Statistica Sinica **13**(2003), 207-226

A GENERALIZED LINEAR MODEL FOR REPEATED ORDERED CATEGORICAL RESPONSE DATA

K. S. Chan and B. Munoz-Hernandez

University of Iowa and University of Costa Rica

Abstract: We proposed a new approach to model longitudinal data consisting of transitional frequencies classified according to an ordered categorical response variable. Following an approach of Kalbfleisch and Lawless (1985), the responses are assumed to be sampled from an underlying continuous-time finite-state-space Markov chain, with the further assumption that direct transitions are strictly between adjacent states, owing to the ordered categorical nature of the response variable. The model admits a parsimonious parameterization in terms of the transition probability rates (intensity parameters) between adjacent states over an infinitesimal period. It is assumed that after a suitable transformation (link function), the intensity parameters are linear functions of some (possibly time-dependent) covariates. We show that under very mild regularity conditions including a full-rank condition on the "design" matrix, the maximum likelihood (ML) estimators are consistent and asymptotically normal. We also show that, under the same set of regularity conditions and under the null hypothesis of no model misspecification, the likelihood goodness-of-fit test is asymptotically equivalent to the Pearson Chi-square goodness-of-fit test, with the usual limiting Chi-square distribution. We illustrate the new approach with two data sets.

Key words and phrases: Asymptotic normality, continuous-time finite-state-space Markov chain, goodness-of-fit test, maximum likelihood estimation.

1. Introduction

The analysis of repeated measurements of ordered categorical response variable can be broadly classified as:

1. transitional models which model the transition probabilities, and
2. marginal models which model the marginal probabilities.

See Agresti (1989, 1999) for surveys and Kaufmann (1987) for some asymptotic results for marginal models. For transitional modeling, there are a number of approaches. Goodman (1962) considered the use of homogeneous Markov chain of first and higher orders. However, for the case of first order Markov chain this approach requires as many as $K(K-1)$ parameters for each transition probability matrix, where K is the number of categories. Also, there is no natural way to incorporate covariates in this approach.

Göttlein and Pruscha (1992) modeled the cumulative transition probabilities by adapting the method suggested by McCullagh (1980). Their model can include covariates. However their method requires specifying a baseline probability transition matrix which is often specified subjectively.

Kalbfleisch and Lawless (1985) considered analyzing a panel of categorical data by assuming that the data are obtained from sampling a latent continuous-time finite-state-space Markov process; this approach is also referred to as the multi-state Markov model and has found applications in biomedical studies. See, e.g., Kay (1986), Gentleman, Lawless, Lindsey, and Yan (1994), Lee and Kim (1998), and Perez-Ocon, Ruiz-Castro and Gamiz-Perez (2001). For a continuous-time Markov process the transition intensity parameters determine the transition probabilities. See Section 2 for a brief account of the theory. Kalbfleisch and Lawless (1985) pointed out that in some cases the transition intensity matrix may have a simple structure which admits a parsimonious parameterization. For example, Kay (1986) considered the case where direct transitions of the underlying Markov chain must be between adjacent states, or between any state and an absorbing state (death). Similarly, Lee and Kim (1998) considered the case where the states are ordered and the Markov chain proceeds irreversibly and sequentially from a lower state to a higher state with the final state (death) being an absorbing state. The multi-state Markov model allows for the incorporation of covariates via a link function of the intensity parameters. However, as far as we know, this approach has not been adapted to modeling ordered categorical data.

Here, we model ordered categorical panel data using an approach similar to the approach developed by Kalbfleisch and Lawless (1985). However, it is assumed that for the underlying continuous-time finite-state-space Markov chain, any transition over an infinitesimal period must occur between adjacent categories. Hence the intensity matrix is tridiagonal resulting in at most $2(K-1)$ non-zero intensity parameters. It is assumed that after applying a link transformation (e.g., the log-transformation), the vector of intensity parameters is a linear function of some covariates. We propose to estimate the unknown coefficients by the method of Maximum Likelihood (ML).

The tridiagonal form of the intensity matrix has the additional advantage of rendering the implementation of the ML method more stable than a general intensity matrix does (see Section 2). Moreover, the intensities are the rates of transitions between adjacent categories, and hence a model parameterized in terms of the intensity parameters may be interpreted readily. In Section 3, we derive some formulas useful for numerically optimizing the likelihood function. In fact, we consider the likelihood function conditional on the covariates and the initial frequency distribution of the categories. We then discuss the asymptotic distribution of the parameter estimates, as the number of subjects tends

to infinity and the initial relative frequency distribution tends to a fixed positive frequency distribution. We show that, under very mild regularity conditions, the ML estimators of the unknown coefficients are asymptotically jointly normal with the true parameter vector as the mean and the inverse of the Fisher information matrix as the covariance matrix. Moreover, the goodness of fit of the model may be assessed via the likelihood ratio statistic or the Pearson goodness-of-fit statistic which, under the null hypothesis that the model is not misspecified, are shown to be asymptotically equivalent, and asymptotically χ^2 distributed. The degree of freedom of the limiting χ^2-distribution equals the number of free parameters in the saturated model minus the number of parameters in the fitted model. We note that, for the general case where direct transition can be between two arbitrary states, Gentleman, Lawless, Lindsey, and Yan (1994) advocated the use of Pearson goodness-of-fit statistic for assessing the model fit, but they did not give sufficient conditions under which the asymptotic null χ^2-distribution holds. Finally, we illustrate the method with two data sets in Section 4. All proofs are collected in an appendix.

2. A Markov Model

We consider the modeling of longitudinal data consisting of transitional frequencies classified according to an ordered categorical response variable. The responses may be obtained from natural or controlled experiments, which are repeatedly measured at different time points and take values from a set of finitely many ordered categories. Without loss of generality let the categories be denoted by $1, \ldots, K$ and the subjects be sampled over times $0, \ldots, T$. The series of responses, $Y(t)$, $t = 0, \ldots, T$, for each individual in the sample forms a time series. A set of static (time-independent) or dynamic (time-dependent) covariates may also be available for every subject at each sampling time point. Some examples of static covariates are sex, marital status, and the dummy variable indicating which experiment a subject receives. Ordinarily, the static covariates are discrete and their values stratify the population into strata within each of which the population is homogeneous and subject to the same temporal changes. We assume that the subjects are accordingly (post-) stratified into G groups. Because the analysis will be done conditional on the covariates, with no loss of generality, we assume that we have G independent samples from the G populations.

Although the response variable is measured at a finite set of time points, it is often plausible that the responses are sampled from an underlying continuous time process. In other words, we assume that there is an underlying latent process $\{Y(t),\ t \in R,\ t \geq 0\}$ but the process is only observed at $t = 0, \ldots, T$. Henceforth, it is assumed that the responses of any subject from each of the

G populations are obtained from sampling a continuous-time (inhomogeneous) latent Markov process whose transition mechanism depends on a set of dynamic covariates utilized for modeling trend, seasonality and/or intervention.

Quite often, we only have aggregate data in the form of transition frequencies between consecutive time points. Therefore, we need to determine the transition probabilities in terms of the transition parameters. First, we briefly summarize some useful results of continuous-time finite-state space Markov chain; see Cox and Miller (1968) for a systematic account. For simplicity, we assume that all the dynamic covariates are constant between two consecutive sampling time points and that, within this section, $G = 1$; that is, there is only one population. All probabilities and expectations are conditioned on the covariates and the initial frequency distribution of $Y(0)$. Let

$$\begin{aligned} p_{uvst} &= P(\text{the response is in state } v \text{ at time } t \text{ given that it is in state } u \text{ at time } s) \\ &= P(Y(t) = v | Y(s) = u). \end{aligned}$$

The transition probabilities are determined by the q_{uv}'s, the intensity parameters over short time intervals:

$$p_{uvt,t+\Delta t} = q_{uv}\Delta t + o(\Delta t), \qquad u \neq v \tag{1}$$

$$p_{uut,t+\Delta t} = 1 + q_{uu}\Delta t + o(\Delta t), \tag{2}$$

where q_{uv}, $u \neq v \in \{1, \ldots, K\}$, are nonnegative numbers. The larger the intensity parameter q_{uv} is, the greater is the probability of transition from category u to v over a short time period. For simplicity, we assume that the q_{uv}'s are constants (in practice, they depend on the covariates and hence are piecewise constant). It follows from (1) and (2) that $q_{uu} + \sum_{u \neq v} q_{uv} = 0$; consequently, $q_{uu} = -\sum_{u \neq v} q_{uv}$.

