likelihood.

5.1.3.1 Parameter estimation

If ordinal categorical variables are coded as equally spaced, consider the generalized linear model structure

$$\begin{aligned} & g(P(Y_k = y_{p+1}|Y_{k-p} = y_1 \cdots Y_{k-1} = y_p)) \\ =& g(\pi^{(p)}_{y^+_{p+1}|y^+_1 \cdots y^+_p}) = \beta_0 + \beta_1 y_1 + \cdots + \beta_p y_p \\ =& \boldsymbol{\beta}' \mathbf{y}_p, \end{aligned} \tag{5.4}$$

where g is a given one-to-one link function such as logit, probit, log-log and so on, $\boldsymbol{\beta} \equiv [\beta_0, \beta_1, \ldots, \beta_p]'$ and $\mathbf{y}_p \equiv [1, y_1, \ldots, y_p]'$. In many cases, continuous cumulative distribution functions can be used as link functions. Note that for binary data, (5.4) is of the same structure as binary autoregressive model (5.3).

By (2.17) and (5.4), the likelihood function is

$$\begin{aligned} L(\boldsymbol{\beta}, g) &\propto \prod_{(y_1,\ldots,y_{p+1}) \in C^+_{p+1}} \left(\pi^{(p)}_{y^+_{p+1}|y^+_1 \cdots y^+_p}\right)^{\sum_{k=p+1}^n N_{+\cdots+y_1^{(k-p)}\cdots y_{p+1}^{(k)}+\cdots+}} \\ &\propto \prod_{(y_1,\ldots,y_{p+1}) \in C^+_{p+1}} \left(g^{-1}(\boldsymbol{\beta}' \mathbf{y}_p)\right)^{\sum_{k=p+1}^n N_{+\cdots+y_1^{(k-p)}\cdots y_{p+1}^{(k)}+\cdots+}}, \end{aligned}$$

yielding the log-likelihood function

$$l(\boldsymbol{\beta}, g) \equiv \log(L(\boldsymbol{\beta}, g)) = \sum_{(y_1,\ldots,y_{p+1})\in C^+_{p+1}} \left(\sum_{k=p+1}^{n} N_{+\cdots+y_1^{(k-p)}\cdots y_{p+1}^{(k)}+\cdots+} \right) \log\left(g^{-1}(\boldsymbol{\beta}'\mathbf{y}_p)\right)$$

The d-th component of the score function $\dot{\mathbf{l}}(\boldsymbol{\beta})$ is

$$\begin{aligned} \dot{l}_d(\boldsymbol{\beta}, g) &\equiv \frac{\partial l(\boldsymbol{\beta}, g)}{\partial \beta_d} \\ &= \sum_{(y_1,\ldots,y_{p+1})\in C^+_{p+1}} \left(\sum_{k=p+1}^{n} N_{+\cdots+y_1^{(k-p)}\cdots y_{p+1}^{(k)}+\cdots+} \right) \frac{1}{g^{-1}(\boldsymbol{\beta}'\mathbf{y}_p)} (g^{-1})'(\boldsymbol{\beta}'\mathbf{y}_p) y_d, \end{aligned}$$

where $d = 0, \ldots, p$ and $y_0 = 1$. If there is an explicit form of $\dot{l}_d(\boldsymbol{\beta}, g)$, the maximum likelihood estimator of β_d can be obtained by directly solving the equation $\dot{l}_d(\boldsymbol{\beta}, g) = 0$; otherwise, a Newton-Raphson algorithm could be applied to obtained the maximum likelihood estimator of $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}}^{(j+1)} = \hat{\boldsymbol{\beta}}^{(j)} + \left[\mathbb{J}(\hat{\boldsymbol{\beta}}^{(j)}, g)\right]^{-1} \dot{l}(\hat{\boldsymbol{\beta}}^{(j)}, g),$$

where

$$\mathbb{J}(\boldsymbol{\beta}, g) \equiv - \begin{bmatrix} \frac{\partial^2 l(\boldsymbol{\beta}, g)}{\partial \beta_0^2} & \frac{\partial^2 l(\boldsymbol{\beta}, g)}{\partial \beta_0 \beta_1} & \cdots & \frac{\partial^2 l(\boldsymbol{\beta}, g)}{\partial \beta_0 \beta_p} \\ \frac{\partial^2 l(\boldsymbol{\beta}, g)}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 l(\boldsymbol{\beta}, g)}{\partial \beta_1^2} & \cdots & \frac{\partial^2 l(\boldsymbol{\beta}, g)}{\partial \beta_1 \beta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 l(\boldsymbol{\beta}, g)}{\partial \beta_p \partial \beta_0} & \frac{\partial^2 l(\boldsymbol{\beta}, g)}{\partial \beta_p \partial \beta_1} & \cdots & \frac{\partial^2 l(\boldsymbol{\beta}, g)}{\partial \beta_p^2} \end{bmatrix}$$

is the observed information (negative Hessian) matrix.

Let $\mathbb{I} = E(\mathbb{J})$ be the Fisher information matrix. Then we have the following asymptotic property for $\boldsymbol{\beta}$:

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \rightsquigarrow \mathrm{N}(0,\ \mathbb{I}^{-1}).$$

5.1.3.2 Link selection

After obtaining the maximized log-likelihood under each link function, we may select the best link by penalized likelihood criteria; likelihood ratio testing is not applicable since models having different links do not nest in one another. But since

the number of coefficients (penalty term) is the same for any given link, selecting the link with the largest log-likelihood is equivalent.

5.2 Strict stationarity

5.2.1 Likelihood ratio and score tests

Recall from Chapter 2 that the mles of transition probabilities under a strictly stationary AD(p) model cannot be expressed in closed form, but they can be obtained numerically using Lang's (2004) algorithm.

