Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uasa20

The Analysis of Panel Data under a Markov Assumption

J. D. Kalbfleisch [a] & J. F. Lawless [a]
[a] Department of Statistics and Actuarial Science , University of Waterloo , Waterloo , Ontario , N2L 3G1 , Canada
Published online: 12 Mar 2012.

To cite this article: J. D. Kalbfleisch & J. F. Lawless (1985) The Analysis of Panel Data under a Markov Assumption, Journal of the American Statistical Association, 80:392, 863-871

To link to this article: http://dx.doi.org/10.1080/01621459.1985.10478195

The Analysis of Panel Data Under a Markov Assumption

J. D. KALBFLEISCH and J. F. LAWLESS*

Methods for the analysis of panel data under a continuous-time Markov model are proposed. We present procedures for obtaining maximum likelihood estimates and associated asymptotic covariance matrices for transition intensity parameters in time homogeneous models, and for other process characteristics such as mean sojourn times and equilibrium distributions. Generalizations to handle covariance analysis and to the fitting of certain nonhomogeneous models are presented, and an example based on a longitudinal study of the smoking habits of school children is discussed. Questions of embeddability and estimation are examined.

KEY WORDS: Markov processes; Maximum likelihood estimation; Regression analysis; Longitudinal data; Embeddability.

1. INTRODUCTION

Continuous-time Markov models have found wide application in the social sciences, especially in the study of data that record life history events (e.g., social mobility studies) for individuals. A recent review of this work is given by Bartholomew (1983); see also, for example, Bartholomew (1982), Singer and Spilerman (1976a,b), and Tuma et al. (1979). In many instances, a full history of the individual processes is not available. In this article we consider "panel data" in which the observations consist of the states occupied by the individuals under study at a sequence of discrete time points; we assume no information to be available about the timing of events between observation times.

Methods for the analysis of panel data under a continuous-time Markov model have been discussed by Bartholomew (1983), Singer and Spilerman (1976a,b), Wasserman (1980), and others. Most of the methodology so far proposed has severe limitations, such as (a) inability to handle observation times that are not equally spaced; (b) inability to produce standard errors, tests, and interval estimates for all model characteristics of interest; and (c) inability to generalize easily to handle covariance analysis. We propose here algorithms for maximum likelihood estimation that overcome these difficulties and that furthermore provide a very efficient way of obtaining maximum likelihood estimates (MLE's).

Section 2 sets notation and reviews a number of results associated with continuous-time Markov processes. Section 3 presents the basic algorithm. Estimates of asymptotic covariance matrices for parameters and estimates of other process characteristics such as equilibrium distributions are also discussed. Some further aspects of maximum likelihood estimation are discussed in Section 4, and Section 5 considers generalizations to covariance analysis and to the fitting of certain nonhomogeneous models. The remainder of the article contains an example based on a longitudinal study of the smoking habits of school children (Section 6), a discussion of the embedding problem for continuous-time Markov processes and its relationships to estimation (Section 7), and some concluding comments.

2. CONTINUOUS-TIME MARKOV PROCESSES

Suppose individuals independently move among k states, denoted by $1, \ldots, k$, according to a continuous-time Markov process. Let $X(t)$ be the state occupied at time t by a randomly chosen individual. For $0 \leq s \leq t$, let $P(s, t)$ be the $k \times k$ transition probability matrix with entries

$$p_{ij}(s, t) = \Pr\{X(t) = j \mid X(s) = i\},$$

for $i, j = 1, \ldots, k$. As is well known (e.g., see Cox and Miller 1965), this process can be specified in terms of the transition intensities,

$$q_{ij}(t) = \lim_{\Delta t \to 0} p_{ij}(t, t + \Delta t)/\Delta t, \qquad i \neq j.$$

For convenience, we also define

$$q_{ii}(t) = -\sum_{j \neq i} q_{ij}(t), \qquad i = 1, \ldots, k,$$

and let $Q(t)$ be the $k \times k$ transition intensity matrix with entries $q_{ij}(t)$.

This article is primarily concerned with time-homogeneous models in which $q_{ij}(t) = q_{ij}$ independent of t. In this case, the process is stationary and we write

$$P(t) = P(s, s + t) = P(0, t)$$

and $Q = (q_{ij})$ to denote the transition intensity matrix. Note that $q_{ij} \geq 0$ for $i \neq j$ and that $\sum_{j=1}^{k} q_{ij} = 0$. It is well known (e.g., see Cox and Miller 1965) that

$$P(t) = e^{Qt} = \sum_{r=0}^{\infty} Q^r t^r / r!.$$

Consider the time-homogeneous model, and suppose that $q_{ij} = q_{ij}(\boldsymbol{\theta})$ depends on b functionally independent parameters $\theta_1, \ldots, \theta_b$, with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_b)$ for each $i, j = 1, \ldots, k$. As a simple example, consider a Markov model for marital stability discussed by Tuma et al. (1979, p. 824). Here, an individual is in state 1 (not married), state 2 (married), or state 3 (has "emigrated" from the study for some reason);

* J. D. Kalbfleisch and J. F. Lawless are Professors, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada. Research was supported in part by grants from the Natural Sciences and Engineering Research Council of Canada. The authors thank K. S. Brown for the data on school children in Section 6 and for discussions concerning these data and the methods proposed here. They also thank W. M. Vollmer and E. Kelly for work on the development of the computer algorithms.

Journal of the American Statistical Association
December 1985, Vol. 80, No. 392, Theory and Methods

state 3 is absorbing. The transition intensity matrix has the form

$$Q = \begin{pmatrix} -\lambda_1-\mu_1 & \lambda_1 & \mu_1 \\ \lambda_2 & -\lambda_2-\mu_2 & \mu_2 \\ 0 & 0 & 0 \end{pmatrix}, \qquad (2.1)$$

and we have $\boldsymbol{\theta} = (\lambda_1, \mu_1, \lambda_2, \mu_2)$, say. The matrix $P(t)$ is a complicated function of these parameters. For example (see Tuma et al. 1979),

$$p_{11}(t) = \frac{(a_1 + \delta_2)e^{\delta_1 t} - (a_1 + \delta_1)e^{\delta_2 t}}{\delta_2 - \delta_1}, \qquad (2.2)$$

where $a_1 = \lambda_1 + \mu_1$, $a_2 = \lambda_2 + \mu_2$, and

$$\delta_1 = -\tfrac{1}{2}(a_1 + a_2) + \tfrac{1}{2}[(a_1 - a_2)^2 + 4\lambda_1\lambda_2]^{1/2}$$
$$\delta_2 = -\tfrac{1}{2}(a_1 + a_2) - \tfrac{1}{2}[(a_1 - a_2)^2 + 4\lambda_1\lambda_2]^{1/2}. \qquad (2.3)$$

We comment further on this model below.

3. MAXIMUM LIKELIHOOD ESTIMATION

Suppose for now that each of a random sample of n individuals is observed at times $t_0, t_1, \ldots, t_m$. (We will discuss the possibility of immigration or emigration from the group of individuals observed later.) If n_{ijl} denotes the number of individuals in state i at t_{l-1} and j at t_l, and if we condition on the distribution of individuals among states at t_0, then the likelihood function for $\boldsymbol{\theta}$ is

$$L(\boldsymbol{\theta}) = \prod_{l=1}^{m} \left\{ \prod_{i,j=1}^{k} p_{ij}(t_{l-1}, t_l)^{n_{ijl}} \right\}. \qquad (3.1)$$

In the time-homogeneous case with $w_l = t_l - t_{l-1}$, $l = 1, \ldots, m$, (3.1) gives the log-likelihood,

$$\log L(\boldsymbol{\theta}) = \sum_{l=1}^{m} \sum_{i,j=1}^{k} n_{ijl} \log p_{ij}(w_l). \qquad (3.2)$$

Throughout, we suppress the dependence of $p_{ij}(w_l; \boldsymbol{\theta})$ on $\boldsymbol{\theta}$.