Assuming that equations (1) and (2) hold for all t, it can be shown that the derivatives of p_{uvst} (with respect to t) satisfy the Chapman-Kolmogorov equation: $p'_{uvst} = \sum_r p_{urst} q_{rv}$, or in matrix notation

$$\boldsymbol{P}'(s,t) = \boldsymbol{P}(s,t)\boldsymbol{Q}, \tag{3}$$

where the (u, v) entry of $\boldsymbol{P}(s,t)$ and $\boldsymbol{Q}$ are respectively p_{uvst} and q_{uv}. In fact, $\boldsymbol{P}(s,t)$ also satisfies the backward equation:

$$\boldsymbol{P}'(s,t) = \boldsymbol{Q}\boldsymbol{P}(s,t). \tag{4}$$

The initial condition for both the forward and backward equations is $\boldsymbol{P}(s,s) = \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix. It can be shown that

$$\boldsymbol{P}(s,t) = \exp((t-s)\boldsymbol{Q}), \tag{5}$$

where, for any square matrix $\boldsymbol{M} = (m_{ij})$,

$$\exp(\boldsymbol{M}) = \sum_{n=0}^{\infty} \frac{\boldsymbol{M}^n}{n!}. \tag{6}$$

For all $c > 0$, the partial sums on the RHS of (6) converges uniformly for $|\boldsymbol{M}| \leq c$ where $|\boldsymbol{M}|^2 = \sum_{i,j} m_{i,j}^2$ is the squared Euclidean norm of $\boldsymbol{M}$; see Horn and Johnson (1991). Suppose that $\boldsymbol{M}$ is diagonalizable and $\boldsymbol{M} = \boldsymbol{H}\Delta H^{-1}$, where $\boldsymbol{\Delta}$ is a diagonal matrix of the eigenvalues of $\boldsymbol{M}$ and $\boldsymbol{H}$ is the matrix of the corresponding eigenvectors. Then $\exp(\boldsymbol{M}) = \boldsymbol{H}\exp(\boldsymbol{\Delta})\boldsymbol{H}^{-1}$, where $\exp(\boldsymbol{\Delta})$ is the diagonal matrix whose (j,j) entry equals the exponential of the (j,j) entry of $\boldsymbol{\Delta}$, $\forall j$.

The ordering of the categories implies that over an infinitesimal period the continuous-time Markov chain can only jump between adjacent categories, resulting in a tridiagonal intensity transition matrix:

$$\boldsymbol{Q} = \begin{bmatrix} -q_{12} & q_{12} & 0 & \cdots & 0 & 0 & 0 \\ q_{21} & -q_{21} - q_{23} & q_{23} & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & q_{K-1,K-2} & -q_{K-1,K-2} - q_{K-1,K} & q_{K-1,K} \\ 0 & 0 & 0 & \cdots & 0 & q_{K,K-1} & -q_{K,K-1} \end{bmatrix}.$$

Interpretation of the q_{uv}'s is aided by the following two well-known results: Given that $Y(t) = k$, the waiting time for the next transition is exponential with the reciprocal of $-q_{kk}$ as its mean. The second result is that, given a transition out of category k and assuming k is an intermediate category, the odds of the transition being to category $k+1$ is $q_{k,k+1}/q_{k,k-1}$.

Let $\boldsymbol{q} = (q_{12}, q_{23}, \ldots, q_{K-1,K}, q_{21}, q_{32}, \ldots, q_{K,K-1})^T$, where the superscript T denotes taking the transpose; that is, $\boldsymbol{q}$ consists of the first super-diagonal elements of $\boldsymbol{Q}$, followed by the first sub-diagonal elements.

It is assumed that there is a link function h, e.g., the logarithmic transformation, such that

$$h(\boldsymbol{q}) = \boldsymbol{X}(t)\boldsymbol{\theta}, \tag{7}$$

where h is applied entry-wise to $\boldsymbol{q}$, $\boldsymbol{X}(t)$ is a known $2(K-1) \times m$ covariate matrix, and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^T$ is an $m \times 1$ vector of unknown parameters.

Let $\boldsymbol{P}(t, t+1)$ be the transition probability matrix whose (u,v) entry equals $p_{uvt,t+1}$. It follows from (5) that

$$\boldsymbol{P}(t, t+1) = \exp(\boldsymbol{Q}), \quad t = 0, \ldots, T-1. \tag{8}$$

In practice, $\boldsymbol{X}(t)$ is piecewise constant, being constant between two consecutive sampling points. However, (8) still holds with $\boldsymbol{Q}$ replaced by $\boldsymbol{Q}(t, \boldsymbol{\theta})$, its value

at time t. The tridiagonal structure of $\boldsymbol{Q}$ implies that it can be diagonalized (see Exercise 5, p.174 in Horn and Johnson (1985)). Hence $\boldsymbol{P}(t,t+1)$ can be computed easily. Note that when the intensity parameters are small in magnitude, $\boldsymbol{P}(t,t+1) \approx \boldsymbol{I}+\boldsymbol{Q}(t,\boldsymbol{\theta})$. This crude approximation may be used to obtain starting values for ML estimation. As a function of θ, the transition matrix $\boldsymbol{P}(t,t+1)$ enjoys two properties stated below.

Theorem 2.1. *Assume the link function $h:(0,\infty)\to R$ is continuously differentiable and its first derivative never vanishes. Then, for fixed t, the transition matrix $\boldsymbol{P}(t,t+1):\theta\in R^m\to\exp(\boldsymbol{Q}(t,\theta))$ is an element-wise positive and continuously differentiable function of θ.*

We note that the condition of the preceding theorem is satisfied if we take the logarithm function as the link function. To compute the scores, we need the partial derivatives of $\boldsymbol{P}(t,t+1)$ w.r.t. θ_j, the j-th component of θ, which satisfy an equation obtained by differentiating both sides of the backward equation w.r.t. θ_j: for s, t between two consecutive integers,

$$\frac{\partial \boldsymbol{P}'}{\partial\theta_j}(s,t)=\frac{\partial \boldsymbol{Q}}{\partial\theta_j}\boldsymbol{P}(s,t)+\boldsymbol{Q}\frac{\partial \boldsymbol{P}}{\partial\theta_j}(s,t). \tag{9}$$

Noting that $\frac{\partial^2}{\partial t\partial\theta_j}=\frac{\partial^2}{\partial\theta_j\partial t}$ and the initial condition $\frac{\partial \boldsymbol{P}}{\partial\theta_j}(s,s)=0$, the preceding equation can be solved to obtain (see, e.g., Graham (1986, p.10))

$$\frac{\partial \boldsymbol{P}}{\partial\theta_j}(t,t+1)=\boldsymbol{H}\{\boldsymbol{G}\circ(\boldsymbol{H}^{-1}\frac{\partial \boldsymbol{Q}}{\partial\theta_j}(t,\theta)\boldsymbol{H})\}\boldsymbol{H}^{-1} \tag{10}$$

where, for two matrices of identical dimension, $\boldsymbol{A}\circ\boldsymbol{B}=(a_{ij}b_{ij})$ denotes the Hadamard product of $\boldsymbol{A}$ and $\boldsymbol{B}$; λ_i's are the eigenvalues of $\boldsymbol{Q}(t,\theta)$; $\boldsymbol{H}$ is the matrix whose i^{th} column vector is the eigenvector corresponding to λ_i; and $\boldsymbol{G}$ is a matrix whose (i,j) element equals $[\exp(\lambda_j-\lambda_i)-1]/(\lambda_j-\lambda_i)]$ (defined as $\exp(\lambda_i)$ if $\lambda_i=\lambda_j$). See Kalbfleisch and Lawless (1985) and Horn and Johnson (formula (6.6.28), (1991)) for alternative derivations of (10).