Let $\hat{\pi}^{s(p)}_{y_1\cdots y_p+\cdots+}$ and $\hat{\pi}^{s(p)}_{y^+_{p+1}|y^+_1\cdots y^+_p}$ be the mles of the parameters so obtained. Then, based on likelihood (5.1), the likelihood ratio and score test statistics for testing for strict stationarity against nonstationarity of arbitrary type are

$$G^2_s = -2\Bigg[\sum_{(y_1,\ldots,y_p)\in C^+_p} N_{y_1\cdots y_p+\cdots+}\log\left(\frac{\hat{\pi}^{s(p)}_{y_1\cdots y_p+\cdots+}}{\hat{\pi}^{(p)}_{y_1\cdots y_p+\cdots+}}\right) + \sum_{(y_1,\ldots,y_p,y_{p+1})\in C^+_{p+1}}\sum_{k=p+1}^{n} N_{+\cdots+y_1^{(k-p)}\cdots y_{p+1}^{(k)}+\cdots+}\log\left(\frac{\hat{\pi}^{s(p)}_{y^+_{p+1}|y^+_1\cdots y^+_p}}{\hat{\pi}^{(p)}_{y^{(k)}_{p+1}|y^{(k-p)}_1\cdots y^{(k-1)}_p}}\right)\Bigg]$$

and

$$X^2_s = \sum_{(y_1\cdots y_p)\in C^+_p}\frac{\left(N_{y_1\cdots y_p+\cdots+} - N\hat{\pi}^{s(p)}_{y_1\cdots y_p+\cdots+}\right)^2}{N\hat{\pi}^{s(p)}_{y_1\cdots y_p+\cdots+}} + \sum_{k=p+1}^{n}\sum_{(y_1\cdots y_{p+1})\in C^+_{p+1}}\frac{\left(N_{+\cdots+y_1^{(k-p)}\cdots y_{p+1}^{(k)}+\cdots+} - N_{+\cdots+y_1^{(k-p)}\cdots y_p^{(k-1)}+\cdots+}\hat{\pi}^{s(p)}_{y^+_{p+1}|y^+_1\cdots y^+_p}\right)^2}{N_{+\cdots+y_1^{(k-p)}\cdots y_p^{(k-1)}+\cdots+}\hat{\pi}^{s(p)}_{y^+_{p+1}|y^+_1\cdots y^+_p}},$$

respectively. Note that in contrast to the testing procedure for transition probability stationarity under AD(p), we realize that $\hat{\pi}^{s(p)}_{y_1\cdots y_p+\cdots+} \neq \hat{\pi}^{(p)}_{y_1\cdots y_p+\cdots+}$, the mles of joint probabilities of the initial p time points under strict stationarity AD(p) and AD(p) are not equal any more. The limiting null distribution of each test statistic is chi-square with $(c-1)[n-(p+1)]c^p + (c^p-1)$ degrees of freedom.

5.2.2 Simulation

Define binary AD(1) random variables Y_1, Y_2, Y_3, Y_4 as follows:

$$Y_1 = \begin{cases} 1 & \text{w.p. } \frac{1}{2} \\ 0 & \text{w.p. } \frac{1}{2} \end{cases}, \begin{cases} P(Y_t = 1|Y_{t-1} = 1) = \frac{\theta_3}{3} \\ P(Y_t = 1|Y_{t-1} = 0) = \frac{\theta_3}{2} \end{cases} \text{for } t = 2, 3, 4 \text{ and } 0.4 \leq \theta_3 \leq 1.8 \tag{5.5}$$

Note that for arbitrary $\theta_3 \neq 0$, $P(Y_t = 1|Y_{t-1} = 0)$ is always 50% larger than $P(Y_t = 1|Y_{t-1} = 1)$. Clearly, process (5.5) is transition stationary since the transition probability is free of time point t. In particular,

$$\begin{aligned} P(Y_2 = 1) =& P(Y_2 = 1|Y_1 = 1)P(Y_1 = 1) + P(Y_2 = 1|Y_1 = 0)P(Y_1 = 0) \\ =& \frac{\theta_3}{3}\frac{1}{2} + \frac{\theta_3}{2}\frac{1}{2} = \frac{5\theta_3}{12} \end{aligned}$$

Thus, when $\theta_3 = \frac{6}{5}$, $P(Y_2 = 1) = \frac{1}{2} = P(Y_1 = 1)$, which guarantees that process (5.5) is of strict stationarity by Lemma 2.2.3; otherwise, it is transition probability stationary. We simulated the process 10000 times for each of several values of θ_3 ranging between 0.4 and 1.8, and performed the likelihood ratio and score tests for transition probability stationarity and strict stationarity respectively. The empirical rejection rates are listed in Table 5.2. When θ_3 is far away from 1.2 and the sample size is insufficiently large, there are too many empty cells in the simulated data, so that the test for strict stationarity could not be performed by either the likelihood ratio test or score test due to zero expected cell counts. We use "X" in Table 5.2 to indicate such cases. Rejection rates for small sample size ($N = 50$), moderate size ($N = 200$) and large size ($N = 1000$) are displayed in separate boxes. Lower bounds (LB) and upper bounds (UB) of 95% Wald-based confidence limits of empirical sizes (rejection rate when $\theta_3 = 1.2$) are given at the bottom of each box. We can see that

the likelihood ratio test is better for testing for strict stationarity than the score test since the empirical size of the likelihood ratio test is not significantly different from the nominal size 0.05 except when the sample size is small, and the power of the likelihood ratio test is uniformly higher than that of the score test. Since the likelihood ratio test performs well here, we do not develop on any modification to it, as we did for determining the order of antedependence. As expected, since process (5.5) is transition probability stationary, the empirical rejection rates for testing for transition probability stationarity by both the likelihood ratio test and score test for all values of θ_3 are not significantly higher than the nominal size 0.05. In contrast, for testing for strict stationarity, the further θ_3 is away from 1.2, the higher the empirical rejection rate is. From Table 5.2, we see that for large and moderate size data, the performance of likelihood ratio test and that of score test are very close to each other: for data of large or moderate size, their empirical sizes are both insignificantly different from 0.05 and they are both powerful; for small size data, likelihood ratio test has uniformly higher empirical rejection rates than that of score test for all values of θ_3. In particular, likelihood ratio test is a bit over sensitive under null hypothesis.