The MLE $\hat{\boldsymbol{\theta}}$ or, equivalently, the MLE $\hat{Q} = Q(\hat{\boldsymbol{\theta}})$ of Q, is obtained by maximizing (3.2). One approach (see Wasserman 1980) utilizes a numerical algorithm that requires no derivatives of $\log L(\boldsymbol{\theta})$. There are various algorithms that could be used (e.g., see Chambers 1977, chap. 6). We propose here a more efficient quasi-Newton procedure that uses first derivatives of $\log L(\boldsymbol{\theta})$. This leads to faster convergence and an estimate of the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$. Our approach is made feasible by the provision of an efficient algorithm for the computation of $P(t; \boldsymbol{\theta})$ and its derivative with respect to $\boldsymbol{\theta}$. This leads to simple computation of $\log L(\boldsymbol{\theta})$ and its derivatives.

To compute $P(t; \boldsymbol{\theta}) = \exp\{Q(\boldsymbol{\theta})t\}$ for a given $\boldsymbol{\theta}$, we use a canonical decomposition: if, for the given $\boldsymbol{\theta}$, $Q(\boldsymbol{\theta})$ has distinct eigenvalues $d_1, \ldots, d_k$ and A is the $k \times k$ matrix whose jth column is a right eigenvector corresponding to d_j, then $Q = ADA^{-1}$, where $D = \operatorname{diag}(d_1, \ldots, d_k)$. Then

$$P(t) = A \operatorname{diag}(e^{d_1 t}, \ldots, e^{d_k t})A^{-1}, \qquad (3.3)$$

where the dependence of Q, $P(t)$, A and the d_j's on $\boldsymbol{\theta}$ is suppressed for notational convenience. When Q has repeated eigenvalues, an analogous decomposition of Q to Jordan canonical form is possible (see Cox and Miller 1965). This is rarely necessary, since for most models of interest, $Q(\boldsymbol{\theta})$ has distinct eigenvalues for almost all $\boldsymbol{\theta}$.

Derivatives can be computed in a similar way. In particular, the matrix with entries $\partial p_{ij}(t; \boldsymbol{\theta})/\partial\theta_u$ can be obtained as

$$\partial P(t)/\partial\theta_u = AV_uA^{-1}, \qquad u = 1, \ldots, b, \qquad (3.4)$$

where V_u is a $k \times k$ matrix with (i, j) entry

$$g_{ij}^{(u)}(e^{d_i t} - e^{d_j t})/(d_i - d_j), \qquad i \neq j,$$
$$g_{ii}^{(u)} t e^{d_i t}, \qquad i = j,$$

and $g_{ij}^{(u)}$ is the (i, j) entry in $G^{(u)} = A^{-1}(\partial Q/\partial\theta_u)A$. A derivation of this result, which also appears in Jennrich and Bright (1976), is given in Appendix A. The derivatives $\partial Q/\partial\theta_u$ are usually very simple, and calculation of (3.4) is easy once A and D are obtained. It is generally the case that $P(t)$ is a complicated function of $\boldsymbol{\theta}$; the above formulas, however, allow us to avoid use of any explicit representation of the elements $p_{ij}(t)$ as functions of $\boldsymbol{\theta}$.

3.1 The Scoring Procedure

The quasi-Newton (or scoring) procedure described below is an efficient and simple method for obtaining the MLE of $\boldsymbol{\theta}$. From (3.2) we find

$$S_u(\boldsymbol{\theta}) = \frac{\partial \log L}{\partial\theta_u} = \sum_{l=1}^{m} \sum_{i,j=1}^{k} n_{ijl} \frac{\partial p_{ij}(w_l)/\partial\theta_u}{p_{ij}(w_l)}, \qquad u = 1, \ldots, b, \qquad (3.5)$$

$$\frac{\partial^2 \log L}{\partial\theta_u\partial\theta_v} = \sum_{l=1}^{m} \sum_{i,j=1}^{k} n_{ijl} \times \left\{ \frac{\partial^2 p_{ij}(w_l)/\partial\theta_u\partial\theta_v}{p_{ij}(w_l)} - \frac{\partial p_{ij}(w_l)/\partial\theta_u \partial p_{ij}(w_l)/\partial\theta_v}{p_{ij}^2(w_l)} \right\}.$$

Direct use of a Newton–Raphson algorithm would require the evaluation of both first and second derivatives. We use instead the scoring device in which the second derivatives are replaced by estimates of their expectations. This leads to an algorithm in which only first derivatives are required.

Let $N_i(t_{l-1}) = \sum_{j=1}^{k} n_{ijl}$ represent the number of individuals in state i at time t_{l-1}. By first taking the expectation of n_{ijl} conditional on $N_i(t_{l-1})$ and then using the fact that $\sum_{j=1}^{k} \partial^2 p_{ij}(w_l)/\partial\theta_u\partial\theta_v = 0$, we find that

$$E\left(-\frac{\partial^2 \log L}{\partial\theta_u\partial\theta_v}\right) = \sum_{l=1}^{m} \sum_{i,j=1}^{k} \frac{E\{N_i(t_{l-1})\}}{p_{ij}(w_l)} \frac{\partial p_{ij}(w_l)}{\partial\theta_u} \frac{\partial p_{ij}(w_l)}{\partial\theta_v},$$

which can be estimated by

$$M_{uv}(\boldsymbol{\theta}) = \sum_{l=1}^{m} \sum_{i,j=1}^{k} \frac{N_i(t_{l-1})}{p_{ij}(w_l)} \frac{\partial p_{ij}(w_l)}{\partial\theta_u} \frac{\partial p_{ij}(w_l)}{\partial\theta_v}. \qquad (3.6)$$

Computation of (3.5) and (3.6) for any given θ is facilitated by the results (3.3) and (3.4).

The algorithm is now simply described. Let $\boldsymbol{\theta}_0$ be an initial estimate of $\boldsymbol{\theta}$, $S(\boldsymbol{\theta})$ be the $b \times 1$ vector $(S_u(\boldsymbol{\theta}))$, and $M(\boldsymbol{\theta})$ the $b \times b$ matrix $(M_{uv}(\boldsymbol{\theta}))$. Then an updated estimate is obtained as

$$\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0 + M(\boldsymbol{\theta}_0)^{-1}S(\boldsymbol{\theta}_0),$$

where it is assumed that $M(\boldsymbol{\theta}_0)$ is nonsingular. The process is now repeated with $\boldsymbol{\theta}_1$ replacing $\boldsymbol{\theta}_0$. With a good initial estimate $\boldsymbol{\theta}_0$, the algorithm produces $\hat{\boldsymbol{\theta}}$ upon convergence, and

$M(\hat{\boldsymbol{\theta}})^{-1}$ is an estimate of the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$. If the true value of $\boldsymbol{\theta}$ is an interior point of the parameter space, $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ has a limiting multivariate normal distribution as the number of individuals (n) under observation approaches infinity.

3.2 Illustration

Consider the model (2.1) with $\boldsymbol{\theta} = (\lambda_1, \mu_1, \lambda_2, \mu_2)$. To implement the algorithm, we determine the eigenvalues and eigenvectors of $Q(\boldsymbol{\theta}_0)$, where $\boldsymbol{\theta}_0$ is the trial value. In this case, the eigenvalues can be obtained algebraically; they are 0, δ_1, and δ_2, where δ_1 and δ_2 are defined in (2.3). The eigenvectors are more complicated functions of $\boldsymbol{\theta}$. It is important to note, however, that no use is made of these algebraic expressions; for the given $\boldsymbol{\theta}_0$ the numerical values of the eigenvectors and eigenvalues can be easily obtained with standard software.