3. Maximum Likelihood Estimation

Let n_{guvst} (p_{guvst}) be the number of subjects (probability of a subject) in the gth group and whose response at time t is v, given that at time s the response is u. We abbreviate $n_{guvt,t+1}$ ($p_{guvt,t+1}$) by n_{guvt} (p_{guvt}). Then the conditional log-likelihood function is

$$l(\theta)=\sum_{g=1}^{G}\sum_{t=0}^{T-1}\sum_{u,v=1}^{K}n_{guvt}\log(p_{guvt}(\theta)). \tag{11}$$

The score is

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j} = \sum_{g=1}^{G} \sum_{t=0}^{T-1} \sum_{u,v=1}^{K} \frac{n_{guvt}}{p_{guvt}} \frac{\partial p_{guvt}}{\partial \boldsymbol{\theta}_j}, \qquad j = 1, \ldots, m,$$

the second derivatives are

$$\frac{\partial^2 l(\theta)}{\partial \theta_i \partial \theta_j} = \sum_g \sum_t \sum_{u,v} \frac{n_{guvt}}{p_{guvt}} \frac{\partial^2 p_{guvt}}{\partial \theta_i \partial \theta_j} - \sum_g \sum_t \sum_{u,v} \frac{n_{guvt}}{p_{guvt}^2} \frac{\partial p_{guvt}}{\partial \theta_i} \frac{\partial p_{guvt}}{\partial \theta_j},$$

and the observed Fisher information matrix $I_{\text{obs}}(\theta) = (-\frac{\partial^2 l(\theta)}{\partial \theta_i \partial \theta_j})$. Kalbfleisch and Lawless (1985) approximated the observed Fisher information matrix by its expectation. Here, we adopt a slightly different approximation for the Fisher information matrix by dropping the first term on the right side of the last formula. These approximations are asymptotically equivalent. The ML estimator can be obtained numerically via the method of scoring.

The saturated model treats all the transition probabilities p_{guvt} as parameters. Because $\forall t$, $\forall u$, $\sum_v p_{guvt} = 1$, there are $GTK(K-1)$ free parameters in the saturated model. The non-parametric ML estimator of p_{guvt} is $r_{guvt} = \frac{n_{guvt}}{n_{gut}}$, where $n_{gut} = \sum_v n_{guvt}$. The expected number of subjects whose responses are u at time t equals $E(n_{gut}) = \sum_k n_{gk0} p_{gku0t}(\theta)$. Because all the transition probabilities are positive, so are the p_{gku0t}'s. Let θ^* be the true parameter, n be the number of subjects, and assume that $\phi_{gk0} = \lim_{n\to\infty} n_{gk0}/n$, $\forall k$, g, exist and are positive. Then,

$$\lim_{n\to\infty} E(n_{gut})/n = \sum_k \phi_{gk0} p_{gku0t}(\theta^*). \tag{12}$$

The RHS of (12) will be denoted as $\gamma_{gut} = \gamma_{gut}(\theta^*)$. We prove that $\hat{\theta}_n$, the ML estimator of θ^*, is consistent and asymptotically normal. The goodness of fit of the model may be assessed by the likelihood ratio statistic which is defined as follows:

$$\mathcal{G}^2 = 2(l(\text{saturated model}) - l(\hat{\theta}_n)) = -2 \sum_g \sum_t \sum_{u,v} n_{guvt} \log(p_{guvt}(\hat{\theta}_n)/r_{guvt}).$$

Alternatively, we may use the Pearson goodness-of-fit statistic,

$$X^2 = \sum_g \sum_t \sum_{u,v} (e_{guvt} - n_{guvt})^2 / e_{guvt}, \tag{13}$$

where $e_{guvt} = n_{gut} p_{guvt}(\hat{\theta}_n)$ is the expected count of transitions in the gth group from state u at time t to state v at time $t+1$. We show below that the likelihood ratio statistic and the Pearson goodness-of-fit statistic are asymptotically equivalent under the null hypothesis that the model is not misspecified, and that they

are asymptotically chi-square with $GTK(K-1)-m$ degree of freedom. (Recall that the dimension of θ is m.)

Theorem 3.1. *Let n_{gk0} be the number of subjects in the gth group for which $Y(0)=k$. Write $n_g=\sum_k n_{gk0}$ and $n=\sum_{g=1}^{G} n_g$, assume the following.*

(A1) $\lim_{n\to\infty} n_{gk0}/n=\phi_{gk0}>0$, $\forall k$, g.

(A2) *rank*$(\mathcal{X})=\dim(\theta)=m$, *where $\mathcal{X}$ is the matrix formed by stacking $\boldsymbol{X}_g(t)$ vertically, $g=1,\ldots,G$, $t=0,\ldots,T-1$. The* $\dim(\mathcal{X})$ *is $2(K-1)TG\times m$.*

(A3) *The link function $h:(0,\infty)\to R$ is continuously differentiable and its first derivative never vanishes.*

Then (1) *the model is identifiable and* (2) *the ML estimator $\hat{\theta}_n$ exists and is asymptotically $N(\theta^*,\mathcal{I}^{-1}(\theta^*))$, where $\mathcal{I}(\theta^*)=E(\sum_g\sum_t\sum_{u,v}\frac{n_{guvt}}{p^2_{guvt}}\frac{\partial p_{guvt}}{\partial\theta_i}\frac{\partial p_{guvt}}{\partial\theta_j})$ evaluated under the true model. Finally, under the null hypothesis of no misspecification of the model, the likelihood ratio statistic and the Pearson goodness-of-fit statistic are asymptotically equivalent, and asymptotically χ^2 with $GTK(K-1)-m$ degree of freedom.*

Remarks. If the link function is twice continuously differentiable, then $\mathcal{I}(\theta^*)=E(I_{\text{obs}}(\theta^*))$ is the expected Fisher information matrix. Assumption (**A1**) means that there is a positive limiting fraction of subjects in each group and whose response at $t=0$ equals any fixed but arbitrary category. Consider the associated regression model

$$\boldsymbol{q}_t=\boldsymbol{X}_g(t)\theta+\epsilon_t,\qquad g=1,\ldots,G,\quad t=0,\ldots,T-1.$$

Assumption (**A2**) is equivalent to the condition that the design matrix of the preceding regression model is of full rank; that is, the preceding linear regression model is identifiable. Assumption (**A3**) is the same condition assumed in Theorem 2.1. It can be seen from the proof of the theorem that the conclusions of Theorem 3.1 holds for a general Markov chain model, as outlined in the following. Let $\mathbf{p}(\theta)$ denote the vector consisting of $p_{uvt}(\theta)$, $1\le u,v\le K$, $1\le t\le T-1$ in "lexicographical" order (see the definition below Lemma A.3). Then the conclusions of Theorem 3.1 hold if (A1) and the following two assumptions hold.

(A4) $\mathbf{p}(\theta)$ is a one-to-one function and is continuously differentiable;

(A5) $\partial p(\theta^*)/\partial\theta^T$ is of full-rank, with rank $m=\dim(\theta)$.

4. Examples

It can be verified that conditions (**A1**)−(**A3**) are satisfied for all the models fitted for the two examples discussed below.

Example 1. This example is taken from Agresti (1989). In a double-blind clinical trial, an active hypnotic drug and a placebo were randomly administrated to two independent samples of patients with insomnia. Each individual was asked at the start and at the end of a two-week treatment period the question: "How quickly did you fall asleep after going to bed?". The responses were classified into the one of the four categories: "< 20", "$20 - 30$","$30 - 60$" and "> 60" (in minutes).

Table 1. Time to fall asleep (minutes), by treatment and occasion.