Because the likelihood ratio test performs well as it stands, we do not develop a modified version of it.

The results of this section could also be used to test for strict stationarity under AD(p) against the hypothesis of time-invariant pth-order transition probabilities under AD(p) in an obvious way.

	$N = 50$				$N = 200$				$N = 1000$			
	transition		strict		transition		strict		transition		strict	
θ_3	LRT	score	LRT	score	LRT	score	LRT	score	LRT	score	LRT	score
0.4	X	X	X	X	X	X	X	X	0.048	0.049	1	1
0.6	X	X	X	X	0.049	0.050	0.999	0.999	0.047	0.046	1	1
0.8	X	X	X	X	0.052	0.048	0.845	0.839	0.047	0.047	1	1
1	0.057	0.054	0.096	0.089	0.048	0.046	0.243	0.236	0.047	0.047	0.903	0.903
1.2	0.057	0.054	0.058	0.054	0.053	0.051	0.051	0.050	0.049	0.050	0.049	0.050
1.4	0.061	0.055	0.099	0.089	0.051	0.048	0.234	0.224	0.045	0.046	0.877	0.877
1.6	X	X	X	X	0.052	0.049	0.772	0.765	0.045	0.045	1	1
1.8	X	X	X	X	X	X	X	X	0.050	0.050	1	1
LB	0.052	0.050	0.053	0.049	0.048	0.046	0.047	0.046	0.045	0.045	0.045	0.046
UB	0.061	0.058	0.062	0.058	0.057	0.055	0.055	0.055	0.053	0.054	0.053	0.054

Table 5.2: Empirical rejection rates for tests of two types of stationarity for (5.5)

CHAPTER 6
EXAMPLES

6.1 Labor force data

The labor force data (Lindsey, 1993, p. 185) are annual observations of the employment status of 1583 women based on a wide variety of demographic, social and economic variable from 1967 to 1971 from a cross section of dwellings in the United States. It is interesting to determine how the economic trend during the years 1967 through 1971 affects the employment status of married women. A detailed description of the data is contained in Heckman and Willis (1977). The data are displayed in a multinomial format in Table 6.1, where we ignore all the explanatory variables. Here, Y_1 through Y_5 correspond to the years 1967 through 1971; no data are missing. In the body of data, 1 stands for being employed while 0 stands for being unemployed. For these data, the sample size is quite large and the number of measurement times is relatively small. The sample proportions corresponding to the $2^5 = 32$ cells indicate that employment status is rather persistent over the time period: in particular, the events of being employed for all 5 years or unemployed for all 5 years have relatively high proportions, 0.269 and 0.353 respectively. All remaining cells have relatively small proportions (less than 0.05). The column count contains the number of women with the employment status in the corresponding row during these 5 years.

The penalized log-likelihood model selection procedure based on Theorem 3.1.1 was applied to these data. It selects $\mathrm{AD}(0, 1, 2, 3, 3)$ (or equivalently $\mathrm{AD}(3)$)

Y_1	Y_2	Y_3	Y_4	Y_5	Count	Y_1	Y_2	Y_3	Y_4	Y_5	Count
1	1	1	1	1	426	0	1	1	1	1	73
1	1	1	1	0	38	0	1	1	1	0	11
1	1	1	0	1	16	0	1	1	0	1	7
1	1	1	0	0	47	0	1	1	0	0	17
1	1	0	1	1	11	0	1	0	1	1	9
1	1	0	1	0	2	0	1	0	1	0	3
1	1	0	0	1	12	0	1	0	0	1	5
1	1	0	0	0	28	0	1	0	0	0	24
1	0	1	1	1	21	0	0	1	1	1	54
1	0	1	1	0	7	0	0	1	1	0	16
1	0	1	0	1	0	0	0	1	0	1	6
1	0	1	0	0	9	0	0	1	0	0	28
1	0	0	1	1	8	0	0	0	1	1	36
1	0	0	1	0	3	0	0	0	1	0	24
1	0	0	0	1	5	0	0	0	0	1	35
1	0	0	0	0	43	0	0	0	0	0	559

Table 6.1: Labor Force Data

when AIC is the criterion and selects AD$(0, 1, 2, 2, 3)$ when BIC is the criterion. As is shown in Table 6.2, score tests and unmodified likelihood ratio tests are all highly significant ($P < 10^{-8}$) for AD(0) against AD(1), AD(1) against AD(2), and AD(2) against AD(3), but those for AD(3) against AD(4) are not significant ($P = 0.418$ for the score test and $P = 0.340$ for the likelihood ratio test). Thus

we conclude that AD(3) is the best constant-order AD model. Next, we carry out the two stationarity tests under the assumption of an AD(3) model. As is shown in Table 6.3, there is not significant evidence that the transition probabilities are nonstationary ($P = 0.253$ for the score test and $P = 0.201$ for the likelihood ratio test) and we list the maximum likelihood estimators of all the nonredundant stationary transition probabilities under AD(3) in Table 6.4. For example, the maximum likelihood estimator of the conditional probability that the women were employed during the year either 1970 or 1971 conditioning on the event that they were employed at all previous three years is 0.896. Note that among the conditional probabilities that the women were employed given that they were unemployed only once during the preceding three years, $\hat{\pi}^{(3)}_{1^+|0^+,1^+,1^+}$ and $\hat{\pi}^{(3)}_{1^+|1^+,0^+,1^+}$ are close and are much higher than $\hat{\pi}^{(3)}_{1^+|1^+,1^+,0^+}$, showing that the employment statuses of those women at each year were more closely related to those at the year right before than those at further preceding years. Similarly, $\hat{\pi}^{(3)}_{1^+|1^+,0^+,0^+}$ and $\hat{\pi}^{(3)}_{1^+|0^+,1^+,0^+}$ are close and are much lower than $\hat{\pi}^{(3)}_{1^+|0^+,0^+,1^+}$.