Once the canonical decomposition of $Q(\boldsymbol{\theta}_0)$ is obtained, the matrices $P(w_l; \boldsymbol{\theta}_0)$ are obtained from (3.3) and the derivatives from (3.4). The matrices $G^{(u)}$ in (3.4) are simply computed; for example,

$$G^{(1)} = A^{-1}\begin{pmatrix} -1 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} A.$$

The entire procedure is easily programmed and executes quickly on a computer. A listing of a FORTRAN subroutine that implements the general methodology of this article is available from the authors. An example is given in Section 6.

As is remarked further below, there is an advantage here to parameterize in terms of $\boldsymbol{\theta} = (\log \lambda_1, \log \mu_1, \log \lambda_2, \log \mu_2)$, since the parameter space for $\boldsymbol{\theta}$ is then unrestricted. This guarantees that the result of an iteration will not fall outside the parameter space. Estimates that lie on the boundary of the parameter space do cause difficulty and are discussed in Section 4.

3.3 Estimates of Derived Quantities

Often, specific characteristics of the model are of interest, such as $-q_{ii}(\boldsymbol{\theta})^{-1}$, the mean sojourn time in state i. In general, such quantities are estimated by replacing $\boldsymbol{\theta}$ with $\hat{\boldsymbol{\theta}}$, and the asymptotic variance is estimated from the multivariate delta theorem (see Rao 1973, p. 388). For example, the mean sojourn time in state i is estimated by $-q_{ii}(\hat{\boldsymbol{\theta}})^{-1}$ and its estimated asymptotic variance is $q_{ii}(\hat{\boldsymbol{\theta}})^{-4}$ times

$$\sum_{u=1}^{b}\sum_{v=1}^{b} \frac{\partial q_{ii}(\boldsymbol{\theta})}{\partial \theta_u}\frac{\partial q_{ii}(\boldsymbol{\theta})}{\partial \theta_v} M^{uv}(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}},$$

where $M^{uv}(\boldsymbol{\theta})$ is the u, v element of $M(\boldsymbol{\theta})^{-1}$. In like manner, for example, the transition probabilities $p_{ij}(t; \boldsymbol{\theta})$ for fixed i, j, and t can be estimated. The required derivatives are computed using (3.3) and (3.4).

One important characteristic of ergodic models is the equilibrium probability vector $\boldsymbol{\pi}' = (\pi_1, \ldots, \pi_k)$, specified as the unique solution to the equation $Q'\boldsymbol{\pi} = \mathbf{0}$ that satisfies $\sum_{i=1}^{k} \pi_i = 1$. In order to determine an asymptotic covariance matrix for $\hat{\boldsymbol{\pi}}$ via the delta theorem, we require derivatives $\partial \pi_i / \partial \theta_j$. In Appendix B we prove that the following procedure yields an asymptotic covariance matrix for $\hat{\pi}$.

Let $Q_1(\boldsymbol{\theta})$ be the $k \times (k-1)$ matrix obtained by deleting the last column of Q, and define the $(k-1) \times (k-1)$ matrix $B(\boldsymbol{\theta})$ with (i, j) element

$$B_{ij}(\boldsymbol{\theta}) = q_{ji}(\boldsymbol{\theta}) - q_{ki}(\boldsymbol{\theta}), \qquad i, j = 1, \ldots, k-1.$$

In addition, let $C(\boldsymbol{\theta})$ be the $(k-1) \times b$ matrix whose j'th column is given by

$$(\partial Q_1 / \partial \theta_j)'\boldsymbol{\pi}. \tag{3.7}$$

The derivatives of $\pi_1, \ldots, \pi_{k-1}$ with respect to $\theta_1, \ldots, \theta_b$ are contained in the matrix W with (i, j) element $\partial \pi_i / \partial \theta_j$, where

$$W(\boldsymbol{\theta}) = -B(\boldsymbol{\theta})^{-1}C(\boldsymbol{\theta}). \tag{3.8}$$

An estimate of the asymptotic covariance matrix of $\hat{\boldsymbol{\pi}}_1' = (\hat{\pi}_1, \ldots, \hat{\pi}_{k-1})$ is

$$V_{\pi} = W(\hat{\boldsymbol{\theta}})VW(\hat{\boldsymbol{\theta}})', \tag{3.9}$$

where $V = M(\hat{\boldsymbol{\theta}})^{-1}$ is the estimated covariance matrix for $\hat{\boldsymbol{\theta}}$. It will be noticed that the matrices $B(\hat{\boldsymbol{\theta}})$ and $C(\hat{\boldsymbol{\theta}})$ are easily calculated from quantities already computed when obtaining $\hat{\boldsymbol{\theta}}$.

3.4 Some Simple Extensions

In many studies, some of the individuals are observed only over part of the observation period. If individuals who enter or leave the observation period are like those who stay in the study in all relevant respects (the same transition probabilities apply to both groups), then the methods apply without change. Note, however, that $N_i(t_{l-1})$ of (3.5) is the number of individuals in state i at t_{l-1} for whom the state occupied at t_l is known.

It is similarly unnecessary that all individuals be observed over the same set of time points. The amount of computation, however, increases linearly with the number of distinct time intervals w_l in the sample.

These methods can be extended in a simple way to fit certain nonhomogeneous Markov models. Suppose, for example, that $\{X(t)\}$ is a Markov process with time-dependent intensity matrix $Q(t) = Q_0 \cdot g(t; \lambda)$, where Q_0 is a fixed intensity matrix with unknown entries (q_{ij}), and $g(t; \lambda)$ is a known function of time up to an unspecified parameter λ. In this case, $g(t; \lambda)$ defines, for given λ, an operational time. For given λ, let $s = \int_0^t g(u; \lambda)\, du$ and define $Y(s) = X(t)$. The process $\{Y(s): 0 < s < \infty\}$ is then a homogeneous Markov process with intensity matrix Q_0. Thus for any given λ we replace t_l with $s_l = \int_0^{t_l} g(u; \lambda)\, du$ and w_l with $w_l^* = s_l - s_{l-1}$. The parameters of Q_0 are estimated by the methods outlined above. In addition, the maximized log-likelihood can be obtained for that λ. By varying λ, this additional parameter can be estimated by observing the effect on the maximized log-likelihood. Our approach to variance estimation and confidence intervals is to consider λ as fixed (usually at the value $\hat{\lambda}$, which maximizes the maximized log-likelihood) and then to use the estimated covariance matrix $M(\hat{\boldsymbol{\theta}})^{-1}$ for the corresponding estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$. The associate editor has pointed out the similarity between the interpretation and treatment of the λ parameter here and that of the "conditional" approach to the estimation of the transformation parameter λ in the transformation family of Box and Cox (1964); see Hinkley and Runger (1984) for a review and discussion of this topic.

A second type of time-dependent model can also be treated. In this, the transition matrix $Q(t)$ is allowed to change at the observation times but is constant between observation times. Thus, for example, we might consider

$$\begin{aligned} Q(t) &= Q_1, \qquad 0 \le t < t_{m_1} \\ &= Q_2, \qquad t_{m_1} \le t < t_m, \end{aligned}$$

for some specified m_1 less than m. Estimation of Q_1 and Q_2 proceeds by treating the first m_1 and the second $m_2 - m_1$ time intervals separately.

4. SOME FURTHER ASPECTS OF ESTIMATION

In this section we discuss some points associated with implementing the algorithm for maximum likelihood estimation. Many of these comments have to do with the shape of the likelihood function and the information available about the parameters.