	Initial	Follow-up occasion			
Treatment	occasion	< 20	$20 - 30$	$30 - 60$	> 60
Active drug	< 20	7	4	1	0
	$20 - 30$	11	5	2	2
	$30 - 60$	13	23	3	1
	> 60	9	17	13	8
Placebo	< 20	7	4	2	1
	20-30	14	5	1	0
	30-60	6	9	18	2
	> 60	4	11	14	22

The data are reproduced in Table 1. Although the response variable, "time to fall asleep", could hardly be considered a continuous-time process, the lag time of 14 days between the two measurements may render the continuous time process a useful approximation of the transition mechanism. Table 1 shows that for those subjects whose initial responses belong to the first two categories, there seems to be no difference between the treatment and the control group in terms of the transitions frequencies. The effectiveness of the drug may, however, be argued for subjects initially classified into the last two categories; that is, those with greater sleeping difficulty. It appears that for the treatment group, there are more transitions from the third category to the first two categories, than there are for the control group. Also, there are more transitions from the fourth category to the lower categories in the treatment group than are in the control group. Otherwise, the transition frequencies seem very similar for the two groups. Therefore, we set $\log(q_{32}) = \theta_5 + I[\delta = 1]\theta_7$ and $\log(q_{43}) = \theta_6 + I[\delta = 1]\theta_8$, where δ is 1 for the treatment group and 0 otherwise, and $I[\ ]$ the indicator variable of the enclosed expression. Hence the $\boldsymbol{Q}$ intensity matrix below is modeled on the log-scale by assigning θ_1 to θ_3 to the upper off-diagonal, then θ_4, etc. on the

lower off-diagonal:

$$\boldsymbol{Q} = \begin{bmatrix} -e^{\theta_1} & e^{\theta_1} & 0 & 0 \\ e^{\theta_4} & -e^{\theta_4} - e^{\theta_2} & e^{\theta_2} & 0 \\ 0 & e^{\theta_5+\theta_7 I[\delta=1]} & -e^{\theta_5+\theta_7 I[\delta=1]} - e^{\theta_3} & e^{\theta_3} \\ 0 & 0 & e^{\theta_6+\theta_8 I[\delta=1]} & -e^{\theta_6+\theta_8 I[\delta=1]} \end{bmatrix}. \tag{14}$$

A set of starting values was obtained using the procedure mentioned just above Theorem 2.1. The MLE's are shown in Table 2. Note that both $\hat{\theta}_7$ and $\hat{\theta}_8$ are positive and significant, suggesting that the treatment is effective. Thus, the fitted model suggests that over the period of this experiment the sleeping problem tends to get less severe for patients in the control group that have great sleeping difficulty. This may be due to the placebo effect or some uncontrolled factors such as, for example, a change to a more comfortable weather towards the end of the experiment. Since some observed cells are small or zero, we prefer to assess the goodness of fit of the model via the likelihood ratio statistic $\mathcal{G}^2$. The Pearson goodness-of-fit statistic and the $\mathcal{G}^2$ are distributed asymptotically as chi-square with $G(K-1)K - \dim(\boldsymbol{\theta})$ degrees of freedom. However for moderate sample size, the accuracy of the asymptotic distribution for the Pearson goodness-of-fit statistic is known to suffer when there are some very small counts, see Cochran (1952). Fienberg (1979) suggested that previous simulation results indicate that the $\mathcal{G}^2$ statistic may be more conservative than the Pearson goodness-of-fit statistic. In this and the next example, the Pearson goodness-of-fit statistics are all larger than the $\mathcal{G}^2$ statistics, perhaps due to the existence of several very small expected counts. Further comparison on the performance of these two statistics seems desirable. For these data the $\mathcal{G}^2$ statistic turns out to be 25.81 on $24 - 8 = 16$ d.f., which is insignificant at 5% level. The expected cell counts are shown in Table 3. Comparing these expected values with the data in Table 1, it can be seen that the estimated model provides a reasonably good fit to the two transition matrices.

Table 2. MLE of the parameters and their standard errors for the model (14).

Parameters	MLE	Standard Error
$\hat{\theta}_1$	0.0376	0.47
$\hat{\theta}_2$	-0.472	0.47
$\hat{\theta}_3$	-1.23	0.48
$\hat{\theta}_4$	0.389	0.30
$\hat{\theta}_5$	0.0818	0.24
$\hat{\theta}_6$	0.0185	0.21
$\hat{\theta}_7$	0.981	0.29
$\hat{\theta}_8$	0.612	0.28

Table 3. Time to fall asleep (minutes), by treatment and placebo: expected counts based on the estimated model (14).

	Initial	Follow-up occasion			
Treatment	occasion	< 30	$20-30$	$30-60$	> 60
Treatment	< 20	7.2	4.1	0.7	0.1
	$20-30$	9.7	8.2	1.8	0.2
	$30-60$	15.3	17.4	5.9	1.4
	> 60	10.3	17.1	10.8	8.8
Placebo	< 20	8.2	4.4	1.3	0.1
	$20-30$	8.9	7.1	3.4	0.5
	$30-60$	7.7	10.3	13.4	3.6
	> 60	3.9	7.9	18.1	21.0

Example 2. The second data set is taken from a panel survey of potential voters in Erie County, Ohio, 1940. A group of 445 people responded to six interviews from May to October. They were asked for their vote intention and their answers were classified as either "Republican" (R), "Do not know" (U) or "Democratic" (D); see Anderson and Goodman (1957), Goodman(1962) and Bishop, Fienberg and Holland (1975).

The data are reproduced in Table 4. Although the three categories R, U and D are not, strictly speaking, ordered, it might be argued that the category U is intermediate between R and D. Thus, the tridiagonal structure of the intensity matrix seems plausible. The response variable can be considered as a continuous-time process. As the election got closer, people might gradually firm up their decision. Thus, a simple model including a trend term in the transition rates is first fitted. The parameterization for $\boldsymbol{Q}$ is as follows:

$$\boldsymbol{Q} = \begin{bmatrix} -e^{\theta_5 t+\theta_1} & e^{\theta_5 t+\theta_1} & 0 \\ e^{\theta_6 t+\theta_3} & -e^{\theta_6 t+\theta_3} - e^{\theta_5 t+\theta_2} & e^{\theta_5 t+\theta_2} \\ 0 & e^{\theta_6 t+\theta_4} & -e^{\theta_6 t+\theta_4} \end{bmatrix}. \tag{15}$$

The estimates are in Table 5. This model is called the six-parameter model whereas another model studied below will be referred to as the nine-parameter model. As $\hat{\theta}_5$ and $\hat{\theta}_6$ are (marginally) significant and negative, this confirms that the transition intensities get smaller over time, that people were firming up their decisions as the election got closer.

The likelihood ratio statistic turns out to be 95.00, on $30 - 6 = 24$ d.f. and provides evidence of lack of fit at 5% level of significance ($\chi^2_{24,0.05} = 36.415$).

One plausible explanation for the bad fit of the model is that we have not considered the possible intervention of the Democratic convention held between July and August. This intervention may partially explain the difference between the June-July, the July-August and the August-September transition matrices in the case of the "undecided" people. Anderson (see Goodman (1962)) suggested that the July-August matrix shows changes up to twice the speed of the August-September transition matrix. A second model is now fitted with a parameter accounting for the faster speed on the political decision induced by the Democratic convention for the July-August matrix. A closer examination shows that the May-June and the June-July matrices are similar. So are the August-September and the September-October matrices. Finally the following parameterization for $\boldsymbol{Q}$ is adopted, with the convention that $t = 1$ for May, $t = 2$ for June, etc.

$$\boldsymbol{Q} = \begin{bmatrix} -e^{\theta_1+\theta_9 I[t=3]+\theta_5 I[t\geq 3]} & e^{\theta_1+\theta_9 I[t=3]+\theta_5 I[t\geq 3]} & 0 \\ e^{\theta_3+\theta_9 I[t=3]+\theta_7 I[t\geq 3]} & * & e^{\theta_2+\theta_9 I[t=3]+\theta_6 I[t\geq 3]} \\ 0 & e^{\theta_4+\theta_9 I[t=3]+\theta_8 I[t\geq 3]} & -e^{\theta_4+\theta_9 I[t=3]+\theta_8 I[t\geq 3]} \end{bmatrix}, \quad (16)$$

where "$*$" denotes the negative of the sum of the $(2,1)$ and the $(2,3)$ entries.

