Published in final edited form as:

Stat Med. 2013 August 30; 32(19): . doi:10.1002/sim.5763.

Antedependence Models for Nonstationary Categorical Longitudinal Data with Ignorable Missingness: Likelihood-Based Inference

Yunlong Xie^a and Dale L. Zimmerman^b

Yunlong Xie: yunlong.xie@nih.gov

^aBiostatistics and Bioinformatics Branch, Division of Epidemiology Statistics and Prevention Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD 20892, USA

^bDepartment of Statistics and Actuarial Science, 241 Schaeffer Hall, University of Iowa, Iowa City, IA 52242, USA

Abstract

Time index-ordered random variables are said to be antedependent (AD) of order $(p_1, p_2, \dots, p_n)$ if the k th variable, conditioned on the p_k immediately preceding variables, is independent of all further preceding variables. Inferential methods associated with AD models are well developed for continuous (primarily normal) longitudinal data, but not for categorical longitudinal data. In this article, we develop likelihood-based inferential procedures for unstructured AD models for categorical longitudinal data. Specifically, we derive maximum likelihood estimators (mles) of model parameters; penalized likelihood criteria and likelihood ratio tests for determining the order of antedependence; and likelihood ratio tests for homogeneity across groups, time-invariance of transition probabilities, and strict stationarity. Closed-form expressions for mles and test statistics, which allow for the possibility of empty cells and monotone missing data, are given for all cases save strict stationarity. For data with an arbitrary missingness pattern, we derive an efficient restricted EM algorithm for obtaining mles. The performance of the tests is evaluated by simulation. The methods are applied to longitudinal studies of toenail infection severity (measured on a binary scale) and Alzheimer's disease severity (measured on an ordinal scale). The analysis of the toenail infection severity data reveals interesting nonstationary behavior of the transition probabilities and indicates that an unstructured first-order AD model is superior to stationary and other structured first-order AD models that have previously been fit to these data. The analysis of the Alzheimer's severity data indicates that the antedependence is second-order with time-invariant transition probabilities, suggesting the use of a second-order autoregressive cumulative logit model.

Keywords

Likelihood ratio test; Markov models; Missing data; Transition models

Copyright © 2012 John Wiley & Sons, Ltd.

Correspondence to: Yunlong Xie, yunlong.xie@nih.gov.

9. Supplementary Materials

Web Appendices A and B referenced in Sections 3.1.3 and 5.1, and Web Tables 1 and 2 referenced in Section 7 are available under the Paper Information link at the *Statistics in Medicine* website [http://onlinelibrary.wiley.com/journal/10.1002/\(ISSN\)1097-0258](http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1097-0258).

1. Introduction

Longitudinal data are ubiquitous in biological 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 ([1]): 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 possibly values of explanatory variables in the present and (recent) past, is modeled in terms of the given quantities. Specifically, time index-ordered random variables $Y_1, \dots, Y_n$ are said to be *antedependent of order* $p_1, p_2, \dots, p_n$, or $AD(p_1, p_2, \dots, p_n)$, if Y_k , given at least p_k immediately preceding variables, is independent of all further preceding variables for $k = 1, 2, \dots, n$ ([2], [3]). For example, if $n = 4$, if Y_3 is conditionally independent of Y_1 given Y_2 , and if Y_4 is independent of (Y_1, Y_2, Y_3) , then Y_1, Y_2, Y_3, Y_4 are $AD(0, 1, 1, 0)$. Antedependence models can be useful for nonstationary longitudinal data because they require nothing beyond the aforementioned conditional independencies; in fact, even the order of antedependence, p_k , let alone its strength, is allowed to vary over time. Note that $0 \leq p_k \leq k - 1$ necessarily, and that $AD(p_1, \dots, p_n)$ models are partially nested in the sense that $AD(p_1, \dots, p_n) \subset AD(p_1 + q_1, \dots, p_n + q_n)$ if $q_k \geq 0$ for all k . The special case in which $p_k = \min(k - 1, p)$ for some constant p is known as p -th-order antedependence and denoted as $AD(p)$. $AD(p)$ models are completely nested; that is, $AD(0) \subset AD(1) \subset \dots \subset AD(n - 1)$, with $AD(0)$ being equivalent to mutual independence and $AD(n - 1)$ being equivalent to completely general dependence. Antedependence (of specified order) is said to be *structured* if additional restrictions (such as stationarity) are imposed; otherwise it is *unstructured*.

Statistical inference procedures associated with fitting AD models to normally-distributed longitudinal data are well-developed; for a summary see [4]. A large variety of structured AD models, including stationary autoregressive models, are available, and likelihood-based inference has been widely studied for both structured models and the unstructured model. Likelihood-based inference is especially “nice” for the latter model, owing to the existence of closed-form expressions for maximum likelihood estimators (mles) of the parameters (from complete data) and the fact that various model selection and hypothesis testing procedures, such as those used to determine the order of antedependence, can be expressed simply in terms of residual sums of squares from regressions of response variables on their immediate predecessors. The procedures can be extended easily to accommodate ignorably missing data.

In contrast, inference procedures associated with fitting AD models to categorical longitudinal data are relatively underdeveloped. There are several structured AD models (most of which are stationary) available for such data, including binary Markov models ([5]), Markov generalized linear models ([6]; [1], pp. 190–207; [7], pp. 236–238), mixture transition models ([8]), and marginalized transition models ([9]; [10]; [11]; [12]). But missing from the existing toolkit are categorical-data analogues of the aforementioned inference procedures for unstructured AD models for normally-distributed longitudinal data. Although Anderson and Goodman [13] derived mles for the special case of an unstructured $AD(1)$ model for complete categorical data with no empty cells and considered some related testing problems, general expressions for mles of unstructured AD model parameters in categorical settings do not yet exist, nor are methods available for determining the order of antedependence, for determining whether transition probabilities are time-invariant, or for dealing with missing data or empty cells. The goal of this article is to fill in these gaps, so as to help lift transition-model methodology for nonstationary categorical longitudinal data to the same developmental level as its continuous (normal) counterpart.

The remainder of the article is organized as follows. In Section 2, we set notation and define several parameterizations convenient for use with AD models of categorical longitudinal data. In Section 3, we derive maximum likelihood estimators of AD model parameters in various scenarios, assuming initially that the data are complete and then relaxing this to allow for missing data. Penalized likelihood-based model selection methods and likelihood ratio tests of various hypotheses of interest are presented in Sections 4 and 5, respectively. Section 6 describes simulation studies of two of the hypothesis tests. Section 7 presents two examples that highlight the usefulness of AD models for categorical longitudinal data. Section 8 is a brief discussion.

2. Notation and parameterizations

Let N denote a predetermined, nonrandom number of subjects on which a categorical characteristic of interest is observed repeatedly over time. We assume that n times of observation are intended to be common across subjects, but we allow for missing data, i.e. the possibility that some subjects are not actually observed at some of the appointed times. Let $\mathbf{Y}_i \equiv (Y_{i1}, \dots, Y_{in})'$ denote the vector of values of the characteristic at the n observation times for the i th subject, only some of which may actually be observed. Let $1, \dots, c-2$ denote the characteristic's categories (which are assumed not to change over time). Hence $\mathbf{Y}_i$ has c^n possible outcomes, each of which corresponds to a cell in a $c \times c \times \dots \times c$ (n times) contingency table. Let Y_k denote the characteristic's value at time point k for a generic subject. For each possible outcome $(y_1, \dots, y_n)$, let $\pi_{y_1 \dots y_n} \equiv P(Y_1 = y_1, \dots, Y_n = y_n)$ denote the true cell probability with corresponding cell count $N_{y_1 \dots y_n}$ (which may or may not be observed), and put $\boldsymbol{\pi} = (\pi_{y_1 \dots y_n})$. Accordingly, $N = \sum_{(y_1, \dots, y_n) \in C_n} N_{y_1 \dots y_n}$, where $C_n \equiv \{1, \dots, c\}^n$ is the set of all c^n possible outcomes. We assume that the $\mathbf{Y}_i$'s are independently and identically distributed as Multinomial(1, $\boldsymbol{\pi}$) and that covariates, apart from indicator variables for groups, are either unavailable or not used in the analysis.