1. An initial estimate $\boldsymbol{\theta}_0$ of $\boldsymbol{\theta}$ can usually be obtained in an ad hoc way by examining the transition counts n_{ijl}; an example is given in Section 6. An alternative approach would involve a preliminary tabulation of the likelihood surface.

2. There is an advantage to parameterizing the model by writing $q_{ij} = \exp(a_{ij})$, $i \neq j$. This is because the parameters a_{ij} can take any real value whereas $q_{ij} \ge 0$. This reparameterization avoids problems that can arise when an iteration results in parameter vectors outside the parameter space. Note also that it is possible to have $q_{ij} = 0$. When this happens, successive iterates of a_{ij} will typically become large and negative. In this situation we have found it useful to set the corresponding $q_{ij} = 0$ and fit the model, against using the a_{ij}'s, with one less parameter. The resulting estimate is then compared with one in which q_{ij} is taken as a small positive value.

3. When the times w_l between successive observations are large, it is clear on intuitive grounds that not all parameters will be well estimated. The result can be a likelihood surface for the q_{ij}'s or a_{ij}'s that has ridges defined by certain parameters that are imprecisely estimated. For example, if the w_l's are very large and the process is ergodic, then $p_{ij}(w_l) \doteq \pi_j$, where $\boldsymbol{\pi}' = (\pi_1, \ldots, \pi_k)$ is the vector of equilibrium probabilities. In this case, a large number of individuals under study will allow precise estimation of $\boldsymbol{\pi}$, but individual q_{ij}'s will be imprecisely estimated.

4. If $w_l = w$ for all l, it may be possible to determine $\hat{\boldsymbol{\theta}}$ in a relatively simple way. The empirical transition matrix $\tilde{P}(w)$ with $\tilde{p}_{ij.}(w) = n_{ij.}/n_{i..}$, where $n_{ij.} = \sum_{l=1}^{m} n_{ijl}$ and $n_{i..} = \sum_{j=1}^{k} n_{ij.}$, provides an estimate of $P(w)$. If the equation $\tilde{P}(w) = \exp(Qw)$ admits a solution $\hat{Q} = Q(\hat{\boldsymbol{\theta}})$, then $\hat{\boldsymbol{\theta}}$ is a MLE of $\boldsymbol{\theta}$. The conditions under which this approach can be used are very restrictive. There is a close relationship with the so-called embeddability problem (see Singer and Spilerman 1976a), which is discussed in Section 7.

An Illustrative Example. Consider an ergodic two-state process with

$$Q = \begin{pmatrix} -\alpha & \alpha \\ \beta & -\beta \end{pmatrix},$$

where $\alpha, \beta > 0$. It is easily shown that

$$\begin{aligned} P(w) &= (p_{ij}(w)) \\ &= \begin{pmatrix} 1 - \pi + \pi e^{-\psi w} & \pi(1 - e^{-\psi w}) \\ (1 - \pi)(1 - e^{-\psi w}) & \pi + (1 - \pi)e^{-\psi w} \end{pmatrix}, \end{aligned} \tag{4.1}$$

where $\pi = \alpha/(\alpha + \beta)$ and $\psi = \alpha + \beta$. Suppose that $w_l = w$ $(l = 1, \ldots, m)$ so that the likelihood (3.1) can be written as

$$L = [1 - p_{12}]^{n_{11}} p_{12}^{n_{12}} p_{21}^{n_{21}} [1 - p_{21}]^{n_{22}}, \tag{4.2}$$

where $p_{ij} = p_{ij}(w)$. For fixed w, the parameter space $\{(\pi, \psi): 0 < \pi < 1, \psi > 0\}$ is mapped one to one by (4.1) onto $A = \{(p_{12}, p_{21}): 0 < p_{12} < 1, 0 < p_{21} < 1, p_{12} + p_{21} < 1\}$. We can maximize L for $(p_{12}, p_{21}) \in A$ and then use the relationship (4.1) to obtain $\hat{\pi}, \hat{\psi}$.

Let $\tilde{p}_{12} = n_{12.}/n_{1..}$, $\tilde{p}_{21} = n_{21.}/n_{2..}$, which is the unconstrained maximum of (4.2). If $(\tilde{p}_{12}, \tilde{p}_{21}) \in A$, (4.1) can be used to show that

$$\hat{\pi} = \frac{\tilde{p}_{12}}{\tilde{p}_{12} + \tilde{p}_{21}}, \qquad \hat{\psi} = -\frac{1}{w} \log[1 - \tilde{p}_{12} - \tilde{p}_{21}].$$

If, however, $\tilde{p}_{12} + \tilde{p}_{21} \ge 1$, the supremum of L in A lies on the boundary $p_{12} + p_{21} = 1$ and we find

$$\hat{p}_{12} = 1 - \hat{p}_{21} = (n_{11.} + n_{21.})/n_{...},$$

where $n_{...} = \sum_{i,j} n_{ij.}$. Thus $\pi = (n_{11.} + n_{21.})/n_{...}$ and $\hat{\psi} = \infty$. The parameter π (which incidentally determines the equilibrium distribution) has an admissible MLE, but the likelihood is, for fixed π, bounded but increasing as $\psi \to \infty$.

Note that $p_{12}(w) + p_{21}(w) = 1 - e^{-\psi w}$. If ψw is large, then $p_{12} + p_{21}$ will be close to one, and it is clear that observations for which $\tilde{p}_{12} + \tilde{p}_{21} \ge 1$ will arise much more frequently than when ψw is small. Finally, we remark that for this two-state case there is a close connection with the embeddability problem in that there is no finite MLE of ψ precisely when the empirical transition matrix $\tilde{P}(w)$ is nonembeddable, but this does not appear to be a general phenomenon. We discuss this further in Section 7.

5. THE INCORPORATION OF COVARIATES

In many applications, there are measured covariates on each individual under study, and interest centers on the relationship between these covariates and the intensities q_{ij} in the Markov model. One advantage of the methods in Section 3 is that they generalize in a straightforward way to allow for the regression modeling of Q, though with many distinct covariate values in the sample, computations may be too extensive to be easily implemented directly. Implementation may require that the covariates be grouped.

Suppose that each individual has an associated vector of s covariates, $\mathbf{z}' = (z_1, z_2, \ldots, z_s)$, where $z_1 = 1$. For given $\mathbf{z}$, we suppose that the process is homogeneous Markov with transition intensity matrix

$$Q(\mathbf{z}) = (q_{ij}(\mathbf{z})),$$

where

$$q_{ij}(\mathbf{z}) = \exp(\mathbf{z}'\boldsymbol{\beta}_{ij}), \qquad i \neq j, \tag{5.1}$$

and $q_{ii}(\mathbf{z}) = -\sum_{j \neq i} q_{ij}(\mathbf{z})$. In (5.1), $\boldsymbol{\beta}_{ij} = (\beta_{1ij}, \ldots, \beta_{sij})'$ is a vector of s regression parameters relating the instantaneous rate of transitions from state i to state j to the covariates $\mathbf{z}$. In most applications, some of the regression parameters are specified (usually taken to be zero) and only a subset is to be estimated; the relationship between components of $\mathbf{z}$ and a subset of the transition rates would usually be of interest.

The particular parametric form in (5.1), a log-linear model for the Markov rates $q_{ij}(\mathbf{z})$, is chosen primarily for analytical convenience; other parameterizations may be more appropriate in particular applications. This model has, however, the attractive feature of yielding nonnegative transition intensities for any $\mathbf{z}$ and $\boldsymbol{\beta}_{ij}$'s, and it has been suggested by several authors. Tuma and Robins (1980, pp. 1034–1035) provided an example.