Table 4. Vote intention in Erie County, Ohio.

	June					July			
May	R	U	D	Totals	June	R	U	D	Totals
R	125	16	5	146	R	124	16	3	143
U	11	142	18	171	U	22	142	9	173
D	7	15	106	128	D	6	14	109	129
Totals	143	173	129	445	Totals	152	172	121	445

	August					September			
July	R	U	D	Totals	August	R	U	D	Totals
R	146	4	2	152	R	184	7	1	192
U	40	96	36	172	U	10	82	12	104
D	6	4	111	121	D	4	5	140	149
Totals	192	104	149	445	Totals	198	153	94	445

	October			
September	R	U	D	Totals
R	192	5	1	198
U	11	71	12	94
D	2	5	146	153
Totals	205	81	159	445

Table 5. MLE of the parameters and their standard errors, six-parameter model.

Parameters	MLE	Standard Error
$\hat{\theta}_1$	-2.07	0.19
$\hat{\theta}_2$	-1.46	0.19
$\hat{\theta}_3$	-1.39	0.18
$\hat{\theta}_4$	-1.86	0.18
$\hat{\theta}_5$	-0.126	0.058
$\hat{\theta}_6$	-0.101	0.053

This parameterization implies that the $\boldsymbol{Q}$ matrix is the same for the May-June and June-July periods. Also the $\boldsymbol{Q}$ matrix is the same for the August-September and September-October periods. The corresponding transition matrix for July-August is equal to the August-September transition matrix raised to the power e^{θ_9}. This is because $\boldsymbol{Q}(3,\theta) = \exp(\theta_9)\boldsymbol{Q}(4,\theta)$, hence $\boldsymbol{P}(3,4) = \exp(\exp(\theta_9)\boldsymbol{Q}(4,\theta)) = (\exp(\boldsymbol{Q}(4,\theta))^{\exp(\theta_9)} = \boldsymbol{P}(4,5)^{\exp(\theta_9)}$. The MLE's are displayed in Table 6. The parameters θ_1 to θ_4 describe the baseline transition pattern. All their estimates are negative and significant. Based on the magnitude of the estimates, Republicans were less likely than Democrats to switch to the undecided category, whereas in the case of a change of mind, an undecided voter was more likely to switch to the democratic position with an odds $\exp(2.14 - 1.89) = 1.28$. Both $\hat{\theta}_6$ and $\hat{\theta}_7$ are insignificant, suggesting that the undecided have similar transition patterns over time, except possibly with a faster speed over the July-August period. On the other hand, both $\hat{\theta}_5$ and $\hat{\theta}_8$ are significant and negative, suggesting that both the Republicans and the Democrats were less likely to change their mind over the August to October period than over the earlier period of May to July.

Table 6. MLE of the parameters and its standard errors, nine-parameter model.

Parameters	MLE	Standard Error
$\hat{\theta}_1$	-1.820	0.16
$\hat{\theta}_2$	-2.140	0.17
$\hat{\theta}_3$	-1.890	0.15
$\hat{\theta}_4$	-1.650	0.16
$\hat{\theta}_5$	-1.600	0.28
$\hat{\theta}_6$	0.250	0.23
$\hat{\theta}_7$	0.101	0.22
$\hat{\theta}_8$	-1.250	0.26
$\hat{\theta}_9$	0.698	0.15

Anderson's hypothesis implies that $\theta_9 = \log(2) = 0.693\ldots$. Since $\hat{\theta}_9$ is not significantly different from $\log(2)$, Anderson's hypothesis could not be rejected. A test for the goodness of fit of the model is performed and $\mathcal{G}^2$ equals 20.2 on $30 - 9 = 21$ d.f., which is insignificant ($\chi^2_{21,0.05} = 35.2$). This suggests that the fitted model provides a good fit to the data. The nine-parameter model assumes that the Democratic convention affects the three categories equally. This common-effect assumption may be tested by fitting a larger model with the first (second) occurrence of θ_9 on the second row of Q replaced by θ_{10} (θ_{11}), and θ_9 on the third row of Q replaced by θ_{12} so that the Democratic convention may have different effects on the three categories. Twice the increase in the log-likelihood from the 9-parameter model to the 12-parameter model equals 1.844 which is insignificant ($\chi^2_{3,0.05} = 7.815$), suggesting the validity of the common-effect assumption.

5. Conclusion and Acknowledgment

We have developed a new model for analyzing longitudinal data consisting of ordered categorical response data and a set of covariates. The large-sample properties of the ML estimator of the new model have been derived under mild regularity conditions. We illustrated the potential usefulness of the proposed approach via two examples in the preceding section. However, the Markov assumption fundamental to this approach is a strong assumption. It is of interest to develop methods for checking the Markov assumption, and to develop new frameworks for analyzing data in the case that the Markov assumption fails. So far, we have assumed that the covariates are categorical. An interesting problem is to extend the model to include continuous covariates. We thank an associate editor for helpful comments.

Appendix: Proofs of Theorems 2.1 and 3.1

Proof of Theorem 2.1. Since the first derivative of h never vanishes, $h^{-1} : R \to (0, \infty)$ exists and is continuously differentiable. Therefore, $\boldsymbol{q} = h^{-1}(\boldsymbol{X}(t)\theta)$ is positive. (Again, h^{-1} is applied to $\boldsymbol{X}(t)\theta$ element-wise.) Because $\boldsymbol{q}$ is positive, there is positive probability that the underlying process moves from any state to any of its adjacent states over any fixed but arbitrary short time interval. Consequently, the Markov chain moves with positive probability from any state to any other state over a finite time interval, so $\boldsymbol{P}(t, t+1)$ is positive.

First some definitions. A function $g : \Omega \subset R^k \to R$ is *analytic* if (1) it is infinitely differentiable, (2) Ω is an open set and (3) g equals its Taylor series expansion locally, that is, for all $\boldsymbol{x}_0 \in \Omega$, there exists an $r > 0$ such that if the Euclidean norm $|\boldsymbol{x} - \boldsymbol{x}_0| < r$, $g(\boldsymbol{x}) = \sum_\alpha \boldsymbol{D}^\alpha g(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0)^\alpha/\alpha!$, where $\boldsymbol{x} = (x_1, \ldots, x_k)^T$, $\alpha = (\alpha_1, \ldots, \alpha_k)^T$ is a k-tuple of non-negative integers, $\boldsymbol{D}^\alpha g$

is the partial derivative $\partial^{\alpha_1}\cdots\partial^{\alpha_k} g/\partial x_1^{\alpha_1}\cdots\partial x_k^{\alpha_k}$, $\alpha! = \prod_{i=1}^k \alpha_i$ and $\boldsymbol{x}^\alpha = \prod x_i^{\alpha_i}$. A matrix valued function is *analytic* if it is analytic element-wise. Henceforth in this proof, we identify a $K \times K$ matrix with a K^2 by 1 vector formed by stacking its K columns. It follows from (6) that $\exp(\boldsymbol{M})$ is an analytic function of $\boldsymbol{M}$. Since $\boldsymbol{Q}$ is a continuously differentiable function of θ and the composition of two such functions is still continuously differentiable, $\boldsymbol{P}(t, t+1) = \exp(\boldsymbol{Q}(\theta))$ is continuously differentiable in θ. This completes the proof of the theorem.