Since antedependence is defined in terms of certain conditional independencies, it is convenient to reparameterize in terms of certain conditional probabilities. Define $\pi_{y_k|y_1 \dots y_{k-1}} \equiv P(Y_k = y_k | Y_1 = y_1, \dots, Y_{k-1} = y_{k-1})$ for $k = 2, \dots, n$ and $(y_1, \dots, y_k) \in C_k$. It is easily verified that the mapping from the nonredundant cell-probability parameterization $\mathbf{Y} \equiv \{\pi_{y_1 \dots y_n} : (y_1, \dots, y_n) \in C_n \setminus \{c, \dots, c\}\}$ to the nonredundant "sequential conditional-probability" parameterization $\Theta \equiv \{\pi_{y_1 + \dots + y_1 = 1, \dots, c-1}\} \cup \{\pi_{y_k|y_1 \dots y_{k-1}} : k = 2, \dots, n; y_k = 1, \dots, c-1; (y_1, \dots, y_{k-1}) \in C_{k-1}\}$ is one-to-one. (Here and subsequently, we indicate summation over a subscripted index by replacing that index with a "+.") Moreover, under an AD($p_1, \dots, p_n$) model, for each k such that $p_k = 1$ and $k - p_k = 2$ and each fixed $(y_{k-p_k}, \dots, y_{k-1}) \in C_{p_k}$,

$$\{\pi_{y_k|y_1 \dots y_{k-p_k-1} y_{k-p_k} \dots y_{k-1}} : (y_1, \dots, y_{k-p_k-1}) \in C_{k-p_k-1}\} \text{ are equal; } \quad (1)$$

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

$$\Theta^{(p)} \equiv \bigcup_{k \ni p_k=0} \{\pi_{+\dots+y_k+\dots+y_k=1, \dots, c-1}\} \\ \bigcup_{k \ni p_k \geq 1} \{\pi_{y_k|+\dots+y_{k-p_k} \dots y_{k-1}} : y_k=1, \dots, c-1; (y_{k-p_k}, \dots, y_{k-1}) \in C_{p_k}\},$$

which we call the *transition-probability parameterization*. (Here and throughout, we may sometimes write $\mathbf{p}$ for $(p_1, \dots, p_n)$.) It is easily verified that the dimensionality of $\Theta(\mathbf{p})$ is $(c-1)\sum_{k=1}^n c^{p_k}$. Finally, let $\bar{\Theta}(\mathbf{p})$ denote the extended set of parameters obtained by adding to $\Theta(\mathbf{p})$ the redundant probabilities, i.e. probabilities of the same form as those in $\Theta(\mathbf{p})$ but with $y_k = c$.

To clarify the notation, consider a case in which binary ($c = 2$) longitudinal data are observed on 4 occasions. In this case,

$$\Upsilon = \{\pi_{1111}, \pi_{1112}, \pi_{1121}, \pi_{1122}, \pi_{1211}, \pi_{1212}, \pi_{1221}, \pi_{1222}, \pi_{2111}, \pi_{2112}, \pi_{2121}, \pi_{2122}, \pi_{2211}, \pi_{2212}, \pi_{2221}\} \text{ and} \\ \Theta = \{\pi_{1+++}, \pi_{1|1}, \pi_{1|2}, \pi_{1|11}, \pi_{1|12}, \pi_{1|21}, \pi_{1|22}, \pi_{1|111}, \pi_{1|112}, \pi_{1|121}, \pi_{1|122}, \pi_{1|211}, \pi_{1|212}, \pi_{1|221}, \pi_{1|222}\}.$$

Furthermore, under an AD(0,1,1,1) [or equivalently AD(1)] model,

$$\pi_{1|11} = \pi_{1|21} (\equiv \pi_{1|+1}), \quad \pi_{1|12} = \pi_{1|22} (\equiv \pi_{1|+2}), \quad \pi_{1|111} = \pi_{1|121} = \pi_{1|211} = \pi_{1|221} (\equiv \pi_{1|++1}), \quad \text{and} \\ \pi_{1|112} = \pi_{1|122} = \pi_{1|212} = \pi_{1|222} (\equiv \pi_{1|++2}).$$

Thus,

$$\bar{\Theta}^{(0,1,1,1)} = \{\pi_{1+++}, \pi_{2+++}, \pi_{1|1}, \pi_{2|1}, \pi_{1|2}, \pi_{2|2}, \pi_{1|+1}, \pi_{2|+1}, \pi_{1|+2}, \pi_{2|+2}, \pi_{1|++1}, \pi_{2|++1}, \pi_{1|++2}, \pi_{2|++2}\} \text{ and} \\ \bar{\Theta}^{(0,1,1,1)} = \{\pi_{1+++}, \pi_{2+++}, \pi_{1|1}, \pi_{2|1}, \pi_{1|2}, \pi_{2|2}, \pi_{1|+1}, \pi_{2|+1}, \pi_{1|+2}, \pi_{2|+2}, \pi_{1|++1}, \pi_{2|++1}, \pi_{1|++2}, \pi_{2|++2}\}.$$

3. Maximum likelihood estimation

In this section, we give several results pertaining to maximum likelihood estimation of the transition-probability parameterization of an AD($p_1, \dots, p_n$) process under the multinomial sampling framework described previously. Should mles of the corresponding cell-probability parameterization be desired, they may be obtained using the relation

$$\pi_{y_1 \dots y_n} = \prod_{k=1}^n \left(I(p_k=0) \pi_{+\dots+y_k+\dots+} + I(p_k \geq 1) \pi_{y_k|+\dots+y_{k-p_k} \dots y_{k-1}} \right) \text{ [where } I(\cdot) \text{ is an indicator function]} \text{ and the parameterization-invariance of maximum likelihood estimation.}$$

Initially we assume that the data are complete, subsequently extending to permit missing data; throughout, however, we allow for empty cells.

3.1. Complete data

3.1.1. Unstructured antedependence

Theorem 1: Under AD($p_1, \dots, p_n$), complete-data mles of the parameters of $\Theta(\mathbf{p})$ are as follows: for k such that $p_k = 0$, $\hat{\pi}_{+\dots+y_k+\dots+}^{(\mathbf{p})} = N_{+\dots+y_k+\dots+}/N$; for other k ,

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

Proof: The likelihood function in terms of the parameters of $\Theta^{(\bar{p}1 \cdots pn)}$ is proportional to

$$= \prod_{k=1}^n \left[\left(I(p_k=0) \prod_{y_k=1}^c \pi_{+ \cdots + y_k + \cdots +}^{N_{+ \cdots + y_k + \cdots +}} + I(p_k \geq 1) \prod_{(y_{k-p_k}, \dots, y_k) \in C_{p_k+1}} \pi_{y_k | + \cdots + y_{k-p_k} \cdots y_{k-1}}^{N_{y_1 \cdots y_n}} \right) \right] \quad (2)$$

For k such that $p_k = 0$, the k th term of the outermost product in (2) is the kernel of the likelihood of a saturated c -nomial distribution with cell probabilities $\{\pi_{+ \cdots + y_k + \cdots +} : y_k = 1, \dots, c\}$; for other k , the k th term is the product of c^{p_k} independent likelihood kernels, each corresponding to a saturated c -nomial distribution with cell probabilities $\{\pi_{y_k | + \cdots + y_{k-p_k} \cdots y_{k-1}} : y_k = 1, \dots, c\}$. The cell probabilities for each kernel sum to one and lie within $[0, 1]$, but are not otherwise constrained. Thus, by well-known results on maximum likelihood estimation for saturated multinomial distributions, we have the

following: (a) for those k such that $p_k = 0$, $\hat{\pi}_{+ \cdots + y_k + \cdots +}^{(p)} = N_{+ \cdots + y_k + \cdots +} / N$; (b) for other k , if $N_{+ \cdots + y_k - p_k \cdots y_{k-1} + \cdots +} = 0$ then $N_{+ \cdots + y_k - p_k \cdots y_k + \cdots +} = 0$, implying that

$$\hat{\pi}_{y_k | + \cdots + y_{k-p_k} \cdots y_{k-1}}^{(p)} = 0, \text{ and if } N_{+ \cdots + y_k - p_k \cdots y_{k-1} + \cdots +} = 0 \text{ then } \hat{\pi}_{y_k | + \cdots + y_{k-p_k} \cdots y_{k-1}}^{(p)} = N_{+ \cdots + y_k - p_k \cdots y_k + \cdots +} / N_{+ \cdots + y_k - p_k \cdots y_{k-1} + \cdots +}. \text{ This completes the proof.}$$

Observe that for each fixed $k \in \{1, \dots, n\}$, the mles of the marginal probabilities $\{\pi_{+ \cdots + y_k + \cdots +} : y_k = 1, \dots, c\}$ and conditional probabilities $\{\pi_{y_k | + \cdots + y_{k-p_k} \cdots y_{k-1}} : (y_{k-p_k}, \dots, y_{k-1}, y_k) \in C_{p_k+1}\}$ under $\text{AD}(p_1, \dots, p_n)$ depend on $(p_1, \dots, p_n)$ only through the value of p_k . Also note that upon substituting $\min(k-1, p)$ for p_k ($k = 1, \dots, n$) in Theorem 1 and writing $\Theta^{(\bar{p})}$ for $\Theta^{(0,1, \dots, p, p, \dots, p)}$, we obtain the following corollary for $\text{AD}(p)$ variables. The special case of this corollary when $p = 1$ and no cells are empty was given by [13].