However, strict stationarity is strongly rejected ($P = 1.65 \times 10^{-5}$ for the score test and $P = 8.96 \times 10^{-6}$ for the likelihood ratio test). An examination of each year's marginal probability of employment suggests that the rejection of strict stationarity is due largely to a significantly smaller level of employment among surveyees in 1967 ($\hat{\pi}_{1++++} = 0.427$) and a significantly larger level in 1969 ($\hat{\pi}_{++1++} = 0.490$) than in the three other years ($\hat{\pi}_{+1+++} = 0.461$, $\hat{\pi}_{+++1+} = 0.469$, $\hat{\pi}_{++++1} = 0.457$).

Next, since the order of antedependence has been determined to be three, and the transition probabilities appear to be stationary, we fit an AR(3) model. We consider the four most commonly used links in this context: logit, probit, log-log and complementary log-log. Maximum likelihood estimators of coefficients $\boldsymbol{\beta}$ for

each link and the maximized likelihood under each structure are summarized in Table 6.5. From the table, we see that the model with maximized likelihood among logit AR(3), probit AR(3), log-log AR(3) and complementary log-log AR(3) is

$$\begin{aligned}&\mathbf{\Phi}^{-1}(E(Y_k|\mathcal{F}_{k-1}))\\ =&\mathbf{\Phi}^{-1}(P(Y_k=1|\mathcal{F}_{k-1}))=-1.338+1.711Y_{k-1}+0.402Y_{k-2}+0.427Y_{k-3}\end{aligned}$$

where $k=4,5$ and $\mathbf{\Phi}$ is the cumulative distribution function of the standard normal distribution. So the value evaluated by inverse function of cumulative distribution function of the standard normal distribution at the probability that those women were employed during the year either 1970 or 1971 is equal to -1.338 when those women are unemployed during all the previous three years, and is changed by 1.711, 0.402 and 0.427 by employment status during the lag-one year, lag-two year and lag-three year respectively.

Hypotheses tested	Score	Likelihood ratio
AD(0) vs AD(1)	0	0
AD(1) vs AD(2)	0	0
AD(2) vs AD(3)	0	8.08×10^{-9}
AD(3) vs AD(4)	0.418	0.340
Hypotheses tested	Score	Likelihood ratio
AD(0) vs AD(4)	0	0
AD(1) vs AD(4)	0	0
AD(2) vs AD(4)	2.61×10^{-12}	1.05×10^{-9}
AD(3) vs AD(4)	0.418	0.340

Table 6.2: P-values for testing for order of antedependence of the labor force data

Hypotheses tested	Score	Likelihood ratio
Stationary transition probabilities in AD(3) vs AD(3)	0.253	0.201
Strict stationarity in AD(3) vs AD(3)	1.65×10^{-5}	8.96×10^{-6}

Table 6.3: P-values for testing for stationarity under AD(3) for labor force data

$\hat{\pi}^{(3)}_{1^+\|1^+,1^+,1^+}$	0.896
$\hat{\pi}^{(3)}_{1^+\|1^+,1^+,0^+}$	0.257
$\hat{\pi}^{(3)}_{1^+\|1^+,0^+,1^+}$	0.774
$\hat{\pi}^{(3)}_{1^+\|1^+,0^+,0^+}$	0.219
$\hat{\pi}^{(3)}_{1^+\|0^+,1^+,1^+}$	0.772
$\hat{\pi}^{(3)}_{1^+\|0^+,1^+,0^+}$	0.214
$\hat{\pi}^{(3)}_{1^+\|0^+,0^+,1^+}$	0.651
$\hat{\pi}^{(3)}_{1^+\|0^+,0^+,0^+}$	0.077

Table 6.4: Stationary transition probabilities under AD(3) for labor force data

$g(z)$	$\log\left(\frac{z}{1-z}\right)$	$\mathbf{\Phi}^{-1}(z)$	$-\log(-\log(z))$	$\log(-\log(1-z))$
β_0	-2.301	-1.338	-1.036	-2.190
β_1 (for Y_{k-1})	2.866	1.711	1.978	2.081
β_2 (for Y_{k-2})	0.721	0.402	0.491	0.504
β_3 (for Y_{k-3})	0.77	0.427	0.517	0.544
log-likelihood	-83.049	$\mathbf{-82.647}$	-90.438	-92.459

Table 6.5: Link selection for AR(3) in the labor force data

6.2 Wheeze data

The Wheeze data, as described by Agresti (2002, p. 478), come from a longitudinal study at Harvard of effects of air pollution on respiratory illness in 1019 children. The children were examined annually from age 9 to 12 and classified according to the presence or absence of wheeze (1 = presence 0 = absence). The data are displayed in a multinomial format in Table 6.6. Here, Y_9 through Y_{12} correspond to the ages from 9 to 12; no data are missing. The data do not include possibly relevant covariates such as parental smoking behavior. The observations corresponding to the $2^4 = 16$ cells indicate that more than half (572 out of 1019) children did not have wheeze from age 9 to 12 at all.

Y_9	Y_{10}	Y_{11}	Y_{12}	Count	Y_9	Y_{10}	Y_{11}	Y_{12}	Count
1	1	1	1	94	0	1	1	1	19
1	1	1	0	30	0	1	1	0	15
1	1	0	1	15	0	1	0	1	10
1	1	0	0	28	0	1	0	0	44
1	0	1	1	14	0	0	1	1	17
1	0	1	0	9	0	0	1	0	42
1	0	0	1	12	0	0	0	1	35
1	0	0	0	63	0	0	0	0	572

Table 6.6: Wheeze data

The penalized log-likelihood model selection procedure based on Theorem

3.1.1 selects AD$(0, 1, 2, 3)$ (or equivalently AD(3)) when AIC is the criterion, and selects AD$(0, 1, 2, 2)$ (or equivalently AD(2)) when BIC is the criterion.