The algorithm of Section 3 requires a separate canonical decomposition of $Q(\mathbf{z})$ for each of the r distinct covariate vectors $\mathbf{z}$ in the sample. Let these be denoted by $\mathbf{z}_h = (z_{1h}, \ldots, z_{sh})$ with $z_{1h} = 1$, and let

$$Q_h = Q(\mathbf{z}_h) = (q_{ij}(\mathbf{z}_h)), \qquad h = 1, \ldots, r,$$

as defined in (5.1). Let $n_{ijl}^{(h)}$ be the number of individuals with covariate values $\mathbf{z}_h$ that are in state i at t_{l-1} and state j at t_l. The likelihood is then a product of terms like (3.1), where the hth term arises from data collected on a homogeneous model with intensity matrix Q_h. Thus the log-likelihood is

$$\log L(\boldsymbol{\theta}) = \sum_{h=1}^{r} \sum_{l=1}^{m} \sum_{i,j=1}^{k} n_{ijl}^{(h)} \log p_{ij}(w_l; \mathbf{z}_h),$$

where

$$P_h(t) = \exp(Q_h t) = (p_{ij}(t; \mathbf{z}_h)).$$

The parameter $\boldsymbol{\theta}$ is being used to indicate the vector of parameters in $\boldsymbol{\beta}_{ij}$ $(i \neq j)$ that are to be estimated. The maximum likelihood equations involve the sum of r terms, one for each $\mathbf{z}_h$. Thus the score vector is

$$S(\boldsymbol{\theta}) = \sum_{h=1}^{r} S^{(h)}(\boldsymbol{\theta}),$$

where $S^{(h)}(\boldsymbol{\theta})$ is computed as outlined in Section 3 [see Equation (3.5)]. In like manner, the Fisher scoring matrix $M(\boldsymbol{\theta}) = \sum M^{(h)}(\boldsymbol{\theta})$ is obtained by calculating the scoring matrix $M^{(h)}(\boldsymbol{\theta})$ for each h using (3.6) and formulas (3.3) and (3.4). The algorithm requires the derivatives of $P_h(t)$ with respect to the elements of $\boldsymbol{\theta}$, so a separate diagonalization of each Q_h is required. Note that the computation of $G^{(u)}$ in (3.4) requires calculation of the derivatives of $Q(\mathbf{z}_h)$ with respect to θ_u. Since $\theta_u = \beta_{cij}$ for some c, i, j, these derivatives are easily calculated; we find

$$\partial Q(\mathbf{z}_h)/\partial \beta_{cij} = z_{ch} q_{ij}(\mathbf{z}_h) L,$$

where L is a $k \times k$ matrix with all elements 0 except $L_{ij} = 1$ and $L_{ii} = -1$.

The specification and fitting of models with covariates is discussed further in Section 6.

6. ILLUSTRATIONS

In a study on smoking behavior, children from two Ontario counties (Waterloo and Oxford) who were entering sixth grade at time $t_0 = 0$ were surveyed at times (in years) $t_1 = .15$, $t_2 = .75$, $t_3 = 1.10$, and $t_4 = 1.90$. The study was designed to compare a control group from each county with a treatment group who received educational material on smoking during the first two months of study. A part of the information obtained at each follow-up time was the "smoking status" of each child, which is classified into three states:

State 1—child has never smoked.
State 2—child is currently a smoker.
State 3—child has smoked, but has now quit.

We consider a homogeneous Markov process with intensity matrix

$$Q = \begin{pmatrix} -\theta_1 & \theta_1 & 0 \\ 0 & -\theta_2 & \theta_2 \\ 0 & \theta_3 & -\theta_3 \end{pmatrix},$$

where $\theta_i > 0$ $(i = 1, 2, 3)$ as a model for the underlying continuous-time process for "smoking status." The purpose of our discussion is to illustrate the methodology developed in earlier sections, and not to present a thorough analysis of the data.

Figure 1 presents the transition counts n_{ijl} $(i, j = 1, 2, 3; l$

	1	2	3	
1	93 (96.0)	3 (1.7)	2 (.3)	98
2	0 (0)	8 (12.9)	10 (5.1)	18
3	0 (0)	1 (.5)	8 (8.5)	9

(t_0, t_1)

	1	2	3	
1	89 (85.7)	2 (4.2)	2 (3.1)	93
2	0 (0)	7 (4.0)	5 (8.0)	12
3	0 (0)	5 (2.8)	15 (17.2)	20

(t_1, t_2)

	1	2	3	
1	83 (84.9)	3 (2.9)	3 (1.2)	89
2	0 (0)	9 (6.8)	5 (7.2)	14
3	0 (0)	2 (2.3)	20 (19.7)	22

(t_2, t_3)

	1	2	3	
1	76 (74.5)	3 (4.3)	4 (4.3)	83
2	0 (0)	6 (3.7)	8 (10.3)	14
3	0 (0)	0 (4.3)	28 (23.7)	28

(t_3, t_4)

Figure 1. Transition Counts for Four Time Intervals. Expected transition counts are in parentheses.

	1	2	3	
1	61 (62.4)	1 (1.4)	2 (.2)	64
2	0 (0)	8 (11.5)	8 (4.5)	16
3	0 (0)	1 (.5)	7 (7.5)	8

	1	2	3	
1	59 (57.3)	1 (2.6)	1 (1.2)	61
2	0 (0)	7 (4.7)	3 (5.3)	10
3	0 (0)	3 (1.9)	14 (15.1)	17

	1	2	3	
1	56 (56.1)	2 (2.1)	1 (.8)	59
2	0 (0)	8 (5.9)	3 (5.1)	11
3	0 (0)	2 (1.7)	16 (16.3)	18

	1	2	3	
1	51 (51.2)	2 (2.9)	3 (1.9)	56
2	0 (0)	6 (4.4)	6 (7.6)	12
3	0 (0)	0 (2.6)	20 (17.4)	20

Figure 2. Transition Counts for High-Risk Individuals. Expected transition counts are in parentheses.

$= 1, \ldots, 4$) for the "Oxford treatment" group. From these data, the scoring algorithm of Section 3.1 gives the MLE of Q as

$$\hat{Q} = \begin{pmatrix} -.136 & .136 & 0 \\ 0 & -2.28 & 2.28 \\ 0 & .470 & -.470 \end{pmatrix}.$$

Initial estimates for the algorithm can be obtained in several ways. One possibility is to use a single interval (e.g., t_2, t_3) and to assume that individuals make 0 or 1 transitions in the interval. Thus, for example, we obtain an initial estimate of q_{11} from $\exp(q_{11}(t_3 - t_2)) = 83/89$. The corresponding estimates, .20, 1.26, and .27 for θ_i ($i = 1, 2, 3$) are adequate. The asymptotic covariance matrix for $\hat{\boldsymbol{\theta}}$ can also be obtained from (3.6) and used to give approximate confidence intervals for the θ's.

Questions of fit of the Markov model can, to some extent, be assessed by comparing observed transition frequencies, n_{ijl}, with expected frequencies, $e_{ijl} = n_{i.l}\hat{p}_{ij}(w_l)$, where $n_{i.l} = \sum_{j=1}^{k} n_{ijl}$. A likelihood ratio, or asymptotically equivalent Pearson chi-squared statistic to test the fit of the Markov model is readily obtained by methods similar to those used in Markov chains (e.g., Anderson and Goodman 1957). If none of the $p_{ij}(w_l)$'s is restricted to be zero, the likelihood ratio statistic is

$$\Lambda = 2 \sum_{l=1}^{m} \sum_{i,j=1}^{k} n_{ijl} \log(n_{ijl}/e_{ijl}),$$

which is asymptotically (m fixed, $n \to \infty$) a chi-squared variate on $mk(k - 1) - b$ degrees of freedom. The related Pearson statistic is

$$\chi^2 = \sum_{l=1}^{m} \sum_{i,j=1}^{k} (n_{ijl} - e_{ijl})^2/e_{ijl}.$$

For the example at hand, $p_{21}(w_l) = p_{31}(w_l) = 0$, and the degrees of freedom become $4m - b = 13$. We find that $\Lambda = 33.2$ and $\chi^2 = 19.9$. Although these are based on some small frequencies, it is apparent that there is fairly strong evidence against the Markov model.