Corollary 1: Under $\text{AD}(p)$, complete-data mles of the parameters of $\Theta^{(\bar{p})}$ are as follows: if $p = 0$, $\hat{\pi}_{+ \cdots + y_k + \cdots +}^{(p)} = N_{+ \cdots + y_k + \cdots +} / N$ for $k = 1, \dots, n$; if $p \geq 1$, $\hat{\pi}_{y_1 + \cdots +}^{(p)} = N_{y_1 + \cdots +} / N$,

$$\hat{\pi}_{y_k | y_1 \cdots y_{k-1}}^{(p)} = \begin{cases} 0 & \text{if } N_{y_1 \cdots y_{k-1} + \cdots +} = 0 \\ \frac{N_{y_1 \cdots y_k + \cdots +}}{N_{y_1 \cdots y_{k-1} + \cdots +}} & \text{otherwise} \end{cases}$$

for $k = 2, \dots, p+1$, and

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

for $k = p+2, \dots, n$.

3.1.2. Time-invariant transition probabilities—Consider a situation with a constant order of antedependence, p . If measurement times are equally spaced, it may be of interest to estimate parameters under $\text{AD}(p)$ with a stationarity property imposed. One such property is that of time-invariant transition probabilities. If $p \geq 1$ and $\pi_{y_{p+1}^{(k)} | y_1^{(k-p)} \cdots y_p^{(k-1)}}^{(k)}$ denotes $P(Y_k =$

$y_{p+1}|Y_{k-p}=y_1, \dots, Y_{k-1}=y_p$ for $k = p+1, \dots, n$, the p th-order transition probabilities are said to be time-invariant if

$$\pi_{y_{p+1}|y_1^{(1)} \dots y_p^{(p)}} = \pi_{y_{p+1}^{(2)}|y_1^{(2)} \dots y_p^{(p+1)}} = \dots = \pi_{y_{p+1}^{(n)}|y_1^{(n-p)} \dots y_p^{(n-1)}} \text{ for all } (y_1, \dots, y_{p+1}) \in C_{p+1}. \quad (3)$$

Theorem 2: Under AD(p) with $p \geq 1$ and time-invariant p th-order transition probabilities, the complete-data mle of $\pi_{y_1 \dots y_p + \dots +}^{(p)}$ is $\hat{\pi}_{y_1 \dots y_p + \dots +}^{(p)} = N_{y_1 \dots y_p + \dots +} / N$, and the complete-data mle of the common p th-order transition probability, denoted by $\hat{\pi}_{y_{p+1}|y_1 \dots y_p}^{(p)}$, is as follows:

$$\hat{\pi}_{y_{p+1}|y_1 \dots y_p}^{(p)} = \begin{cases} 0 & \text{if } \sum_{k=p+1}^n N_{+\dots+y_1^{(k-p)} \dots y_p^{(k-1)} + \dots +} = 0 \\ \frac{\sum_{k=p+1}^n N_{+\dots+y_1^{(k-p)} \dots y_p^{(k-1)} + \dots +}}{\sum_{k=p+1}^n N_{+\dots+y_1^{(k-p)} \dots y_p^{(k-1)} + \dots +}} & \text{otherwise.} \end{cases}$$

Thus, when p th-order transition probabilities are time-invariant, they may be pooled over time to yield the mle of the common p th-order transition probability.

3.1.3. Strict stationarity—Variables $Y_1, \dots, Y_n$ are said to be strictly stationary if the joint probabilities of all events are invariant to time shifts. The following lemma (proved in Web Appendix A) gives necessary and sufficient conditions for strict stationarity under AD(p) in terms of the transition-probability parameterization.

Lemma 1: Under AD(p) with $p \geq 1$, the variables $Y_1, \dots, Y_n$ are strictly stationary if and only if the transition probabilities are time-invariant and

$$\pi_{y_1 \dots y_p + \dots +} = \sum_{y_0=1}^c \pi_{y_p|y_0 y_1 \dots y_{p-1}} \pi_{y_0 y_1 \dots y_{p-1} + \dots +} \text{ for all } (y_1, \dots, y_p) \in C_p \setminus \{c, \dots, c\}. \quad (4)$$

Unfortunately, mles of neither the transition nor cell probabilities of a strictly stationary AD(p) model can be expressed in closed form. However, since the constraints imposed by (1), (3), and (4) may be written as nonredundant homogeneous smooth functions of the expected cell counts, the mles can be obtained numerically using the algorithm of [14].

3.2. Missing data

3.2.1. Monotone missingness—All of the expressions for complete-data mles given to this point 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 ($i = 1, \dots, N$; $k = 2, \dots, n-1$). Let $N^{(k)\bullet}$ denote the number of subjects having complete observations between

time points 1 and k (inclusive), and let $N_{+\dots+y_{k-p_k} \dots y_k + \dots +}^{(k)\bullet}$ denote the number of these subjects for which $Y_{k-p_k} = y_{k-p_k}, \dots, Y_k = y_k$ (regardless of whether $Y_{k+1}, \dots, Y_n$ are observed or missing). Then, the observed-data likelihood is given by an expression identical to (2)

except that $N_{+\dots+y_k + \dots +}^{(k)\bullet}$ and $N_{+\dots+y_{k-p_k} \dots y_k + \dots +}^{(k)\bullet}$ are substituted for the corresponding complete-data counts. Hence, under AD($p_1, \dots, p_n$), the monotone-missing-data mles

(assuming ignorability) of the parameters of $\Theta(\bar{\mathbf{p}})$, denoted by $\hat{\pi}_{+\dots+y_k + \dots +}^{(\bar{\mathbf{p}})\bullet}$ and

$\hat{\pi}_{y_k|+\dots+y_{k-p_k}\dots y_{k-1}}^{(p)\bullet}$, are given by expressions identical to those in Theorem 1 except that $N^{(k)\bullet}$, $N_{+\dots+y_k+\dots+}^{(k)\bullet}$, $N_{+\dots+y_{k-p_k}\dots y_k+\dots+}^{(k)\bullet}$, and $N_{+\dots+y_{k-p_k}\dots y_{k-1}+\dots+}^{(k)\bullet}$ are substituted for the corresponding complete-data counts. Similarly, under AD(p), the monotone-missing-data mles of the parameters of $\Theta^{(p)}$ are obtained by substituting the analogous quantities into Corollary 1; furthermore, monotone-missing-data mles of time-invariant transition probabilities are obtained by substituting those same quantities into Theorem 2.

Mles under AD(p) may also be obtained easily for ignorably 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 ($i = 1, \dots, N$; $k = 2, \dots, n - 1$). For such data, mles are as noted above (with $p_k = \min(k - 1, p)$) but applied to the data in reverse time order. This follows from the fact that AD(p) random variables are also AD(p) when arranged in reverse time order ([4], p. 151). There is not an analogous reversibility result, however, for the more general case of AD($p_1, \dots, p_n$).

3.2.2. Arbitrary missingness—When the data have an arbitrary ignorable pattern of missingness, a restricted EM algorithm [15] may be used to obtain mles under AD($p_1, \dots, p_n$). Customized to the present setting, the algorithm exploits the fact that the AD($p_1, \dots, p_n$) multinomial model can be regarded as a linearly constrained case of the saturated multinomial model, specifically one with linear equality constraints on the parameters of the sequential conditional-probability parameterization, as specified by (1). The constraints may be written as $\mathbf{A}\boldsymbol{\theta} = \mathbf{0}$, where $\boldsymbol{\theta}$ is the vector of nonredundant sequential conditional probabilities and $\mathbf{A}$ is a matrix of ones, zeros, and minus ones corresponding to the prescribed order of antedependence; for example, for a binary AD(0,1,1) model, $\boldsymbol{\theta} = (\pi_{1++}, \pi_{1/1}, \pi_{1/2}, \pi_{1/11}, \pi_{1/21}, \pi_{1/12}, \pi_{1/22})'$ and

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}.$$

Partition the cell counts $\mathbf{N} = (N_{y_1 \dots y_n})$ into $(\mathbf{N}'_{obs}, \mathbf{N}'_{unobs})'$, their observed and unobserved parts. Then the conditional expectation of the complete-data log-likelihood given the observed data and the current parameter estimate may be written as

$$Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(m)}) = \sum_{(y_1, \dots, y_n) \in C_n} E(N_{y_1 \dots y_n} | \mathbf{N}_{obs}, \hat{\boldsymbol{\theta}}^{(m)}) (\log \pi_{y_1+\dots+} + \log \pi_{y_2|y_1} + \log \pi_{y_3|y_1 y_2} + \dots + \log \pi_{y_n|y_1 \dots y_{n-1}}).$$

The restricted EM algorithm, in general form, is then as follows (a detailed illustration for a specific case may be found in Web Appendix B):

E step—Evaluate $Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(m)})$. (The initial restricted estimate, $\hat{\boldsymbol{\theta}}^{(0)}$, may be taken to correspond to uniform cell probabilities, i.e. $\pi_{y_1 \dots y_n} \equiv c^{-n}$.)