Hypotheses tested	Score	Likelihood ratio
AD(0) vs AD(1)	0	0
AD(1) vs AD(2)	0	0
AD(2) vs AD(3)	2.31×10^{-5}	8.51×10^{-5}
Hypotheses tested	**Score**	**Likelihood ratio**
AD(0) vs AD(3)	0	0
AD(1) vs AD(3)	0	0
AD(2) vs AD(3)	2.31×10^{-5}	8.51×10^{-5}

Table 6.7: P-values for testing for order of antedependence of the Wheeze data

$\hat{\pi}^{(3)}_{Y_4=1\|Y_1=1,Y_2=1,Y_3=1}$	0.758
$\hat{\pi}^{(3)}_{Y_4=1\|Y_1=1,Y_2=1,Y_3=0}$	0.349
$\hat{\pi}^{(3)}_{Y_4=1\|Y_1=1,Y_2=0,Y_3=1}$	0.609
$\hat{\pi}^{(3)}_{Y_4=1\|Y_1=1,Y_2=0,Y_3=0}$	0.160
$\hat{\pi}^{(3)}_{Y_4=1\|Y_1=0,Y_2=1,Y_3=1}$	0.559
$\hat{\pi}^{(3)}_{Y_4=1\|Y_1=0,Y_2=1,Y_3=0}$	0.185
$\hat{\pi}^{(3)}_{Y_4=1\|Y_1=0,Y_2=0,Y_3=1}$	0.288
$\hat{\pi}^{(3)}_{Y_4=1\|Y_1=0,Y_2=0,Y_3=0}$	0.058

Table 6.8: MLE of transition probabilities of the Wheeze data under AD(3)

As is shown in Table 6.7, unmodified likelihood ratio tests and score tests are

all highly significant for AD(0) against AD(1), AD(1) against AD(2), and AD(2) against AD(3). Thus we conclude that AD(3) (the saturated model) is the best constant-order AD model and we list the maximum likelihood estimators of all the nonredundant transition probabilities under AD(3) in Table 6.8. For example, the maximum likelihood estimator of the conditional probability that the children had wheeze presence at the age of 12 conditioning on the event that they had wheeze presence at the age of 9 through 11 is 0.758. Note that among the conditional probabilities that the children had wheeze given that they were had wheeze presence only once during the preceding three years, $\hat{\pi}^{(3)}_{Y_4=1|Y_1=0,Y_2=1,Y_3=1}$ and $\hat{\pi}^{(3)}_{Y_4=1|Y_1=1,Y_2=0,Y_3=1}$ are close and are much higher than $\hat{\pi}^{(3)}_{Y_4=1|Y_1=1,Y_2=1,Y_3=0}$, showing that the wheeze presence of those children at each year was more closely related to that at the year right before than that at further preceding years. Similarly, $\hat{\pi}^{(3)}_{Y_4=1|Y_1=1,Y_2=0,Y_3=0}$ and $\hat{\pi}^{(3)}_{Y_4=1|Y_1=0,Y_2=1,Y_3=0}$ are close and are much lower than $\hat{\pi}^{(3)}_{Y_4=1|Y_1=0,Y_2=0,Y_3=1}$.

Accordingly, we do not consider testing for transition probability stationarity. However, we test for strict stationarity against the saturated model and find slight though not quite statistically significant evidence against strict stationarity ($P = 0.0686$ for the score test and $P = 0.0609$ for the likelihood ratio test.) An examination of each year's marginal probability of wheeze presence suggests that the nonstationarity is due largely to a significantly smaller level of wheeze presences among those children at the ages of 11 and 12 ($\hat{\pi}_{++1+} = 0.235$, $\hat{\pi}_{+++1} = 0.212$) than that of those children at the ages of 9 and 10 ($\hat{\pi}_{1+++} = 0.260$, $\hat{\pi}_{+1++} = 0.250$.)

6.3 Toenail infection data

The toenail infection data, as described by Molenberghs and Verbeke (2005, p. 8), were obtained from a study comparing two oral treatments (labeled here as A

and B) for toenail dermatophyte onychomycosis. Two response variables, unaffected nail length (in mm) and the severity of the infection (0 = not severe, 1 = severe) were measured at seven scheduled time points: baseline; months 1, 2, and 3 during treatment; and months 6, 9, and 12 after initial treatment. We focus our attention on the order of antedependence for the severity status variable. As is shown in Tables 6.9 and 6.10, of the 146 and 148 people who received Treatments A and B, respectively, 107 and 117 provided their severity status at all time points. The pattern of missingness is not monotone, so we use the EM algorithm to obtain mles and determine the order of antedependence.

From Table 6.11, we can see that for treatment A, AIC selects AD$(0, 1, 1, 1, 3, 1, 2)$ while BIC selects AD$(0, 1, 1, 1, 1, 1, 1)$ (or equivalently AD(1)); for treatment B, both AIC and BIC select AD$(0, 1, 1, 1, 1, 1, 1)$ (or equivalently AD(1)). Thus it is sensible to test for constant order of antedependence for each treatment. Due to the non-monotone pattern of missingness in this data set, we use the unmodified likelihood ratio test. For both treatments A and B, we conclude that AD(1) is the best constant-order AD model, as is shown in Table 6.12. This result is consistent to the conclusion shown by Molenberghs and Verbeke. We list the maximum likelihood estimators of all the nonredundant stationary transition probabilities under AD(1) in Table 6.13. For example, for all the patients who received Treatment A, the maximum likelihood estimator of the conditional probability that the patients had severe infection at the first month conditioning on the event that they had infection which was not severe at the baseline is 0.014. Note that for both groups, the conditional probabilities that the patients had severe infection for a time given that they had severe infection at the measurement time right before decrease during the first six months (time points 1 to 5) but increase during the rest periods (time points 6

and 7.)

Since this data set is not observed at equally spaced time points, we do not consider tests for stationarity. But since both treatment groups share the same constant order of antedependence, it is sensible to test whether there is homogeneity across the two groups. The P-value for the likelihood ratio test of homogeneity across groups A and B is 0.475, indicating no evidence against homogeneity. From Table 6.13, we can see that the transition probabilities are close between those two groups.

Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Count	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Count
1	1	1	1	1	1	1	5	1	1	1	1	•	0	0	1
1	1	1	1	1	0	0	5	0	1	1	1	0	•	1	1
1	1	1	1	0	0	0	7	1	1	0	0	0	•	1	1
1	1	1	0	0	0	0	9	0	0	0	0	0	•	1	1
1	1	0	1	0	0	0	1	1	1	1	1	1	•	0	1
1	1	0	0	0	0	0	8	1	1	1	1	0	•	0	2
1	0	0	0	0	0	0	3	1	1	1	0	0	•	0	1
0	1	0	0	1	1	0	1	0	0	1	0	0	•	0	1
0	0	1	1	0	0	0	2	1	0	0	0	0	•	0	1
0	0	1	0	0	0	0	1	0	0	0	0	0	•	0	6
0	0	0	1	0	0	0	1	1	1	1	•	1	•	1	1
0	0	0	0	0	1	0	1	1	1	1	0	•	•	0	1
0	0	0	0	0	0	1	2	0	0	1	1	0	0	•	1
0	0	0	0	0	0	0	61	0	0	•	•	0	•	•	1
0	•	0	0	0	0	1	1	0	0	0	0	•	•	•	2
0	0	•	0	0	0	0	1	1	1	1	•	•	•	•	1
1	1	1	•	1	0	0	1	0	0	0	•	•	•	•	2
0	0	1	•	0	0	0	1	0	0	•	•	•	•	•	2
0	0	0	•	0	0	0	1	1	•	•	•	•	•	•	2
1	1	1	1	•	1	1	2	0	•	•	•	•	•	•	2
1	0	0	0	•	1	0	1								

Table 6.9: Toenail data by treatment A

Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Count	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Count
1	1	1	1	1	1	1	3	1	1	1	1	•	0	0	2
1	1	1	1	1	1	0	1	1	1	1	•	•	0	0	1
1	1	1	1	1	0	0	1	0	0	0	0	0	•	0	5
1	1	1	1	0	0	0	14	1	1	0	0	0	•	0	1
1	1	1	0	0	0	0	11	1	1	1	1	•	•	0	1
1	1	0	0	0	0	0	4	1	1	1	1	0	0	•	1
1	0	1	1	0	0	0	1	1	1	1	1	1	1	•	1
1	0	0	0	0	0	0	7	0	0	•	1	1	0	•	1
0	1	1	1	0	0	0	2	0	0	0	0	0	•	•	3
0	0	1	1	0	0	0	1	1	0	0	0	0	•	•	1
0	0	0	0	1	1	1	1	1	1	0	0	0	•	•	1
0	0	0	0	0	1	1	1	0	0	0	•	0	•	•	2
0	0	0	0	0	0	1	1	0	0	0	0	•	•	•	2
0	0	0	0	0	0	0	69	0	0	0	•	•	•	•	2
0	0	0	0	•	0	0	1	1	1	0	•	•	•	•	1
0	0	0	0	•	1	0	1	1	1	•	•	•	•	•	1
1	1	0	0	•	0	0	2	0	•	•	•	•	•	•	1

Table 6.10: Toenail data by treatment B

Penalized likelihood criteria	Best order in treatment A	Best order in treatment B
AIC	AD(0, 1, 1, 1, 3, 1, 2)	AD(0, 1, 1, 1, 1, 1, 1)
BIC	AD(0, 1, 1, 1, 1, 1, 1)	AD(0, 1, 1, 1, 1, 1, 1)

Table 6.11: Order selection by penalized likelihood criteria in the toenail data

Hypotheses tested	Treatment A	Treatment B
AD(0) vs AD(1)	0	0
AD(1) vs AD(2)	0.195	0.740
AD(2) vs AD(3)	0.311	0.989
AD(3) vs AD(4)	0.949	1
AD(4) vs AD(5)	1	1
AD(5) vs AD(6)	1	1

Hypotheses tested	Treatment A	Treatment B
AD(0) vs AD(6)	0	0
AD(1) vs AD(6)	1	1
AD(2) vs AD(6)	1	1
AD(3) vs AD(6)	1	1
AD(4) vs AD(6)	1	1
AD(5) vs AD(6)	1	1

Table 6.12: P-values for order selection by likelihood ratio test for the toenail data

	Treatment A	Treatment B
$\hat{\pi}^{(1)}_{Y_2=1\|Y_1=0}$	0.014	0.027
$\hat{\pi}^{(1)}_{Y_2=1\|Y_1=1}$	0.921	0.810
$\hat{\pi}^{(1)}_{Y_3=1\|Y_2=0}$	0.042	0.025
$\hat{\pi}^{(1)}_{Y_3=1\|Y_2=1}$	0.722	0.889
$\hat{\pi}^{(1)}_{Y_4=1\|Y_3=0}$	0.026	0.001
$\hat{\pi}^{(1)}_{Y_4=1\|Y_3=1}$	0.655	0.676
$\hat{\pi}^{(1)}_{Y_5=1\|Y_4=0}$	0.012	0.011
$\hat{\pi}^{(1)}_{Y_5=1\|Y_4=1}$	0.476	0.217
$\hat{\pi}^{(1)}_{Y_6=1\|Y_5=0}$	0.010	0.009
$\hat{\pi}^{(1)}_{Y_6=1\|Y_5=1}$	0.545	0.833
$\hat{\pi}^{(1)}_{Y_7=1\|Y_6=0}$	0.020	0.009
$\hat{\pi}^{(1)}_{Y_7=1\|Y_6=1}$	0.714	0.833

Table 6.13: MLE of transition probabilities of the toenail data under AD(1)