It is common to find that time-homogeneous Markov models are not strictly appropriate. Fitting Markov models is, nonetheless, a useful and often necessary first step in developing a suitable model. The time-homogeneous Markov model is a convenient baseline; insight can be obtained by examining the nature of departures from this model.

In the present situation, other models provide a somewhat better fit to the data. The population is very heterogeneous, and one possibility is to examine more homogeneous subgroups. Data were collected on each child, which pertained to the exposure he/she received from relatives and friends that smoked. On this basis, each child was assigned to a "high risk" or a "low risk" group. The observed and expected transition rates from the high risk group are given in Figure 2. We find $\Lambda = 22.4$, and the fit of the Markov model is somewhat better.

Other possibilities also arise. Examination of Figures 1 and 2 shows that the number of transitions from State 2 to State 3 is unusually high in the first interval that immediately follows the treatment program. A Markov model describes much more accurately subsequent transition patterns in these data. Another possibility is to consider time dependence in the transition intensities. One approach is to fit separate models to different time intervals and to compare the estimated transition intensities. Another approach is to fit a parametric time-dependent model. For example, we considered models with

$$q_{ij}(t) = q_{ij}e^{-\lambda t} \quad \text{for} \quad i, j = 2, 3$$

to look for a monotone trend in the propensity of individuals to alter their smoking habits as they get older. This gives no significant improvement in this instance.

Comparisons among the four treatment-county groups are also of interest. These can be done using likelihood ratio tests or by the methods of Section 5, which utilize a regression

Table 1. Simulated Data From Model (6.1)

		(t_0, t_1)			(t_1, t_2)			(t_2, t_3)			(t_3, t_4)			(t_4, t_5)		
z		1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
(0,0)	1	13	2	0	14	1	2	12	2	1	11	0	2	11	1	0
	2	4	8	3	1	9	0	1	6	3	1	7	0	0	6	1
	3	0	0	0	0	0	3	0	0	5	0	0	9	0	0	11
(0,1)	1	14	1	0	16	0	1	19	0	0	17	2	1	16	1	2
	2	3	11	1	3	8	1	1	5	2	2	3	0	3	2	0
	3	0	0	0	0	0	1	0	0	3	0	0	5	0	0	6
(1,0)	1	11	4	0	10	2	2	10	2	3	10	2	2	10	1	0
	2	3	11	1	5	9	1	4	7	0	1	6	2	3	5	0
	3	0	0	0	0	0	1	0	0	4	0	0	7	0	0	11
(1,1)	1	11	3	1	12	1	1	17	1	0	17	0	1	13	1	3
	2	3	11	1	6	6	2	1	5	1	0	6	0	1	5	0
	3	0	0	0	0	0	2	0	0	5	0	0	6	0	0	7

NOTE: Entries are the numbers of transitions over the corresponding time interval.

Table 2. True Values, Maximum Likelihood Estimates, and Estimated Standard Errors for Model (6.1) Fitted to the Data in Table 1

Parameter	True Value	MLE	Estimated Standard Error
α_{12}	−2.30	−2.177	.356
β_1	.50	.700	.406
β_2	−.50	−.772	.406
α_{13}	−2.30	−2.659	.235
α_{21}	−1.90	−1.389	.278
β_3	.50	.284	.307
β_4	.50	.111	.304
α_{23}	−2.30	−2.246	.254

model. The Markov model, however, does not fit the data sufficiently well that one would want to base comparisons solely on it.

To illustrate the methods of Section 5, we consider data simulated from a three-state Markov process with transition matrix

$$\begin{pmatrix} q_{11}(z) & e^{\alpha_{12}+z_1\beta_3+z_2\beta_2} & e^{\alpha_{13}} \\ e^{\alpha_{21}+z_1\beta_3+z_2\beta_4} & q_{22}(z) & e^{\alpha_{23}} \\ 0 & 0 & 0 \end{pmatrix}, \tag{6.1}$$

which is a natural regression generalization of the marital status example leading to (2.1). For the simulation, the parameters were $\alpha_{12} = \alpha_{13} = \alpha_{23} = -2.30$, $\alpha_{21} = -1.90$, $\beta_1 = \beta_3 = \beta_4 = .50$, and $\beta_2 = -.50$. It was assumed that 15 individuals began in each of States 1 and 2 at time $t_0 = 0$, for each of four groups with regression vectors $(z_1, z_2) = (0, 0), (0, 1), (1, 0), (1, 1)$, respectively. The simulated data are presented in Table 1.

Convergence to the MLE's was rapid, and the region of convergence appeared to be reasonably broad. Convergence occurred from starting values obtained by a method similar to that used in the previous example. The MLE's and their estimated standard errors are given in Table 2. These results allow simple expression of approximate confidence intervals for the parameters. The overall likelihood ratio statistic for testing the fit of the model is $\Lambda = 78.69$ on $80 - 8 = 72$ df, indicating no lack of fit.

7. EMBEDDABILITY AND ESTIMABILITY

Consider again the homogeneous case of Section 2, and suppose that the observation times are equally spaced so that $w_l = w$ $(l = 1, \ldots, m)$. In addition, suppose that the model is saturated so that the elements of $\boldsymbol{\theta}$ are the intensities q_{ij} $(i \neq j)$ and the parameter space is

$$\mathbf{Q} = \{Q = (q_{ij}): q_{ij} \geq 0, q_{11} = -\sum_{j \neq i} q_{ij}\},$$

the set of admissible intensity matrices. In this case, the likelihood (3.1) can be written as

$$L(\boldsymbol{\theta}) = \prod_{i,j=1}^{k} p_{ij}(w)^{n_{ij.}}, \tag{7.1}$$

where $n_{ij.} = \sum_{l=1}^{m} n_{ijl}$ is the total number of recorded transitions from i to j. Since, for each i, the likelihood (7.1) is of multinomial form, a natural estimate of $P(w)$ is $\tilde{P}(w) = (\tilde{p}_{ij}(w))$, where

$$\tilde{p}_{ij}(w) = n_{ij.} \Big/ \sum_{j=1}^{k} n_{ij.}, \qquad i, j = 1, \ldots, k.$$

If there exists a matrix $\hat{Q} \in \mathbf{Q}$ such that

$$\tilde{P}(w) = \exp(\hat{Q}w), \tag{7.2}$$

if follows that $\hat{Q}$ is a maximum likelihood estimate of Q. When applicable, this method gives a simple procedure for maximizing the likelihood (7.1) with respect to $Q \in \mathbf{Q}$.

This approach leads one to consider those conditions on a stochastic matrix P under which the equation

$$\exp(Q) = P \tag{7.3}$$

admits a solution $Q \in \mathbf{Q}$. This is known as the embeddability problem for homogeneous Markov processes and has been discussed by many authors; for example, see Kingman (1962), Kingman and Williams (1973), Frydman (1980), who extended the problem to include nonhomogeneous processes, and Singer and Spilerman (1976a), who reviewed the area. If $k = 2$, P is embeddable if and only if $\text{tr}(P) > 1$. Simple necessary and sufficient conditions for general k are not known. Equation (7.3) can, however, admit 0, 1 or many solutions in Q for a given P.