Restricted M step

1. Obtain the current unrestricted estimate, $\hat{\boldsymbol{\theta}}_U$, by maximizing $Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(m)})$ using explicit expressions given by Schafer ([16], sec. 7.3) for the estimated cell probabilities of a saturated multinomial distribution and the relationship between $\boldsymbol{\pi}$

and θ described in Section 2, and calculate $\mathbf{I}_U = - \frac{\partial^2 Q(\theta | \hat{\theta}_R^{(m)})}{\partial \theta^2} \bigg|_{\theta = \hat{\theta}_R^{(m)}}$. It is worth noting that $\mathbf{I}_U$ is diagonal due to the linearity of $Q(\theta | \hat{\theta}_R^{(m)})$ in the logs of the elements of θ .

2. Obtain a provisional restricted estimate, $\hat{\theta}_R$, via the equation

$$\hat{\theta}_R = \hat{\theta}_U - \mathbf{I}_U^{-1} \mathbf{A}' (\mathbf{A} \mathbf{I}_U^{-1} \mathbf{A}')^{-1} \mathbf{A} \hat{\theta}_U.$$

3. If $Q(\hat{\theta}_R | \hat{\theta}_R^{(m)}) > Q(\hat{\theta}_R^{(m)} | \hat{\theta}_R^{(m)})$, then $\hat{\theta}_R^{(m+1)} = \hat{\theta}_R$.
4. If not, then do step halving (and, if necessary, step quartering, etc.) on $\Delta \hat{\theta}_U$, where $\Delta \hat{\theta}_U = \hat{\theta}_U - \hat{\theta}_R^{(m)}$.

The iteration between the E-step and restricted M-step is continued, and the estimate of θ updated as long as $Q(\hat{\theta}_R | \hat{\theta}_R^{(m)})$ increases, until a convergence criterion is satisfied.

Restricted EM algorithms similar to the one just described may be devised without difficulty for use with the time-invariant transition probability AD(p) model as well, since for this model the constraints on the parameters likewise are linear. However, the strictly stationary AD($p_1, \dots, p_n$) multinomial model imposes nonlinear constraints (4) in addition to linear constraints (1) and (3), so the same restricted EM algorithm is not applicable to it.

4. Penalized likelihood-based model selection

Consider the same data-model situation as that considered in the previous section, but now suppose that the order ($p_1, \dots, p_n$) of the AD model is not known and it is the goal of the investigator to determine the order that best fits the data, in some sense. Standard likelihood-based hypothesis testing is not entirely suitable for this purpose because AD($p_1, \dots, p_n$) models are not completely nested. As an alternative, in this section we propose the use of penalized likelihood criteria; specifically, we consider information criteria of the form

$$IC(p_1, \dots, p_n) = a(N) (c-1) \sum_{k=1}^n c^{p_k} - 2 \log L(\hat{\pi}^{(p)}), \text{ where } a(N) \text{ is a specified function of } N.$$

Many well-known penalized likelihood criteria, when applied to AD($p_1, \dots, p_n$) models, are cases of this general form, including AIC, BIC, corrected AIC, quasi-AIC, and quasi-BIC. Note that the criteria are expressed in “smaller is better” form. In our example we feature AIC, for which $a(N) = 2$.

Since p_k can be any nonnegative integer less than or equal to $k-1$, the number of 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; for example, 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($p_1, \dots, p_n$) for all 5040 models. However, it turns out that the AD model that minimizes $IC(p_1, \dots, p_n)$ 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, which follows immediately from (2), Theorem 1, and results discussed in Section 3.2. In the theorem and subsequently, put

$$C_{u,k}^+ = \{(y_u, y_{u+1}, \dots, y_k) \in C_{k-u+1:N+\dots+y_u \dots y_k+\dots} > 0\} \text{ and write } C_k^+ \text{ for } C_{k,k}^+.$$

Theorem 3

The minimizer of $IC(p_1, \dots, p_n)$ is given by $\hat{p}_k = \underset{p_k=0, \dots, k-1}{\operatorname{argmin}} IC(p_k)$ for $k = 1, \dots, n$, where, if the data are complete,

$$IC(p_k) = \frac{a(N)}{2}(c-1)c^{p_k} - \left[I(p_k=0) \sum_{y_k \in C_k^+} N_{+\dots+y_k+\dots} \log \left(\frac{N_{+\dots+y_k+\dots}}{N} \right) \right. \quad (5)$$

$$\left. + I(p_k \geq 1) \sum_{(y_k-p_k, \dots, y_k) \in C_{k-p_k, k}^+} N_{+\dots+y_k-p_k \dots y_k+\dots} \log \left(\frac{N_{+\dots+y_k-p_k \dots y_k+\dots}}{N_{+\dots+y_k-p_k \dots y_{k-1}+\dots}} \right) \right]. \quad (6)$$

If the data are ignorably monotone missing, then $N_{+\dots+y_k+\dots}$, $N_{+\dots+y_k-p_k \dots y_k+\dots}$, and $N_{+\dots+y_k-p_k \dots y_{k-1}+\dots}$ in (5) and (6) are replaced by $N_{+\dots+y_k+\dots}^{(k)\bullet}$, $N_{+\dots+y_k-p_k \dots y_k+\dots}^{(k)\bullet}$, and $N_{+\dots+y_k-p_k \dots y_{k-1}+\dots}^{(k)\bullet}$; if the missingness is arbitrary, then those same quantities are replaced by their mles $N_{+\dots+y_k+\dots}$, $N_{+\dots+y_k-p_k \dots y_k+\dots}$, and $N_{+\dots+y_k-p_k \dots y_{k-1}+\dots}$ under $AD(p_1, \dots, p_n)$, which may be obtained using the restricted EM algorithm as described in Section 3.2.2.

5. Hypothesis tests

Next we consider likelihood ratio tests of hypotheses concerning antedependence under the same data-model situation of the previous section. We first consider tests for the order of antedependence, beginning with the most general case of nested variable-order models and then specializing to constant-order models. Then we consider tests for homogeneity of antedependence across groups and for two forms of stationarity. For simplicity of presentation, we assume in this section that the data are complete, unless noted otherwise. For results on limiting distributions of test statistics, we assume standard multinomial sampling asymptotics, including fixed c and n as $N \rightarrow \infty$, and all cell probabilities positive.

5.1. Order of antedependence

Consider testing $AD(p_1, \dots, p_n)$ as a null hypothesis against $AD(q_1, \dots, q_n)$ as the alternative hypothesis, where $p_k < q_k$ for all k and the inequality is strict for at least one k . Using (2) and Theorem 1, it is easily verified that the likelihood ratio test statistic is given by

$$G_{\mathbf{p}, \mathbf{q}}^2 \equiv 2 \sum_{k=2}^n I(q_k > p_k) \sum_{(y_k-q_k, \dots, y_k) \in C_{k-q_k, k}^+} N_{+\dots+y_k-q_k \dots y_k+\dots} \log \left(\frac{\hat{\pi}_{y_k+\dots+y_k-q_k \dots y_{k-1}}^{(q)}}{\hat{\pi}_{y_k+\dots+y_k-q_k \dots y_{k-1}}^{(p)}} \right)$$

Where $\hat{\pi}_{y_k+\dots+y_k-q_k \dots y_{k-1}}^{(\mathbf{p})} = I(p_k=0) \hat{\pi}_{+\dots+y_k+\dots}^{(\mathbf{p})} + I(p_k \geq 1) \hat{\pi}_{y_k+\dots+y_k-p_k \dots y_{k-1}}^{(\mathbf{p})}$. The

limiting null distribution of $G_{\mathbf{p}, \mathbf{q}}^2$ is chi-square with

$d(\mathbf{p}, \mathbf{q}) \equiv \dim(\Theta^{(\mathbf{q})}) - \dim(\Theta^{(\mathbf{p})}) = (c-1) \sum_{k=2}^n (c^{q_k} - c^{p_k})$ degrees of freedom. Hence the approximate size- α likelihood ratio test for $AD(p_1, \dots, p_n)$ versus $AD(q_1, \dots, q_n)$ rejects the former if $G_{\mathbf{p}, \mathbf{q}}^2$ exceeds the $100(1 - \alpha)$ th percentile of this chi-square distribution. However, the simulation study presented in the next section indicates that the Type I error rate of this

test is somewhat larger than α . Aided by the following lemma (proved in Web Appendix B), we use Lawley's [17] strategy of multiplying the likelihood ratio test statistic by a scale factor to obtain a modified version of the likelihood ratio test for which the Type I error rate is closer to its nominal level.