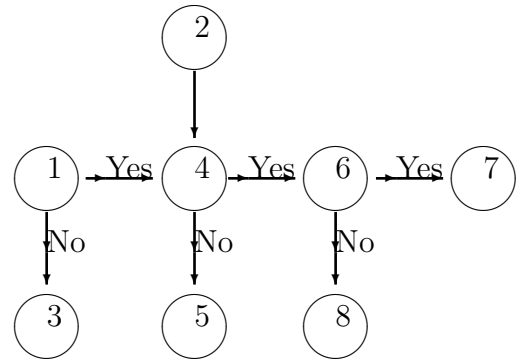
CHAPTER 7
CONCLUSION AND DISCUSSION

We have presented methodology for obtaining the maximum likelihood estimators of transition probabilities under different assumptions, determining the order of antedependence of categorical longitudinal data by penalized likelihood criteria and hypothesis testing and testing for two types of stationarity under constant order of antedependence for both complete data and data with missingness. These methods are intended as a supplement to, not a replacement for, existing transition-model methodology for such data. Determining the order of antedependence and other relevant features (e.g. stationarity) prior to assuming any additional structure should facilitate the subsequent selection of a model from a more structured class of antedependence models for categorical data, such as the stationary Markov generalized linear models described by Diggle et al. (2002, pp. 190-207) and Molenberghs and Verbeke (2005, pp. 236-238) and the marginalized transition models of Heagerty (2002).

7.1 Conclusion with flowchart

In this section, we summarize the procedure of using the methods introduced in this thesis into a flowchart and make some recommendations for practical use by comparison.

7.1.1 Flowchart



- **1**. Constant AD order cannot be assumed.
  Determine the best variable AD order
  by penalized likelihood criteria.
  Whether the best AD order is a
  constant AD order or nearly so?

- **2**. Constant-order AD model can be assumed.

- **3**. Variable-order AD model is concluded, which is inherently nonstationary. Stop.

- **4**. Determine (confirm) the constant AD order p by hypothesis tests (likelihood ratio or score test).
  Are data measured on equally-spaced time points?

- **5**. Stop.

- **6**. Transition probability stationarity under AD(p)?

- **7**. Conclude constant-coefficient AR(p) model and fit generalized linear model. Further check for strict stationarity.

- **8**. Functional-coefficient AR(p) model.

7.1.2 Comparison of the tests

In Chapter 2, we derived methods for determining the maximum likelihood estimator of transition probabilities under arbitrary variable order of antedependence and under transition probability stationary and strictly stationary AD(p) models for complete data, monotone missing data and data with an arbitrary pattern of missingness. In Chapter 3, we introduced methods for determining the best variable order antedependence model by penalized likelihood criteria. We use penalized likelihood criteria for order selection first, as is shown in the flowchart. However, different penalized likelihood criteria may yield different selected variable order antedependence models. Thus, unless the variable orders selected by penalized likelihood criteria vary substantially across time, it may be desirable to further determine a constant order of antedependence by hypothesis testing, especially when the number of time points is not large. In Chapters 4 and 5, we develop hypothesis testing procedures for determining constant order of antedependence and testing for two types of stationarity under AD(p) model. In Tables 7.1 and 7.2, we make comparisons among those tests by different criteria. For determining the order of antedependence, the likelihood ratio and score tests are certainly better than the Wald test. But neither the likelihood ratio test nor score test is a panacea. For testing constant order of antedependence, the score test is powerful with its actual size matching its nominal size, but it cannot be used if the data have an arbitrary pattern of missingness. The likelihood ratio test can be applied to data with any missingness pattern, but it is oversensitive under the null hypothesis. A modified version of the likelihood ratio test, modified by equating the expectation of a scalar multiplied test statistic and its degrees of freedom is an improvement. For testing two types of stationarity, the likelihood ratio test and score test appeared to have

similar performance for large and moderate size data, while the likelihood ratio test has uniformly higher rejection rates than that of score test for small size data. Practically, for complete or monotone missing data, as in Chapter 6, we refer to both of them for order selection and two types of stationarity as long as they are applicable, while for data with arbitrary missingness, likelihood ratio test is the only one applicable.

Criteria	Likelihood ratio	Score	Wald
Actual size	oversensitive [1]	matches nominal size	lower than nominal size
Power	powerful	powerful	not powerful
Convegence rate	fast	fast	slow
Empty cell	not affected	not affected	affected
Monotone missing	closed form	closed form	no closed form
Arbitrary missing	works by EM algorithm	not work	not work

[1]Modification can serve as a remedy

Table 7.1: Comparison among triad for testing AD order

Criteria	Likelihood ratio	Score
Actual size	matches nominal size	matches nominal size
Power	powerful	powerful
Transition stationarity AD(p)	always works	not work for arbitrary missingness
Strict stationarity AD(p)	complete data only	complete data only

Table 7.2: Comparison among triad for testing stationarity under AD(p)

7.1.3 Extension to multivariate categorical longitudinal data

In this thesis, we assumed that the categorical longitudinal data is univariate. However, it is not conceptually difficult to extend out estimation and testing procedure to multivariate data. In fact, Anderson and Goodman (1957 sec. 3.5) extended some of their aforementioned procedure to bivariate categorical longitudinal data. In general, an extension to multivariate categorical longitudinal data can be made by simply relabelling the outcomes so that multivariate outcomes are converted to univariate outcomes. More specifically, suppose that there are u categorical characteristics in the observations and that there are c_1 categories for the first characteristic,..., c_u categories for the uth characteristic. Then, we can simply consider all the $c_1 \times \cdots \times c_u$ possible combinations of u characteristics as a new, single characteristic with $c_1 \times \cdots \times c_u$ categories, and apply the methods developed in this thesis.

7.2 Discussion and open questions

To our knowledge, this thesis represents the second major research effort on inference for antedependence (Markov) models for categorical longitudinal data without covariates, the first being Anderson and Goodman (1957). We extended Anderson and Goodman's methods in several directions by using techniques developed after 1957, such as the EM algorithm to complete data with missingness, penalized likelihood criteria for model selection, using numerical methods to test for strict stationarity under an AD(p) model and so on. One problem we have not yet solved is how to test for strict stationarity under an AD(p) model when the data are incomplete. Since the likelihood function for a strictly stationary AD(p) model cannot

be written as the product of saturated multinomial distributions, we have to find an alternative way to do the E-step if we want to perform the EM algorithm. This could be an interesting open question in the framework of considering categorical longitudinal data without covariates as multinomial outcome.