Proof of Theorem 3.1. With no loss of generality, it is assumed throughout the proof that $G = 1$, that is, there is a single population. Hence, all subscripts indicating the population will be omitted in the notation. For clarity, we sometimes write $\boldsymbol{q}(t,\theta)$, $\boldsymbol{Q}(t,\theta)$ and $\boldsymbol{P}(t, t+1, \theta)$ instead of $\boldsymbol{q}$, $\boldsymbol{Q}$ and $\boldsymbol{P}(t, t+1)$. We write p^*_{uvt}, γ^*_{ut} for $p_{uvt}(\theta^*)$, $\gamma_{ut}(\theta^*)$, etc. (recall that γ^*_{ut} is the expected number of subjects whose responses equal u at time t, under the true model). Assumption (**A2**) implies that for distinct parameters $\theta \neq \theta'$, there exists a $t \in \{0, \ldots, T-1\}$ such that $\boldsymbol{X}(t)\theta \neq \boldsymbol{X}(t)\theta'$, and hence $\boldsymbol{q}(t,\theta) \neq \boldsymbol{q}(t,\theta')$. This is because (**A3**) entails that the link function h is one-to-one. Since the exponential function $\exp(\boldsymbol{M})$ is also a one-to-one function in the matrix argument $\boldsymbol{M}$, we have $\boldsymbol{P}(t, t+1, \theta) \neq \boldsymbol{P}(t, t+1, \theta')$. This demonstrates that the model is identifiable.

The rest of the theorem will be proved through a number of lemmas whose proofs are deferred. Our proof is inspired by Dudley (1976) which provides an elegant exposition on the asymptotic theory of the ML estimation of categorical data models. In particular, the proofs of Lemmas A.1, A.5, and the derivation of the common asymptotic χ^2-distribution for the likelihood and the Pearson goodness-of-fit tests under the null hypothesis of no model misspecification are similar to those of Lemmas 15.1, 15.5 and Theorem 17.4 in Dudley (*op cit.*); hence they are omitted.

Lemma A.1. *$\forall x,\ y > 0$, $\exists w$ between x and y such that $2x\log(x/y) = (x-y)^2/w$.*

Let

$$\tilde{l}(\theta) = \frac{2}{n}\{l(\text{saturated model}) - l(\theta)\} = 2\sum_{t=1}^{T-1}\sum_{u,v}\frac{n_{uvt}}{n}\log\frac{r_{uvt}}{p_{uvt}(\theta)}$$
$$= 2\sum_{t,u}\frac{n_{ut}}{n}\sum_{v} r_{uvt}\log\frac{r_{uvt}}{p_{uvt}}. \tag{17}$$

Lemma A.2. *It holds a.s. that there exists a constant $c > 0$ such that for n sufficiently large, for all $\theta \in R^m$, $\tilde{l}(\theta) \geq c\sum_{t,u}\gamma^*_{ut}\sum_v (r_{uvt} - p_{uvt}(\theta))^2/p^*_{uvt}$.*

Lemma A.3. *The ML estimator $\hat{\theta}_n$ exists and is consistent.*

Let $\mathbf{p}(\theta)$ denote the vector consisting of $p_{uvt}(\theta)$, $1 \le u, v \le K$, $1 \le t \le T-1$, ordered in "lexicographical" order; that is p_{uvt} precedes p_{xys} if and only if either (1) $t < s$ or (2) $t = s$, $u < x$ or (3) $t = s$, $u = x$ and $v < y$. We write $\mathbf{p}(\theta)$ instead of $\mathbf{p}$ to emphasize its dependence on θ. Similarly, $\mathbf{r}$ denotes the vector of all r_{uvt}'s ordered in a manner analogous to that of $\mathbf{p}$. Let $V = \{\mathbf{p}(\theta) : \theta \in R^m\}$. Then, V is a smooth sub-manifold with a single co-ordinate map $\theta \in R^m \to \mathbf{p}(\theta) \in V$. For an introduction to the concept of a manifold, see 16.1−16.6 in Dudley (1976). The derivative of the co-ordinate map at θ is $\boldsymbol{D}\mathbf{p}(\theta) = \partial\mathbf{p}/\partial\theta^T$. Note that $\boldsymbol{D}\mathbf{p}(\theta)$ is a $d \times m$ matrix where $d = TK(K-1)$.

Lemma A.4. *Assume that* (**A2**) *and* (**A3**) *hold, then* $\boldsymbol{D}\mathbf{p}(\theta)$ *is of rank* m *for all* θ*.*

Henceforth, we denote $\boldsymbol{D}\mathbf{p}(\theta^*)$ by $\boldsymbol{D}^*$ or simply by $\boldsymbol{D}$. It follows from Lemma A.4 that V is a m-dimensional manifold. The tangent flat to V at θ^* is the hyper-plane $F = \{\boldsymbol{u}(\theta) = \mathbf{p}(\theta^*) + \boldsymbol{D}(\theta - \theta^*),\ \theta \in R^m\} \subset R^d$. Below, the ambient space R^d is assumed to be endowed with the inner product defined by the following formula where $\boldsymbol{a}$ and $\boldsymbol{b}$ are vectors in R^d with their elements denoted as a_{uvt} and b_{uvt} respectively: $< \boldsymbol{a}, \boldsymbol{b} >_p = \sum_{t,u} \gamma^*_{ut} \sum_v (a_{uvt} - b_{uvt})^2 / p^*_{uvt}$. This inner product induces the vector norm $|\boldsymbol{a}|_p = \sqrt{< \boldsymbol{a}, \boldsymbol{a} >_p}$, for any $\boldsymbol{a} \in R^d$. The Euclidean norm of $\boldsymbol{a}$ is $|\boldsymbol{a}| = \sqrt{\sum_{uvt} a^2_{uvt}}$. Note that the norm $|\cdot|_p$ is equivalent to the Euclidean norm, i.e., there exists two fixed positive constants M_1 and M_2 such that for all $\boldsymbol{a} \in R^d$, $M_1|\boldsymbol{a}| \le |\boldsymbol{a}|_p \le M_2|\boldsymbol{a}|$. It follows from the equivalence of these two norms that $O(|\boldsymbol{a}|) = O(|\boldsymbol{a}|_p)$, and similarly for other asymptotic relations involving O_p, o and o_p. Let $\boldsymbol{f} : \boldsymbol{a} \in R^d \to \boldsymbol{f}(\boldsymbol{a}) \in F$ be the orthogonal projection onto F where $\boldsymbol{f}(\boldsymbol{a}) = \min_{\boldsymbol{b} \in F} |\boldsymbol{a} - \boldsymbol{b}|^2_p$. The projection $\boldsymbol{f}(\boldsymbol{a})$ has an explicit form which is derived below. Let $\boldsymbol{W}$ denote the symmetric matrix for which

$$|\boldsymbol{b}|^2_p = \boldsymbol{b}^T \boldsymbol{W} \boldsymbol{b}, \qquad \forall \boldsymbol{b} \in R^d. \tag{18}$$

Then $\boldsymbol{f}(\boldsymbol{a}) = \min_{\boldsymbol{b} \in F} (\boldsymbol{a} - \boldsymbol{b})^T \boldsymbol{W} (\boldsymbol{a} - \boldsymbol{b})$, which is simply a weighted regression problem whose solution is well-known:

$$\boldsymbol{f}(\boldsymbol{a}) = \mathbf{p}^* + \boldsymbol{D}(\boldsymbol{D}^T \boldsymbol{W} \boldsymbol{D})^{-1} \boldsymbol{D}^T \boldsymbol{W} (\boldsymbol{a} - \mathbf{p}^*). \tag{19}$$

It follows from the definition of the orthogonal projection that $|\boldsymbol{f}(\boldsymbol{a}) - \mathbf{p}^*|^2_p \le |\boldsymbol{a} - \mathbf{p}^*|^2_p$ for all $\boldsymbol{a} \in R^d$.

Lemma A.5. $|\mathbf{p}(\hat{\theta}_n) - \boldsymbol{f}(\mathbf{r})| = o_p(|\mathbf{r} - \mathbf{p}|)$.