Lemma 2— $E(G_{\mathbf{p},\mathbf{q}}^2) = e(\mathbf{p}, \mathbf{q}) + O(N^{-2})$, where

$$e(\mathbf{p}, \mathbf{q}) = 2 \sum_{k=2}^n I(q_k > p_k) \left[\sum_{(y_{k-q_k}, \dots, y_k) \in C_{k-q_k, k}^+} \left(f(N, \pi_{+\dots+y_{k-q_k} \dots y_k + \dots}^{(\mathbf{q})}) - \pi_{y_k | + \dots + y_{k-q_k} \dots y_{k-1}}^{(\mathbf{q})} f(N, \pi_{+\dots+y_{k-q_k} \dots y_{k-1} + \dots}^{(\mathbf{q})}) \right) - \sum_{(y_{k-p_k}, \dots, y_k) \in C_{k-p_k, k}^+} \left(f(N, \pi_{+\dots+y_{k-p_k} \dots y_k + \dots}^{(\mathbf{p})}) - \pi_{y_k | + \dots + y_{k-p_k} \dots y_{k-1}}^{(\mathbf{p})} f(N, \pi_{+\dots+y_{k-p_k} \dots y_{k-1} + \dots}^{(\mathbf{p})}) \right) \right]$$

and

$$f(N, \pi) = N \pi \log(N \pi) + \frac{1-\pi}{2} + \frac{1-\pi^2}{12N\pi}.$$

Based on Lemma 2, we define the modified likelihood ratio test statistic as

$$\bar{G}_{\mathbf{p},\mathbf{q}}^2 = \frac{d(\mathbf{p}, \mathbf{q})}{\hat{e}(\mathbf{p}, \mathbf{q})} G_{\mathbf{p},\mathbf{q}}^2$$

where $\hat{e}(\mathbf{p}, \mathbf{q})$ is obtained by replacing the unknown probabilities in $e(\mathbf{p}, \mathbf{q})$ with their mles under $\text{AD}(q_1, \dots, q_n)$.

In the special case of testing $\text{AD}(p)$ versus $\text{AD}(q)$, where $0 \leq p < q \leq n-1$, the likelihood ratio test statistic becomes

$$G_{p,q}^2 \equiv 2 \left[\sum_{k=p+1}^q \sum_{(y_1, \dots, y_k) \in C_{1,k}^+} N_{y_1 \dots y_k + \dots} + \log \left(\frac{\hat{\pi}_{y_k | y_1 \dots y_{k-1}}^{(q)}}{\hat{\pi}_{y_k | y_1 \dots y_{k-1}}^{(p)}} \right) \right] + \sum_{k=q+1}^n \sum_{(y_{k-q}, \dots, y_k) \in C_{k-q, k}^+} N_{+\dots+y_{k-q} \dots y_k + \dots} + \log \left(\frac{\hat{\pi}_{y_k | + \dots + y_{k-q} \dots y_{k-1}}^{(q)}}{\hat{\pi}_{y_k | + \dots + y_{k-q} \dots y_{k-1}}^{(p)}} \right),$$

for which the limiting null chi-square distribution has $d(p, q) \equiv (n-p-1)c^p - (n-p)c^{p+1} - (n-q-1)c^q + (n-q)c^{q+1}$ degrees of freedom. The case $q = p+1$ is worthy of special mention. It can be easily seen that the likelihood ratio test criteria for the sequence of tests of $\text{AD}(p)$ versus $\text{AD}(p+1)$ ($p = 0, 1, \dots, n-2$) sum to the likelihood ratio test criterion for $\text{AD}(0)$ versus $\text{AD}(n-1)$. Thus, this sequence can be viewed as a decomposition of the well-known likelihood ratio test for complete independence (against the fully saturated alternative) into steps of order one. This suggests two practical strategies for selecting the order of antedependence. Analogously to 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(1), and if AD(0) is rejected then testing 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($n - 2$), etc.]

If data are (ignorably) missing, $G_{p,q}^2$ may be generalized as follows:

$$G_{p,q}^2 = 2 \sum_{k=2}^n I(q_k > p_k) \left(\sum_{(y_{k-q_k}, \dots, y_k) \in C_{k-q_k, k}^+} \hat{N}_{+\dots+y_{k-q_k} \dots y_k + \dots}^{(q)} + \log \hat{\pi}_{y_k | + \dots + y_{k-q_k} \dots y_{k-1}}^{(q)} \right. \\ \left. - \sum_{(y_{k-p_k}, \dots, y_k) \in C_{k-p_k, k}^+} \hat{N}_{+\dots+y_{k-p_k} \dots y_k + \dots}^{(p)} + \log \hat{\pi}_{y_k | + \dots + y_{k-p_k} \dots y_{k-1}}^{(p)} \right)$$

where the superscript on each estimated cell count and transition probability indicates the model under which the restricted EM algorithm is applied to obtain the estimate. Similar generalizations to accommodate missing data may be made to all likelihood ratio tests presented herein, so henceforth we display only complete-data versions of test statistics.

5.2. Homogeneity across groups

Suppose that the subjects can be classified into s distinct groups, and that the orders of antedependence are identically ($p_1, \dots, p_n$) across groups. Let N^g and $N_{y_1 \dots y_n}^g$ denote the sample size and the observed cell count for outcome ($y_1, \dots, y_n$), respectively, for group g , and consider testing whether the parameters are homogeneous across groups, i.e. testing H_0 :

$H_0: \pi_{y_1 + \dots +}^{g, (p)} = \pi_{y_1 + \dots +}^{(p)}$ and $\pi_{y_k | + \dots + y_{k-p_k} \dots y_{k-1}}^{g, (p)} = \pi_{y_k | + \dots + y_{k-p_k} \dots y_{k-1}}^{(p)}$ ($k=2, \dots, n; g=1, \dots, s$) against $H_1: \text{not } H_0$. Under H_0 , the mles are obtained merely by pooling across groups. Under H_1 , the likelihood is obtained as the product of the s AD($p_1, \dots, p_n$) multinomial likelihoods and is maximized by $\hat{\pi}_{+\dots+y_k + \dots +}^{g, (p)} = N_{+\dots+y_k + \dots +}^g / N^g$ and

$$\hat{\pi}_{y_k | + \dots + y_{k-p_k} \dots y_{k-1}}^{g, (p)} = \begin{cases} 0 & \text{if } N_{+\dots+y_{k-p_k} \dots y_{k-1} + \dots}^g = 0 \\ \frac{N_{+\dots+y_{k-p_k} \dots y_{k-1} + \dots}^g}{N_{+\dots+y_{k-p_k} \dots y_{k-1} + \dots}^g} & \text{otherwise} \end{cases}$$

for $g = 1, \dots, s$. Thus the likelihood ratio statistic is

$$G_{hom, p}^2 = 2 \sum_{g=1}^s \sum_{k=1}^n \left[I(p_k = 0) \sum_{y_k \in C_k^{g, +}} N_{+\dots+y_k + \dots}^g + \log \left(\frac{\hat{\pi}_{+\dots+y_k + \dots}^{g, (p)}}{\hat{\pi}_{+\dots+y_k + \dots}^{(p)}} \right) \right] \\ + I(p_k \geq 1) \sum_{(y_{k-p_k}, \dots, y_k) \in C_{k-p_k, k}^{g, +}} N_{+\dots+y_{k-p_k} \dots y_k + \dots}^g + \log \left(\frac{\hat{\pi}_{y_k | + \dots + y_{k-p_k} \dots y_{k-1}}^{g, (p)}}{\hat{\pi}_{y_k | + \dots + y_{k-p_k} \dots y_{k-1}}^{(p)}} \right) \Big].$$

Here $C_k^{g, +}$ and $C_{k-p_k, k}^{g, +}$ have similar meaning as C_k^+ and $C_{k-p_k, k}^+$ but for group g only. The

limiting null distribution of both test statistics is chi-square with $(s-1)(c-1) \sum_{k=1}^n c^{p_k}$ degrees

of freedom. As with likelihood ratio tests for order of antedependence, Type I error rates associated with $G_{hom,p}^2$ are somewhat larger than nominal levels. A result similar to Lemma 2 can be obtained, leading to a modified likelihood ratio test for homogeneity whose actual and nominal sizes agree more closely.

5.3. Time-invariant transition probabilities

Consider testing the null hypothesis of time-invariant p th-order transition probabilities, as specified by (3), against the alternative of an unstructured AD(p) model. By Corollary 1, the unrestricted mle of a generic p th-order transition probability at time k under AD(p) is

$\hat{\pi}_{y_{p+1}|y_1^{(k-p)} \dots y_p^{(k-1)}}^{(p)} = N_{+\dots+y_1^{(k-p)} \dots y_p^{(k-1)} y_{p+1}^{(k)} + \dots + N_{+\dots+y_1^{(k-p)} \dots y_p^{(k-1)} + \dots +}$ (unless the denominator is zero, in which case the mle is equal to 0). Thus, the likelihood ratio test statistic is

$$G_t^2 = -2 \sum_{(y_1, \dots, y_{p+1})} \sum_k N_{+\dots+y_1^{(k-p)} \dots y_p^{(k-1)} y_{p+1}^{(k)} + \dots +} \log \left(\frac{\hat{\pi}_{y_{p+1}|y_1 \dots y_p}^{(p)}}{\hat{\pi}_{y_{p+1}|y_1^{(k-p)} \dots y_p^{(k-1)}}^{(p)}} \right)$$

where the summation is over those $(y_1, \dots, y_{p+1}) \in C_{p+1}$ and $k = p+1, \dots, n$ such that $N_{+\dots+y_1^{(k-p)} \dots y_p^{(k-1)} y_{p+1}^{(k)} + \dots +} > 0$. The limiting null distribution of G_t^2 is chi-square with degrees of freedom equal to the number of nonredundant conditions in (3), or $(c-1)[n-(p+1)]c^p$.