Although characterization of embeddability is an interesting and challenging mathematical problem, the implications for the analysis of panel data would seem to be few. There are several reasons for this:

1. If $\tilde{P}(w)$ is nonembeddable, this does not imply that the continuous-time Markov model gives a poor fit to the data, nor that the parameters in that model are poorly estimated. Although $\tilde{P}$ is nonembeddable, there might be stochastic matrices "close" to $\tilde{P}$ that are embeddable (see Example 7.1 below).
2. If $\tilde{P}(w)$ is embeddable, it is quite possible that many of the parameters of the continuous-time model are nonetheless very poorly estimated (see Example 7.2).
3. The approach of estimating Q by solving (7.2) is applicable only if the model is saturated (parameter space is $\mathbf{Q}$) and the observation times are equally spaced. Models and sampling plans used in many applications do not meet these requirements. In addition, this approach allows no possibility of generalization to incorporate covariates.

These observations suggest that considerations of embeddability of $\tilde{P}(w)$ are largely irrelevant with respect to important statistical questions. In addition, even when the approach embodied in (7.2) is possible, it is preferable to consider maximization of (7.1) over the space $\mathbf{Q}$ directly. By so doing, estimates are obtained even if $\tilde{P}$ is nonembeddable and the covariance matrix M^{-1} provides estimates of precision.

We now give three examples that illustrate and extend points already made.

Example 7.1. (See remark 1 above.) Suppose that 100 individuals, beginning in each of $k = 3$ states, are observed over one transition of length $w = 1$, and the observed transition matrix is

$$\tilde{P} = \tilde{P}(1) = \begin{pmatrix} .8 & .2 & .0 \\ .1 & .8 & .1 \\ .2 & .1 & .7 \end{pmatrix}.$$

For any irreducible Markov process with intensity matrix Q, it is clear that all entries of $\exp(Q)$ are nonzero, and since $\tilde{p}_{13} = 0$, it follows that $\tilde{P}$ is nonembeddable. The approach to estimation embodied in (7.2) fails. The algorithm outlined in Section 2, however, converges to

$$\hat{Q} = \begin{pmatrix} -.237 & .237 & 0 \\ .111 & -.231 & .120 \\ .262 & .102 & -.364 \end{pmatrix},$$

which corresponds to

$$\hat{P} = \exp(\hat{Q}) = \begin{pmatrix} .800 & .189 & .011 \\ .100 & .810 & .090 \\ .200 & .100 & .700 \end{pmatrix},$$

in general good agreement with the observed transition rates. In addition, all elements of Q are well estimated. Although $\tilde{P}$ is nonembeddable, there are stochastic matrices "close" to $\tilde{P}$ that are embeddable.

Example 7.2. (See remark 2 above.) Suppose that

$$\tilde{P} = \begin{pmatrix} .51 & .49 \\ .49 & .51 \end{pmatrix}$$

is an estimate based on 100 individuals beginning in each of States 1 and 2 and observed for a single transition over a period of length $w = 1$. Since $\operatorname{tr}(\tilde{P}) > 1$, $\tilde{P}$ is embeddable, and there exists a unique matrix,

$$\hat{Q} = \log \tilde{P} = \begin{pmatrix} -1.956 & 1.956 \\ 1.956 & -1.956 \end{pmatrix},$$

the maximum likelihood estimate of Q. Application of the algorithm in Section 2 leads to this same estimate along with estimates of standard errors. It is easily seen that, although $\hat{Q}$ is unique, only the equilibrium distribution is estimated with precision. In effect, any Q with equilibrium distribution $(\pi_1, \pi_2) = (.5, .5)$ and with large entries will yield a corresponding P that is consistent with the observed transition frequencies.

Example 7.3. There have been many remarks in the literature concerning the possibility that (7.3) may admit multiple solutions $Q \in \mathbf{Q}$. When this occurs, there are multiple maximum likelihood estimates of Q. In this example, we examine conditions under which this can occur, in the case $k = 3$.

Let $Q = (q_{ij})$ be the intensity matrix of a three-state ergodic process, and let $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)'$ be the equilibrium distribution. Since $\boldsymbol{\pi}$ is the left eigenvector of the latent root $\lambda = 0$ of Q, it follows, after some algebra, that

$$Q = \gamma H^{-1} \begin{pmatrix} 0 & 0 & 0 \\ 0 & -1 - b & c \\ 0 & (a - b^2)/c & -1 + b \end{pmatrix} H,$$

where

$$H = \begin{pmatrix} \pi_1 & \pi_2 & \pi_3 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix} = \begin{pmatrix} 1 & \pi_2 + \pi_3 & -\pi_2 \\ 1 & -\pi_1 & \pi_1 + \pi_3 \\ 1 & -\pi_1 & -\pi_2 \end{pmatrix}^{-1},$$

$$\gamma = -(q_{11} + q_{22} + q_{33})/2,$$

and

$$-1 - b = \frac{q_{11} - q_{31}}{\gamma}, \quad c = \frac{q_{12} - q_{32}}{\gamma}, \quad \frac{a - b^2}{c} = \frac{q_{21} - q_{31}}{\gamma}.$$

If $a < 0$, the remaining two latent roots are complex, and if $a_1^2 = -a$, the relevant 2×2 submatrix is

$$\begin{pmatrix} -1 - b & c \\ (-a_1^2 - b^2)/c & -1 + b \end{pmatrix},$$

with eigenvalues $\lambda = -1 \pm a_1 i$. The left eigenvectors are the rows of

$$K = \begin{pmatrix} \dfrac{-b + a_1 i}{c} & 1 \\ \dfrac{-b - a_1 i}{c} & 1 \end{pmatrix} = c \begin{pmatrix} \dfrac{1}{2a_1 i} & -\dfrac{1}{2a_1 i} \\ \dfrac{b + a_1 i}{2a_1 c i} & \dfrac{-b + a_1 i}{2a_1 c i} \end{pmatrix}^{-1}.$$

For a fixed t, we let $P(t) = \exp(Qt)$ and determine a family of matrices Q_k $(k = 0, 1, 2, \ldots)$, which give rise to the same $P(t) = \exp(Q_k t)$. From the above,

$$Q = \gamma K^{-1} H^{-1} D H K,$$

where $D = \operatorname{diag}(0, -1 + a_1 i, -1 - a_1 i)$ so that

$$P(t) = K^{-1} H^{-1} \exp(\gamma t D) H K.$$

Let k be any integer $(k \neq 0)$ and define $a_1^* = a_1 + 2k\pi/t\gamma$. It can then be seen that $\exp(\gamma t D^*) = \exp(\gamma t D)$, where $D^* = (0, -1 + a_1^* i, -1 - a_1^* i)$. Define also $b^* = b a_1^*/a_1$ and $c^* = c a_1^*/a_1$. In an obvious notation, it follows that $K^* = K$ and the intensity matrix, Q_k say, corresponding to a_1^*, b^*, c^*, H, and γ satisfies

$$\exp(Q_k t) = \exp(Qt).$$

Some additional calculations verify that

$$Q_k = Q + (2k\pi/\gamma a_1 t)(Q + \gamma G),$$

where

$$G = \begin{pmatrix} \pi_2 + \pi_3 & -\pi_2 & -\pi_3 \\ -\pi_1 & \pi_1 + \pi_3 & -\pi_3 \\ -\pi_1 & -\pi_2 & \pi_1 + \pi_2 \end{pmatrix}.$$

If, for example, $\pi_1 = \pi_2 = \pi_3 = \frac{1}{3}$, $a_1 = \frac{1}{4}$, $c = \frac{1}{4}$, $b = 0$, and $\gamma = 1$, then

$$Q = \frac{1}{12} \begin{pmatrix} -9 & 2 & 7 \\ 6 & -7 & 1 \\ 3 & 5 & -8 \end{pmatrix}.$$