We make use of this result to derive the asymptotic distribution of $\hat{\theta}_n$ from that of $\mathbf{r}$. Anderson and Goodman (1957) have shown that for fixed u and t, the random vector $(r_{uvt}, 1 \le v \le K)^T$ has the same asymptotic distribution as

the estimates of the multinomial probabilities $\mathbf{p}^*_{uvt}$'s and with the sample size as if it was $E(n_{ut}) = \gamma^*_{ut}$. Moreover, they showed that the random variables r_{uvt} for two different values of u or two different values of t are asymptotically independent. Anderson and Goodman (*op. cit.*) implies that $\sqrt{n}(\mathbf{r} - \mathbf{p}^*)$ is asymptotically $\mathrm{N}(0, \boldsymbol{C})$ where $\boldsymbol{C}$ is specified below. Recall that $\mathbf{r}$ consists of subvectors $\mathbf{r}_{ut} = (r_{u1t}, \ldots, r_{uKt})^T$, $u = 1, \ldots, K$, $t = 0, \ldots, T-1$, with the indices ut ordered in "lexicographical" order. Similarly defined are the $\mathbf{p}_{ut}$. The matrix $\boldsymbol{C}$ is block diagonal with each block being the asymptotic covariance matrix, of $\sqrt{n}(\mathbf{r}_{ut} - \mathbf{p}^*_{ut})$, which equals $\boldsymbol{C}_{ut} = \gamma^*_{ut}(\mathrm{diag}(\mathbf{p}^*_{ut}) - \mathbf{p}^*_{ut}(\mathbf{p}^*_{ut})^T)$, where $\mathrm{diag}(\mathbf{p}^*_{ut})$ is a diagonal matrix with $\mathbf{p}^*_{ut}$ being the diagonal vector. Clearly, $\boldsymbol{C}_{ut}\mathbf{1} = 0$ where $\mathbf{1}$ is a vector of 1's. Hence $\boldsymbol{C}_{ut}$ is of rank $K-1$ and $\boldsymbol{C}$ is of rank $TK(K-1)$. $\boldsymbol{W}$, defined by (18), can be verified to be a block diagonal matrix whose (u,t) (in "lexicographical" order) block equals $\boldsymbol{W}_{ut} = \gamma^*_{ut}\,\mathrm{diag}(1/\mathbf{p}^*_{ut})$, where division is defined element-wise. It can be verified that

$$\boldsymbol{W}_{ut}\boldsymbol{C}_{ut}\boldsymbol{W}_{ut} = \boldsymbol{W}_{ut} - \gamma^*_{ut}\mathbf{1}\mathbf{1}^T, \tag{20}$$

$$\boldsymbol{C}_{ut}\boldsymbol{W}_{ut}\boldsymbol{C}_{ut} = \boldsymbol{C}_{ut}. \tag{21}$$

However, it follows from the identity $\sum_v p_{uvt}(\theta) \equiv 1$ that for all l, $\sum_v \partial p_{uvt}(\theta)/\partial\theta_l = 0$. Consequently, we get

$$\boldsymbol{D}^T\boldsymbol{W}\boldsymbol{C}\boldsymbol{W} = \boldsymbol{D}^T\boldsymbol{W}, \tag{22}$$

$$\boldsymbol{C}\boldsymbol{W}\boldsymbol{C} = \boldsymbol{C}, \tag{23}$$

with (22) following from (20), and (23) from (21).

It follows from Lemma A.5 and the fact that $\sqrt{n}(\mathbf{r} - \mathbf{p}^*) = O_p(1)$ that $\sqrt{n}(\mathbf{p}(\hat{\theta}_n) - \mathbf{p}^*) = \sqrt{n}(\boldsymbol{f}(\mathbf{r}) - \mathbf{p}^*) + o_p(1)$. Because $|\mathbf{p}(\hat{\theta}_n) - \mathbf{p}^* - \boldsymbol{D}(\hat{\theta}_n - \theta)| = o_p(|\hat{\theta}_n - \theta|)$, we have

$$\sqrt{n}\boldsymbol{D}(\hat{\theta}_n - \theta) = \sqrt{n}(\boldsymbol{f}(\mathbf{r}) - \mathbf{p}^*) + o_p(1). \tag{24}$$

It follows from (19) and Anderson and Goodman (*op. cit.*) that the RHS, and hence the LHS, of (24) is asymptotically $\mathrm{N}(0, \Lambda)$ where $\Lambda = \boldsymbol{D}(\boldsymbol{D}^T\boldsymbol{W}\boldsymbol{D})^{-1}\boldsymbol{D}^T\boldsymbol{W} \times \boldsymbol{C}\boldsymbol{W}\boldsymbol{D}(\boldsymbol{D}^T\boldsymbol{W}\boldsymbol{D})^{-1}\boldsymbol{D}^T$. From (22), Λ becomes $\boldsymbol{D}(\boldsymbol{D}^T\boldsymbol{W}\boldsymbol{D})^{-1}\boldsymbol{D}^T$ whose rank is m. Because $\boldsymbol{D}$ is of full rank, $\sqrt{n}(\hat{\theta}_n - \theta)$ is asymptotically $\mathrm{N}(0, \Sigma)$ for some strictly positive definite matrix Σ. From (24), we have $\boldsymbol{D}\Sigma\boldsymbol{D}^T = \Lambda = \boldsymbol{D}(\boldsymbol{D}^T\boldsymbol{W}\boldsymbol{D})^{-1}\boldsymbol{D}^T$. As $\boldsymbol{D}$ is of full rank, $\Sigma = (\boldsymbol{D}^T\boldsymbol{W}\boldsymbol{D})^{-1}$. It can be verified that $\mathcal{I}(\theta^*)/n = \boldsymbol{D}^T\boldsymbol{W}\boldsymbol{D}$. This completes the proof of the claim on the asymptotic distribution of the ML estimator $\hat{\theta}_n$. This completes the proof of the theorem.

Proof of Lemma A.2. By the Law of Large Numbers, $n_{ut}/n \to \gamma^*_{ut} > 0$ a.s. Hence, for n sufficiently large, $\forall u$, $\forall t$, $n_{ut}/n \geq \gamma^*_{ut}/2$. Combining this result with

Lemma (A.1), we have that for n sufficiently large, $\tilde{l}(\theta) \geq \sum_{t,u} \gamma^*_{ut}/2 \sum_v p^*_{uvt}(r_{uvt} - p_{uvt}(\theta))^2/p^*_{uvt}$. Letting $c = \min\{p^*_{uvt}/2,\ 1 \leq u,\ v \leq K,\ 1 \leq t \leq T-1\}$, we obtain the desired result.

Proof of Lemma A.3. Note that maximizing the log likelihood function is equivalent to minimizing $\tilde{l}(\ )$. Let $\delta > 0$ be an arbitrary but fixed number. Because $\mathbf{p}(\theta)$ is a one-to-one function of θ and is continuous, $\exists \epsilon > 0$ such that $|\theta - \theta^*| > \delta$ implies that there exists u, v, t such that $|p_{uvt}(\theta) - p^*_{uvt}| > \epsilon$. Because $r_{uvt} \to p^*_{uvt}$, it holds a.s. that for n sufficiently large, $|\theta - \theta^*| > \delta$ implies that there exists u, v, t such that $|p_{uvt}(\theta) - r_{uvt}| > \epsilon/2$. Consequently, Lemma A.2 implies that there exists $c_1 > 0$ such that, for n sufficiently large and $|\theta - \theta^*| > \delta$, $\tilde{l}(\theta) \geq c_1\epsilon$. However, $\tilde{l}(\theta^*) \to 0$ a.s. Therefore, the infinmum of $\tilde{l}$ must occur somewhere inside the region $\{\theta, |\theta - \theta^*| \leq \delta\}$, a compact set. So the ML estimator $\hat{\theta}_n$ exists. (In the case that there are several global maxima, $\hat{\theta}_n$ can be taken as any of them.) Moreover, $|\hat{\theta}_n - \theta^*| \leq \delta$ for n sufficiently large a.s. Since $\delta > 0$ is arbitrary, $\hat{\theta}_n$ is consistent.