5.4. Strict stationarity

Recall from Section 3.1.3 that the mles of cell probabilities of an AD(p) model under strict stationarity cannot be expressed in closed form, but they can be obtained numerically using Lang's [14] algorithm. Let $\hat{\pi}^{s,(p)}$ be the vector of mles so obtained. Then, the likelihood ratio statistic for testing for strict stationarity under an unstructured AD(p) model is

$$G_s^2 \equiv -2 \log \left(\frac{L(\hat{\pi}^{s,(p)})}{L(\hat{\pi}^{(p)})} \right). \text{ The limiting null distribution of } G_s^2 \text{ is chi-square with } (c-1)[n-(p+1)]c^p + (c^p - 1) \text{ degrees of freedom.}$$

6. Simulation studies

Here we present simulation studies to evaluate the performance of the likelihood ratio test (and its modified version), with and without missing data, for two of the hypotheses described in Section 5. Studies of the tests for the other hypotheses yield similar conclusions and are not reported here.

6.1. Order of antedependence

Simulations were generated from two binary linear AD processes observed at four time points. The first process is defined as follows:

$$Y_1 = \begin{cases} 1 & \text{w.p. } \frac{1}{2} \\ 2 & \text{w.p. } \frac{1}{2}, \end{cases} \quad Y_2|Y_1 = \begin{cases} Y_1 & \text{w.p. } \frac{1}{3} \\ 1 & \text{w.p. } \frac{1}{3} \\ 2 & \text{w.p. } \frac{1}{3}, \end{cases} \quad (7)$$

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

Here $\theta \in [0, 2]$ controls the degree of departure from first-order antedependence: when $\theta = 0$ the process is AD(1), when $\theta > 0$ the process is AD(3), and as θ increases, the departure from AD(1) is larger in the sense that the conditional (on intervening variables) dependence between variables lagged two or three time points apart increases. Note that the marginal distribution of $Y_k - 1$ is Bernoulli($\frac{1}{2}$) for all k , regardless of θ . The second process differs from the first only by redefining Y_3 and Y_4 as follows:

$$Y_3|(Y_1, Y_2) = \begin{cases} Y_1 & \text{w.p. } \frac{1}{4} \\ Y_2 & \text{w.p. } \frac{1}{4} \\ 1 & \text{w.p. } \frac{1}{4} \\ 2 & \text{w.p. } \frac{1}{4}, \end{cases} \quad Y_4|(Y_1, Y_2, Y_3) = \begin{cases} Y_1 & \text{w.p. } \frac{\theta}{4} \\ Y_2 & \text{w.p. } \frac{2-\theta}{8} \\ Y_3 & \text{w.p. } \frac{2-\theta}{8} \\ 1 & \text{w.p. } \frac{1}{4} \\ 2 & \text{w.p. } \frac{1}{4}. \end{cases} \quad (9)$$

This difference renders the process AD(2) [rather than AD(1)] when $\theta = 0$.

For each process, each of four sample sizes ($N = 50, 100, 200, 1000$), and each θ over a range of values, 10,000 samples were generated, and (nominal) size-0.05 likelihood ratio and modified likelihood ratio tests of AD(1) versus AD(3) for (7)&(8) and of AD(2) versus AD(3) for (7)&(9) were performed. In order to investigate the effect of missing data, we also performed the tests after deleting 25% of the observations in each sample, subject to the constraint that none of the subjects have more than two observations deleted. Empirical rejection rates are displayed in Table 1. These indicate that the actual size of the likelihood ratio test is significantly higher than the nominal size when $N = 200$. In contrast, the actual size of the modified likelihood ratio test is consistently at or below the nominal size. As for power, it is perceptibly smaller for the modified test except when $N = 1000$. For both processes and both tests, the power increases with N and θ , as expected. Also as expected, there is a measurable loss of power when 25% of the observations are missing. Finally, a comparison of the results for the two processes suggests that the power for detecting a departure from a specified order of antedependence is higher when the difference in order for that departure is larger.

6.2. Time-invariant transition probabilities

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

$$Y_1 = \begin{cases} 1 & \text{w.p. } \frac{1}{2} \\ 2 & \text{w.p. } \frac{1}{2}, \end{cases} \quad P(Y_t = 1 | Y_1, \dots, Y_{t-1}) = \begin{cases} \frac{1}{2} + \frac{(t-1)\lambda}{10} & \text{if } Y_{t-1} = 1 \\ \frac{1}{3} + \frac{(t-1)\lambda}{10} & \text{if } Y_{t-1} = 2 \end{cases} \quad (10)$$

where $0 \leq \lambda \leq 5/3$. For each fixed (t, λ) , $P(Y_t = 1 | Y_{t-1} = 1) = P(Y_t = 1 | Y_{t-1} = 2)$; hence the variables are not AD(0) for any λ . When $\lambda = 0$, however, the transition probabilities are time-invariant; they become more variable over time as λ increases. Ten thousand realizations were simulated from (10) for $\lambda = 0, 0.1, \dots, 1.0$ and $N = 50, 200$, and 1000; and the likelihood ratio test for time-invariant transition probabilities was performed, both for complete data and for data with 25% of the observations deleted as before.

Empirical rejection rates for the tests are listed in Table 2. Results are as expected, with power increasing as either λ , N , or the proportion of non-missing data increases.

7. Examples

7.1. Toenail infection data

Molenberghs and Verbeke ([7], p. 8) report on a longitudinal study comparing two oral treatments (labeled here as A and B) for toenail dermatophyte onychomycosis. Severity of the infection (1 = severe, 2 = not severe) was scheduled to be observed at seven time points: baseline, at which time a three-month treatment period began; months 1, 2, and 3 during treatment; and months 6, 9, and 12 after initial treatment. The aims of the study were to compare the effectiveness of the treatments at reducing infection severity and to see how this evolves over time. For both treatment groups, all but 14 of the $2^7 = 128$ cells are empty. Furthermore, of the 146 and 148 people who received treatments A and B, respectively, only 107 and 117 provided their severity status at all time points (see Web Tables 1 and 2), so some data are missing. We assume that the missingness is ignorable.

The pattern of missingness is not monotone, so the restricted EM algorithm described in Section 3.2.2 was used to fit $AD(p_1, \dots, p_n)$ models of all orders. The fitted models were then compared using the methods of Section 4. For treatment A, AIC selects $AD(0, 1, 1, 1, 1, 1, 2)$; for treatment B it selects $AD(0, 1, 2, 1, 1, 1, 1)$. Furthermore, forward selection and backward elimination strategies for building a constant-order AD model, based on size-0.05 modified likelihood ratio tests at each step, both select an $AD(1)$ model for each group. In light of this, we perform the modified likelihood ratio test for homogeneity across groups assuming first-order antedependence, finding no evidence against it ($P = 0.64$) and concluding therefore that the two treatments do not differ significantly in effectiveness. The logical next step, namely testing for time-invariance of the transition probabilities under the homogeneous $AD(1)$ model, is somewhat problematic due to the unequal spacing between measurements. Time-invariance can be tested straightforwardly over the first 4 measurements (which are equally spaced); for this test $P = 0.026$, suggesting that the lag-one transition probabilities up to the fourth measurement occasion are not time-invariant. Invariance with respect to time *index* rather than actual time, based on all 7 occasions, is rejected even more emphatically ($P < 2.0 \times 10^{-8}$) and an examination of the mles of the transition probabilities (Table 3) indicates why. The estimated transition probabilities for severe infection at time index t , given non-severe infection at the previous time, are small (less than 0.05) and do not change much over time. However, the estimated transition probabilities for severe infection at time index t , given severe infection at the previous time, vary substantially: they decrease over the first four time periods but increase considerably over the last two time periods, even though the last two periods are much longer. Thus, it appears that the “time-specific cure rate,” defined as 1.0 minus the probability of severe infection at time index t given that the infection was severe at the previous time, increases by approximately 8 to 12 percent each month during treatment and for three months thereafter, after which the curative effect of treatment rapidly wears off. Arguably, this phenomenon is more interesting than the behavior of the estimated marginal probabilities of severe infection, which decrease steadily over the first six months (including the first 3-month period after the cessation of treatment) before leveling off for the remaining six months.