Yet another interesting question is how to develop methods for AD models for multivariate mixed (some variables continuous, some categorical) longitudinal data with or without covariates. Note that the toenail infection data considered in this prospectus are actually bivariate and mixed, though we considered only the categorical response variable in Section 6.3. One way to analyze mixed data is to categorize the continuous variables to convert the mixed data to multivariate categorical data and use the methods introduced in Section 7.1. However, categorization of continuous data may not be informative enough, and we suggest factoring the likelihood function as an alternative way. Consider, for example, the simplest case of one binary characteristic (Y) and one continuous (normal) characteristic (Z): $\mathcal{F}_n \equiv \big((Y_1, Z_1), (Y_2, Z_2), \ldots, (Y_n, Z_n)\big)$ across n time points. Then, p^{th} order antedependence prescribes that

$$\begin{aligned}
&f((Y_k, Z_k)|\mathcal{F}_{k-1}) \\
=&f((Y_k, Z_k)|\mathcal{F}_{k-1} \setminus \mathcal{F}_{k-1-p}) \\
=&f(Y_k|\mathcal{F}_{k-1} \setminus \mathcal{F}_{k-1-p}) f(Z_k|Y_k, (\mathcal{F}_{k-1} \setminus \mathcal{F}_{k-1-p})),
\end{aligned}$$

where logistic regression could be a candidate structure for $f(Y_k|\mathcal{F}_{k-1} \setminus \mathcal{F}_{k-1-p})$.

However, in such an approach, we have to assume some structured models for $f(Y_k|\mathcal{F}_{k-1} \setminus \mathcal{F}_{k-1-p})$ and $f(Z_k|Y_k, (\mathcal{F}_{k-1} \setminus \mathcal{F}_{k-1-p}))$. This is different from the fundamental multinomial assumption of this thesis. So research in this direction requires a framework quite different from the one used in this thesis.

REFERENCES

[1] Agresti, A. (2002). *Categorical Data Analysis*, 2nd ed. New York: Wiley.

[2] Anderson, T. W. and Goodman, L. A. (1957). Statistical inference about Markov chains. *Annals of Mathematical Statistics*, **28**, 89-110.

[3] Azzalini, A. (1994). Logistic regression for autocorrelated data with application to repeated measures. *Biometrika*, **81**, 767-775.

[4] Berchtold, A. and Raftery, A. (2002). The mixture transition distribution model for high-order Markov chains and non-Gaussian time series. *Statistical Science*, **17**, 328-356.

[5] Bühlmann, P. and Wyner, A.J. (1999). Variable length Markov chains. *Annals of Statistics*, **27**, 480-513.

[6] Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference*, New York: Springer-Verlag.

[7] Cox, D. R. and Snell, E. J. (1989). *Analysis of Binary Data*, 2nd ed. London: Chapman and Hall/CRC Press.

[8] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with Discussion). *Journal of the Royal Statistical Society, Series B* , **39(1)**, 1-38.

[9] Diggle, P. J., Heagerty, P. J., Liang, K. Y., and Zeger, S. L. (2002). *Analysis of Longitudinal Data*, 2nd ed. New York: Oxford University Press.

[10] Gabriel, K. R. (1962). Ante-dependence analysis of an ordered set of variables. *Annals of Mathematical Statistics*, **33(1)**, 201-212.

[11] Heagerty, P. J. (2002). Marginalized transition models and likelihood inference for longitudinal categorical data. *Biometrics*. **58(2)**, 342-351.

[12] Heagerty, P. J. and Zeger, S. L. (1998). Lorelogram: A regression approach to exploring dependence in longitudinal categorical responses. *Journal of the American Statistical Association*, **93**, 150-162.

[13] Heagerty, P. J. and Zeger, S. L. (2000). Marginalized multilevel models and likelihood inference. *Statistical Science*, **15**, 1-19.

[14] Heckman, J. J. and Willis R. J. (1977). A Beta-logistic Model for the Analysis of Sequential Labor Force Participation by Married Women. *The Journal of Political Economy*, **85**, 27-58.

[15] Kedem B. and Fokianos K. (2002). *Regression models for time series analysis*, New York: Wiley.

[16] Lang, J. B. (2004). Multinomial-Poisson homogeneous models for contingency tables. *Annals of Statistics*, **32**, 340-383.

[17] Lee, K. and Daniels, M. J. (2007). A class of Markov models for longitudinal ordinal data. *Biometrics*, **63**, 1060-1067.

[18] Lindsey, J. K. (1993), *Models for Repeated Measurements*, New York: Oxford University Press.

[19] Macchiavelli, R. E. and Arnold, S. F. (1994). Variable-order antedependence models. *Communications in Statistics - Theory and Methods*, **23(9)**, 2683-2699.

[20] Molenberghs, G. and Verbeke G. (2005), *Models for Discrete Longitudinal Data*, New York: Springer.

[21] Schafer, J. L. (1999), *Analysis of Incomplete Multivariate Data*, Boca Raton, Florida: CRC Press.

[22] Weinberger, M.J., Rissanen, J.J., and Feder, M. (1995). A universal finite memory source. *IEEE Transactions on Information Theory*, **41**, 643-652.

[23] Williams, D.A. (1976). Improved likelihood ratio tests for complete contingency tables. *Biometrika*, **63**, 33-37.

[24] Zeger, S. L. and Qaqish, B. (1988). Markov regression models for time series: a quasi-likelihood approach. *Biometrics*, **44**, 1019-1031.

[25] Zimmerman, D. L. and Núñez-Antón, V (2010), *Antedependence Models for Longitudinal Data*, Boca Raton, Florida: CRC Press.