Proof of Lemma A.4. Let $\boldsymbol{A} = [A_1, \ldots, A_s]$ be a matrix consisting of s column vectors. Then $\mathrm{vec}(\boldsymbol{A})$ denotes the vector consisting of all the column vectors of $\boldsymbol{A}$, specifically, $\mathrm{vec}(\boldsymbol{A}) = (A_1^T, \ldots, A_s^T)^T$. The vec operator enjoys the property that $\mathrm{vec}(\boldsymbol{ABC}) = (\boldsymbol{C}^T \otimes \boldsymbol{A})\mathrm{vec}(\boldsymbol{B})$ where all matrices are assumed to be of compatible dimensions and, for any two matrices $\boldsymbol{R}$ and $\boldsymbol{S}$, $\boldsymbol{R} \otimes \boldsymbol{S}$ denotes their Kronecker product defined by $(r_{ij}\boldsymbol{S})$; see Horn and Johnson (1991). It can be readily verified that for any two matrices, say $\boldsymbol{A}$ and $\boldsymbol{B}$ of the same dimension, $\mathrm{vec}(\boldsymbol{A} \circ \boldsymbol{B}) = \mathrm{vec}(\boldsymbol{A}) \circ \mathrm{vec}(\boldsymbol{B})$. (Recall that $\boldsymbol{A} \circ \boldsymbol{B} = (a_{ij}b_{ij})$ denotes the Hadamard product of $\boldsymbol{A}$ and $\boldsymbol{B}$.) It follows from (10) that

$$\mathrm{vec}\Big(\frac{\partial \boldsymbol{P}}{\partial \theta_j}(t, t+1)\Big) = (\boldsymbol{H}^{-T} \otimes \boldsymbol{H})\Big(\mathrm{vec}(\boldsymbol{G}) \circ \Big\{(\boldsymbol{H}^T \otimes \boldsymbol{H}^{-1})\mathrm{vec}\Big(\frac{\partial \boldsymbol{Q}}{\partial \theta_j}(t, \theta)\Big)\Big\}\Big). \quad (25)$$

Assume, for simplicity, that t is fixed and write $\boldsymbol{Q}(t, \theta)$ and $X_{lj}(t)$ (the (l, j) element of $\boldsymbol{X}(t)$) as $\boldsymbol{Q}$ and X_{lj} respectively. We have

$$\frac{\partial \boldsymbol{Q}}{\partial \theta_j} = \sum_{l=1}^{2K-2} \frac{\partial \boldsymbol{Q}}{\partial q_l}\frac{\partial q_l}{\partial \theta_j} = \sum_{l=1}^{2K-2} \frac{\partial \boldsymbol{Q}}{\partial q_l}\frac{1}{h'(q_l)}X_{lj}, \quad (26)$$

where h' denotes the first derivative of h.

Let $\alpha_1, \ldots, \alpha_m$ be m arbitrary numbers. Then

$$\sum_{j=1}^{m} \alpha_j \mathrm{vec}\Big(\frac{\partial \boldsymbol{P}}{\partial \theta_j}\Big) = (\boldsymbol{H}^{-T} \otimes \boldsymbol{H})\Big(\mathrm{vec}(\boldsymbol{G}) \circ \Big\{(\boldsymbol{H}^T \otimes \boldsymbol{H}^{-1})\mathrm{vec}\Big(\sum_{j=1}^{m} \alpha_j \frac{\partial \boldsymbol{Q}}{\partial \theta_j}\Big)\Big\}\Big).$$

But,

$$\sum_{j=1}^{m} \alpha_j \frac{\partial \boldsymbol{Q}}{\partial \theta_j}(t,\theta) = \sum_{l=1}^{2K-2} \frac{\partial \boldsymbol{Q}}{\partial q_l} \frac{1}{h'(q_l)} \sum_{j=1}^{m} \alpha_j X_{lj}(t).$$

It is readily checked that $\{\mathrm{vec}(\frac{\partial \boldsymbol{Q}}{\partial q_l}),\ l = 1, \ldots, (2K-2)\}$ are independent vectors. It follows from (**A2**) that for any non-zero $\alpha = (\alpha_1, \ldots, \alpha_m)^T$, there exist l and t such that $\sum_{j=1}^{m} \alpha_j X_{lj}(t) \neq 0$, and hence $\sum_{j=1}^{m} \alpha_j \frac{\partial \boldsymbol{Q}}{\partial \theta_j}(t) \neq 0$. (Recall that $h'(q_l)$ is always non-zero.) Because all the elements of $\boldsymbol{G}$ are positive and the matrix $\boldsymbol{H}$ is non-singular, $\sum_{j=1}^{m} \alpha_j \mathrm{vec} \frac{\partial \boldsymbol{P}}{\partial \theta_j}(t, t+1) \neq 0$, demonstrating that $\frac{\partial \boldsymbol{P}}{\partial \theta^T}$ is of rank m.

References

Agresti, A. (1989). A survey of models for repeated ordered categorical response data. *Statist. Medicine* **8**, 1209-1224.

Agresti, A. (1999). Modeling ordered categorical data: recent advances and future challenges. *Statist. Medicine* **18**, 2191-2207.

Anderson, T. and Goodman, L. (1957). Statistical inference about Markov chains. *Ann. Math. Statist.* **28**, 89-110.

Bishop Y., Fienberg, S. and Holland P. W. (1975). *Discrete Multivariate Analysis, Theory and Practice*. The MIT Press, Cambridge.

Cochran W. (1952). The χ^2 test of goodness of fit. *Ann. Math. Statist.* **23**, 315-345.

Cox, D. and Miller, H. (1968). *The Theory of Stochastic Processes*. Chapman and Hall, London.

Dudley, R. M. (1976). *Probabilities and Metrics: Convergence of Laws on Metric Spaces, With a View to Statistical Testing*. Aarhus Universitet, Matematisk Institut, Lecture Notes Series No. 45.

Fienberg, S. (1979). The use of chi-squared statistics for categorical data problems. *J. Roy. Statist. Soc. Ser. B* **41**, 54-64.

Gentleman R. C., Lawless J. F., Lindsey, J. C. and Yan, P. (1994). Multistate Markov-models for analyzing incomplete disease history data with illustrations for HIV disease. *Statist. Medicine* **13**, 805-821.

Goodman, L. (1962). Statistical methods for analyzing processes of change. *Amer. J. Sociology* **68**, 57-78.

Göttlein, A. and Pruscha, H. (1992). Ordinal time series models with application to forest damage data. In *Advanced GLIM and Statistical Modeling* (Edited by Fahrmeir et al.). Springer Verlag Lecture Notes in Statistics, No. 28.

Graham, A. (1986). *Kronecker Products and Matrix Calculus with Applications*. Prentice Hall.

Horn, R. A. and Johnson, C. R. (1985). *Matrix Analysis*. Cambridge University Press, Cambridge.

Horn, R. A. and Johnson, C. R. (1991). *Topics in Matrix Analysis*. Cambridge University Press, Cambridge.

Kay, R. (1986). A Markov model for analyzing cancer markers and disease states in survival studies. *Biometrics* **42**, 855-865.

Kalbfleisch, J. and Lawless, J. (1985). The analysis of panel data under a Markov assumption. *J. Amer. Statist. Assoc.* **80**, 863-871.

Kaufmann, H. (1987). Regression models for nonstationary categorical time series: asymptotic estimation theory. *Ann. Statist.* **15**, 1, 79-98.

Lee E. W. and Kim M. Y. (1998). The analysis of correlated panel data using a continuous-time Markov model. *Biometrics* **54**, 1638-1644.

McCullagh, P. (1980). Regression models for ordinal data. *J. Roy. Statist. Soc. Ser. B* **42**, 109-142.

Perez-Ocon R., Ruiz-Castro J. E. and Gamiz-Perez M. L. (2001). Non-homogeneous Markov models in the analysis of survival after breast cancer. *J. Roy. Statist. Soc. Ser. C* **50**, 111-124.

Department of Statistics and Actuarial Sciences, University of Iowa, Iowa city, Iowa 52242-1409, U.S.A.

E-mail: kchan@stat.uiowa.edu

(Received July 2001; accepted June 2002)