It is of interest to compare the homogeneous unstructured $AD(1)$ model we’ve settled on to structured AD models that have been fit previously to these data. Molenberghs and Verbeke ([7], pp. 238–242) fitted three structured $AD(1)$ logistic models, which are listed below. In these models, Y_{ij} represents infection severity (coded as 0 = not severe, 1 = severe) for subject i at time t_{ij} and is assumed to be Bernoulli with parameter μ_{ij} , and T_i is an indicator

variable for Treatment B. The dependence on the previous observation is time-invariant for Model I (hence Molenberghs and Verbeke call it a stationary AR(1) model), but varies over time for Models II and III; in Model II this variation is attributed to the level of the previous outcome and in Model III it is attributed to the difference in spacing of measurements during the treatment and post-treatment periods.

- I. $\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 T_i + \beta_2 t_{ij} + \beta_3 T_i t_{ij} + \alpha_0 y_{i,j-1}$
- II. $\text{logit}(\mu_{ij}) = (\beta_{00} + \beta_{10} T_i + \beta_{20} t_{ij} + \beta_{30} T_i t_{ij}) I(y_{i,j-1} = 0) + (\beta_{01} + \beta_{11} T_i + \beta_{21} t_{ij} + \beta_{31} T_i t_{ij}) I(y_{i,j-1} = 1)$
 $= \beta_{00} + \beta_{10} T_i + \beta_{20} t_{ij} + \beta_{30} T_i t_{ij} + (\alpha_0 + \alpha_1 T_i + \alpha_2 t_{ij} + \alpha_3 T_i t_{ij}) y_{i,j-1}$, where $\alpha_i = \beta_{i1} - \beta_{i0}$
- III. $\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 T_i + \beta_2 t_{ij} + \beta_3 T_i t_{ij} + \alpha_0 y_{i,j-1} + \alpha_1 y_{i,j-1} I(j \in \{1, 2, 3, 4\})$

As it happens, the effects of treatment, time, or their interaction were not statistically significant in any of these models. However, α_0 was significant in all three models, and α_1 in Model III was significant also. Values of AIC for Models I–III and the homogeneous unstructured AD(1) model (based on the likelihood for the available data, conditional on baseline) are, respectively, 519.64, 524.15, 511.26, and 502.44, suggesting that the unstructured model fits best. Moreover, likelihood ratio tests of each of Models I–III against the unstructured model result in emphatic rejection of the structured models ($P < 0.0001$, $P < 0.00001$, and $P < 0.002$ respectively). The relatively superior fit of the unstructured model is likely due, as noted previously, to nonstationarity induced by an accumulating curative effect as treatment continues, which lingers for several months after the cessation of treatment before dissipating. Although Models II and III allow for time-varying dependence on the previous observation, they apparently are not flexible enough to adequately fit this type of nonstationary behavior.

7.2. Alzheimer's disease data

The National Alzheimer's Coordinating Center reposit global clinical dementia ratings (GCDR) data from patients at 29 Alzheimer's Disease Centers throughout the United States. The GCDR is an overall staging of the severity of Alzheimer's disease, computed based on the memory, orientation, judgment and problem solving, community affairs, home and hobbies and personal cares of the patients, and recorded at five levels of impairment (0 = no impairment, 0.5 = questionable impairment, 1.0 = mild impairment, 2.0 = moderate impairment and 3.0 = severe impairment). GCDR was scheduled to be observed annually from 2006 to 2012 for each patient. We are interested in examining how the GCDR for a subject at a given time depends on that subject's GCDRs at previous times. Only 1097 out of 24787 patients had their GCDR observed at all time points, so much of the data is missing. We assume that the missingness is ignorable.

The missingness in these data is not monotone, so the restricted EM algorithm was used to fit $\text{AD}(p_1, \dots, p_n)$ models of all orders, and the fitted models were compared using the methods of Section 4. AIC selects $\text{AD}(0,1,2,2,2,3,2)$, and forward selection and backward elimination methods for constant-order AD models select an $\text{AD}(2)$ model. Although strict stationarity under $\text{AD}(2)$ is rejected ($P < 10^{-5}$), time-invariance of the second-order transition probabilities is not rejected ($P = 0.203$). Accordingly, and with a view toward eventually modeling these data (which are ordinal) in terms of their cumulative logits, we examine the second-order cumulative transition probabilities $P(Y_t = j | Y_{t-2} = y_{t-2}, Y_{t-1} = y_{t-1})$, pooled across time points $t = 3, 4, 5, 6, 7$ (see Table 4); for ease in interpretation and elimination of the arbitrariness of numerical labeling of impairment levels, we also dichotomize previous responses on the basis of whether there is no impairment or uncertain impairment (responses 0 or 0.5) or definitely some impairment (responses 1, 2, 3). The

effect of GCDR in the immediately preceding year on GCDR in a given year is huge and positive, but it is interesting to note (by comparing the probabilities within rows 1 and 3 of the table, and also within rows 2 and 4) that the effect of GCDR two years prior, adjusted for the effect of GCDR in the immediately preceding year, though smaller than the first-order effect, is still quite substantial.

In light of these results, further model-building might reasonably adapt the well-known proportional odds, cumulative logit model (Agresti 2002, sec. 7.2) for use with a second-order autoregression on dichotomized previous responses, i.e.

$$\text{logit}[P(Y_t \leq j | Y_{t-1}, Y_{t-2})] = b_{j0} + b_1 I(Y_{t-1} \leq 0.5) + b_2 I(Y_{t-2} \leq 0.5) \text{ for } j=0, 0.5, 1, 2 \text{ and } t=3, 4, 5, 6, 7. \quad (11)$$

The mles of the coefficients (standard errors) in model (11) can be summarized as follows: $b_{j0} = -6.23(0.09)$, $-3.94(0.08)$, $-0.84(0.04)$ and $0.83(0.04)$ for $j = 0, 0.5, 1$ and 2 respectively; $b_1 = 4.68(0.08)$ and $b_2 = 2.01(0.07)$; and they are all overwhelmingly significant. Table 5 gives the fitted second-order cumulative transition probabilities corresponding to those in Table 4. The agreement between the entries in the two tables is not bad, especially in the first and last rows, and might be improved by adding an interaction effect or other terms to model (11).

8. Discussion

In this article, we have described how to estimate the parameters of unstructured antedependence models for categorical longitudinal data by the method of maximum likelihood. We have also presented likelihood-based methods for determining the order of antedependence by penalized likelihood criteria or formal hypothesis testing; for testing for homogeneity across groups; and for testing for time-invariance of transition probabilities and strict stationarity. Importantly, these methods allow for the possibility of empty cells and missing data. In particular, for data having an arbitrary missingness pattern, mles under unstructured antedependence of any order may be obtained using an efficient restricted EM algorithm. The methods are intended as a prefatory supplement to the common, but in our opinion unwise, practice of proceeding directly to the fitting of structured antedependence models to such data. They focus attention upon transition probabilities, which is often more appropriate than a focus on marginal probabilities. Moreover, using the methods to determine the order of antedependence and identify or refute the presence of other relevant structural features (e.g. time-invariant transition probabilities) may provide useful guidance in selecting (or discarding) a structured antedependence model. For example, our antedependence-based analysis of the toenail infection data, through its focus on the time-specific transition probabilities, led to an interesting interpretation of the curative effects of the treatments over time not noted by any previously published analysis, and indicated that an unstructured first-order AD model is superior to any structured antedependence model that has been fitted by previous authors. Our analysis of the Alzheimer's disease data indicated the importance of using the data to select the order of antedependence rather than arbitrarily adopting a Markov (first-order antedependence) model; the analysis also justified the use of a structured antedependence model in which the transition probabilities are time-invariant.

We assumed throughout that covariates, other than previous responses and indicator variables for treatments or other groups, are unavailable or ignored in the analysis. If categorical covariates are available, they may be handled in the same way as treatments, by allowing for a different transition probability vector at each combination of levels of the covariates. Such models may have many parameters and may therefore require many

observations to yield useful estimates. If, alternatively, the covariates are continuous, then generalized linear models with antedependence structure can be formulated. Research on these latter models is ongoing.

The models that are the subject of this paper, though they allow the order of antedependence to vary over time, are distinct from the variable-length Markov chains (VLMCs) of [18]. In a VLMC, the order of antedependence may depend on the specific realization, but *not* on time. The possibility of parsimoniously modeling categorical longitudinal data by a melded AD-VLMC model that allows for variation in order over time and realizations simultaneously is a topic for future research. Further work is also needed to extend the methods developed herein to multivariate categorical and mixed (some variables categorical, others continuous) data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The data was supported from National Alzheimer's Coordinating Center (NACC grant number, U01 AG016976). This research was supported by the Intramural Research Program of the National Institutes of health, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development. We thank the Center for Information Technology, the National Institutes of Health, for providing access to the high performance computational capabilities of the Biowulf cluster.

References

1. Diggle, PJ.; Heagerty, PJ.; Liang, KY.; Zeger, SL. Analysis of Longitudinal Data. 2. New York: Oxford University Press; 2002.
2. Gabriel KR. Ante-Dependence Analysis of an Ordered Set of Variables. *Annals of Mathematical Statistics*. 1962; 33:201–212.
3. Macchiavelli RE, Arnold SF. Variable-Order Antedependence Models. *Communications in Statistics - Theory and Methods*. 1994; 23:2683–2699.
4. Zimmerman, DL.; Núñez-Antón, V. Antedependence Models for Longitudinal Data. Boca Raton, Florida: CRC Press; 2010.
5. Cox, DR.; Snell, EJ. Analysis of Binary Data. 2. London: Chapman and Hall/CRC Press; 1989.
6. Zeger SL, Qaqish B. Markov Regression Models for Time Series: A Quasi-Likelihood Approach. *Biometrics*. 1988; 44:1019–1031. [PubMed: 3148334]
7. Molenberghs, G.; Verbeke, G. Models for Discrete Longitudinal Data. New York: Springer; 2005.
8. Berchtold A, Raftery A. The Mixture Transition Distribution Model for High-Order Markov Chains and Non-Gaussian Time Series. *Statistical Science*. 2002; 17:328–356.
9. Azzalini A. Logistic Regression for Autocorrelated Data with Application to Repeated Measures. *Biometrika*. 1994; 81:767–775.
10. Heagerty PJ, Zeger SL. Marginalized Multilevel Models and Likelihood Inference. *Statistical Science*. 2000; 15:1–19.
11. Heagerty PJ. Marginalized Transition Models and Likelihood Inference for Longitudinal Categorical Data. *Biometrics*. 2002; 58:342–351. [PubMed: 12071407]
12. Lee K, Daniels MJ. A Class of Markov Models for Longitudinal Ordinal Data. *Biometrics*. 2007; 63:1060–1067. [PubMed: 18078479]
13. Anderson TW, Goodman LA. Statistical Inference About Markov Chains. *Annals of Mathematical Statistics*. 1957; 28:89–110.
14. Lang JB. Multinomial-Poisson Homogeneous Models for Contingency Tables. *Annals of Statistics*. 2004; 32:340–383.

15. Kim DK, Taylor JMG. The Restricted EM Algorithm for Maximum Likelihood Estimation Under Linear Restrictions on the Parameters. *Journal of the American Statistical Association*. 1995; 90:708–716.
16. Schafer, JL. *Analysis of Incomplete Multivariate Data*. Boca Raton, Florida: CRC Press; 1997.
17. Lawley DN. A General Method for Approximating to the Distribution of Likelihood Ratio Criteria. *Biometrika*. 1956; 43:295–303.
18. Bühlmann P, Wyner AJ. Variable Length Markov Chains. *Annals of Statistics*. 1999; 27:480–513.

Table 1

Empirical rejection rates of likelihood ratio test (LRT) and its modification (MLRT) for order of antedependence for data simulated from processes (7)&(8) and (7)&(9). Empirical sizes (rejection rates when $\theta = 0$) more than two estimated standard errors from the nominal size (0.05) are set in bold type.

N	θ	(7)&(8)				(7)&(9)			
		Complete		Missing		Complete		Missing	
50	0	LRT	MLRT	LRT	MLRT	LRT	MLRT	LRT	MLRT
		0.073	0.029	0.066	0.029	0.097	0.038	0.088	0.032
		0.2	0.1	0.047	0.087	0.111	0.044	0.102	0.042
		0.4	0.138	0.068	0.122	0.058	0.14	0.062	0.122
		0.6	0.202	0.113	0.175	0.094	0.183	0.163	0.075
		0.8	0.285	0.181	0.248	0.147	0.24	0.123	0.103
100	1	0.396	0.268	0.337	0.218	0.322	0.185	0.265	0.137
		0.078	0.04	0.076	0.032	0.081	0.046	0.086	0.043
		0.2	0.103	0.057	0.100	0.049	0.06	0.104	0.052
		0.4	0.164	0.099	0.154	0.083	0.146	0.144	0.080
		0.6	0.267	0.183	0.247	0.150	0.221	0.204	0.121
		0.8	0.434	0.328	0.366	0.246	0.334	0.25	0.187
200	1	0.628	0.518	0.512	0.389	0.477	0.385	0.402	0.280
		0.068	0.044	0.072	0.043	0.063	0.049	0.069	0.049
		0.2	0.102	0.073	0.096	0.061	0.091	0.085	0.065
		0.4	0.231	0.184	0.193	0.137	0.186	0.158	0.129
		0.6	0.47	0.406	0.366	0.288	0.351	0.278	0.231
		0.8	0.745	0.687	0.603	0.518	0.565	0.443	0.390
1000	1	0.924	0.895	0.814	0.754	0.772	0.738	0.641	0.588
		0	0.053	0.051	0.051	0.055	0.051	0.057	0.050
		0.2	0.277	0.272	0.213	0.2	0.196	0.151	0.141
		0.4	0.865	0.862	0.731	0.682	0.675	0.544	0.527
		0.6	0.999	0.999	0.988	0.97	0.968	0.906	0.898

<i>N</i>	θ	(7)&(8)				(7)&(9)			
		Complete		Missing		Complete		Missing	
		LRT	MLRT	LRT	MLRT	LRT	MLRT	LRT	MLRT
	0.8	1	1	1	1	0.999	0.999	0.994	0.993
	1	1	1	1	1	1	1	1	1

Table 2

Empirical rejection rates for tests of time-invariant transition probabilities for data simulated from process (10). Empirical sizes more than two estimated standard errors from the nominal size (0.05) are set in bold type.

λ	$N = 50$		$N = 200$		$N = 1000$	
	Complete	Missing	Complete	Missing	Complete	Missing
0	0.058	0.060	0.053	0.055	0.048	0.047
0.1	0.062	0.061	0.061	0.059	0.093	0.086
0.2	0.067	0.063	0.090	0.080	0.260	0.207
0.3	0.079	0.073	0.140	0.117	0.560	0.436
0.4	0.095	0.091	0.220	0.174	0.839	0.708
0.5	0.118	0.106	0.333	0.253	0.969	0.902
0.6	0.149	0.132	0.470	0.363	0.997	0.982
0.7	0.191	0.172	0.620	0.489	1	0.998
0.8	0.238	0.221	0.762	0.618	1	1
0.9	0.299	0.236	0.870	0.746	1	1
1	0.375	0.298	0.941	0.852	1	1

Table 3

Maximum likelihood estimates of marginal and transition probabilities under the homogeneous AD(1) model for the toenail infection data. Time is given in months after initial treatment.

Time	$P(Y_t = 1)$	Transition probabilities	
0	0.369	—	—
1	0.328	$\hat{\pi}_{1 1}^{(1)} = 0.869$	$\hat{\pi}_{1 2}^{(1)} = 0.022$
2	0.280	$\hat{\pi}_{1 +1}^{(1)} = 0.792$	$\hat{\pi}_{1 +2}^{(1)} = 0.042$
3	0.203	$\hat{\pi}_{1 ++1}^{(1)} = 0.699$	$\hat{\pi}_{1 ++2}^{(1)} = 0.015$
6	0.069	$\hat{\pi}_{1 +++1}^{(1)} = 0.361$	$\hat{\pi}_{1 +++2}^{(1)} = 0.009$
9	0.054	$\hat{\pi}_{1 ++++1}^{(1)} = 0.625$	$\hat{\pi}_{1 ++++2}^{(1)} = 0.016$
12	0.058	$\hat{\pi}_{1 +++++1}^{(1)} = 0.722$	$\hat{\pi}_{1 +++++2}^{(1)} = 0.028$

Second-order cumulative transition probabilities $P(Y_t = j | Y_{t-2} = y_{t-2}, Y_{t-1} = y_{t-1})$ collapsed over time points $t = 3, 4, 5, 6$ and 7 and over dichotomized realizations of previous responses.

Table 4

y_{t-2}	y_{t-1}	$j = 0$	$j = 0.5$	$j = 1$	$j = 2$
0.5	0.5	0.616	0.941	0.995	0.999
0.5	1	0.003	0.097	0.717	0.960
1	0.5	0.011	0.459	0.907	0.989
1	1	0.000	0.016	0.330	0.709

Table 5

Fitted [using model (11)] second-order cumulative transition probabilities $\hat{P}(\hat{Y}_t = j | Y_{t-2} = y_{t-2}, Y_{t-1} = y_{t-1})$ collapsed over time points $t = 3, 4, 5, 6$ and 7 and over dichotomized realizations of previous responses.

y_{t-2}	y_{t-1}	$j = 0$	$j = 0.5$	$j = 1$	$j = 2$
0.5	0.5	0.612	0.939	0.997	0.999
0.5	1	0.015	0.126	0.762	0.945
1	0.5	0.175	0.675	0.979	0.996
1	1	0.002	0.019	0.301	0.695