Let $P(t) = \exp(Qt)$. For a given t, the matrices

$$Q_k = \frac{1}{12} \begin{pmatrix} -9 & 2 & 7 \\ 6 & -7 & 1 \\ 3 & 5 & -8 \end{pmatrix} + \frac{8k\pi}{12t} \begin{pmatrix} -1 & -2 & 3 \\ 2 & 1 & -3 \\ -1 & 1 & 0 \end{pmatrix}$$

sastisfy $P(t) = \exp(Q_k t)$, $k = \cdots -1, 0, 1, \ldots$. For $Q_k \in \mathbf{Q}$, we require that all entries of Q_k have the proper sign. For fixed $k < 0$ (>0), this happens only if $t > -16k\pi/6$ $(t > 24k\pi)$. Thus, if $k = -1$, $Q_{-1} \in \mathbf{Q}$ only for $t > 16\pi/6 = 8.38$. There are multiple solutions in $\mathbf{Q}$ to $P(t) = e^{Qt}$ only for large t values. For t as large as 8.38, $P(t)$ is observationally in-

distinguishable from the equilibrium matrix with all rows $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$,

$$P(t) = \begin{pmatrix} .33319 & .33324 & .33357 \\ .33350 & .33332 & .33317 \\ .33331 & .33344 & .33326 \end{pmatrix}$$

to five figures.

This and similar examples suggest that multiple roots of (7.2) will not occur in practical applications, at least in the 3×3 case. Equivalently, when $\tilde{P}$ is embeddable there will not be multiple maximum likelihood estimates of Q unless $\tilde{P}$ is close to the equilibrium matrix. At this time, we have not undertaken a detailed study of higher-dimensioned problems ($k \geq 4$). If, however, multiple MLE's occur, methods similar to those outlined above could be used to determine all of them. Multiple MLE's indicate that the likelihood is badly behaved and point estimates are bound to be misleading.

APPENDIX A: DERIVATION OF (3.4)

Suppose that Q has distinct eigenvalues $d_1, \ldots, d_k$ for all $\boldsymbol{\theta}$ in some open set. The matrix obtained by differentiating each entry in $P(t)$ with respect to θ_u is, from (3.3) and the fact that $Q = ADA^{-1}$,

$$\begin{aligned} \frac{\partial P(t)}{\partial \theta_u} &= \sum_{s=1}^{\infty} \frac{\partial}{\partial \theta_u}\left(\frac{Q^s t^s}{s!}\right) \\ &= \sum_{s=1}^{\infty} \sum_{l=0}^{s-1} Q^l \frac{\partial Q}{\partial \theta_u} Q^{s-1-l} \frac{t^s}{s!} \\ &= \sum_{s=1}^{\infty} \sum_{l=0}^{s-1} AD^l G_u D^{s-1-l} A^{-1} \frac{t^s}{s!}, \end{aligned}$$

where $G_u = A^{-1}(\partial Q/\partial \theta_u) A$. Thus

$$\begin{aligned} \frac{\partial P(t)}{\partial \theta_u} &= A\left(\sum_{s=1}^{\infty} \sum_{l=0}^{s-1} D^l G_u D^{s-1-l} \frac{t^s}{s!}\right) A^{-1} \\ &= AV_u A^{-1}, \end{aligned}$$

where V_u is a $k \times k$ matrix with (i, j) element

$$g_{ij}^{(u)} \sum_{s=1}^{\infty} \sum_{l=0}^{s-1} d_i^{s-1-l} d_j^l \frac{t^s}{s!} = g_{ij}^{(u)} \frac{e^{d_i t} - e^{d_j t}}{d_i - d_j}, \quad i \neq j,$$

$$g_{ii}^{(u)} \sum_{s=1}^{\infty} \sum_{l=0}^{s-1} d_i^{s-1} \frac{t^s}{s!} = g_{ii}^{(u)} t e^{d_i t}, \quad i = j,$$

where $g_{ij}^{(u)}$ is the (i, j) element in G_u. This establishes (3.4).

APPENDIX B: DERIVATIONS OF (3.8) AND (3.9)

The $k \times k$ equilibrium probability vector $\boldsymbol{\pi}' = (\pi_1, \ldots, \pi_{k-1}, 1 - \pi_1 - \cdots - \pi_{k-1})$ can be obtained as the unique solution to the system of equations

$$Q_1' \boldsymbol{\pi} = \mathbf{0}, \tag{B.1}$$

where $Q_1 = Q_1(\boldsymbol{\theta})$ is the $k \times (k-1)$ matrix obtained by dropping the last column of $Q(\boldsymbol{\theta})$. Now, $\boldsymbol{\pi}$ is not a particularly simple function of $\boldsymbol{\theta}$ or Q_1, so we will develop derivatives $\partial \pi_i/\partial \theta_j$ by treating $Q_1'\boldsymbol{\pi} = F(\boldsymbol{\theta}, \boldsymbol{\pi})$ as a function defining $\pi_1, \ldots, \pi_{k-1}$ implicitly in terms of $\theta_1, \ldots, \theta_b$.

To obtain derivatives $\partial \pi_i/\partial \theta_j$ $(i = 1, \ldots, k-1)$ for fixed j, we use implicit differentiation of (B.1) with respect to θ_j to obtain the system of equations

$$B(\boldsymbol{\theta}) \begin{pmatrix} \partial \pi_1/\partial \theta_j \\ \vdots \\ \partial \pi_{k-1}/\partial \theta_j \end{pmatrix} + C_j(\boldsymbol{\theta}) = 0, \tag{B.2}$$

where $B(\boldsymbol{\theta})$ is a $(k-1) \times (k-1)$ matrix with (i, j) element

$$\frac{\partial F(\boldsymbol{\theta}, \boldsymbol{\pi})_i}{\partial \pi_j} = \frac{\partial (Q_1'\boldsymbol{\pi})_i}{\partial \pi_j} = q_{ji}(\theta) - q_{ki}(\boldsymbol{\theta}),$$

and where $C_j(\boldsymbol{\theta})$ is the $(k-1) \times 1$ vector given by

$$\frac{\partial F(\boldsymbol{\theta}, \boldsymbol{\pi})}{\partial \theta_j} = \left(\frac{\partial Q_1(\boldsymbol{\theta})}{\partial \theta_j}\right)' \boldsymbol{\pi}.$$

From (B.2), the derivatives $\partial \pi_i/\partial \theta_j$ $(i = 1, \ldots, k-1)$ are given by

$$(\partial \pi_1/\partial \theta_j, \ldots, \partial \pi_{k-1}/\partial \theta_j)' = -B(\boldsymbol{\theta})^{-1} C_j(\boldsymbol{\theta}).$$

The matrix giving all derivatives of $\pi_1, \ldots, \pi_{k-1}$ with respect to each of $\theta_1, \ldots, \theta_b$ is thus

$$W(\boldsymbol{\theta}) = (\partial \pi_i/\partial \theta_l) = -B(\boldsymbol{\theta})^{-1} C(\boldsymbol{\theta}),$$

where $C(\boldsymbol{\theta})$ is the $(k-1) \times b$ matrix with jth column $C_j(\boldsymbol{\theta})$. This establishes formula (3.8).

Formula (3.9) then follows by an application of the multivariate delta theorem (e.g., Rao 1973, p. 388).

[*Received April 1984. Revised June 1985.*]

REFERENCES

Anderson, T. W., and Goodman, L. A. (1957), "Statistical Inference About Markov Chains," *Annals of Mathematical Statistics*, 28, 89–110.

Bartholomew, D. J. (1982), *Stochastic Models for Social Processes* (3rd ed.), London: John Wiley.

——— (1983), "Some Recent Developments in Social Statistics," *International Statistical Review*, 51, 1–9.

Box, G. E. P., and Cox, D. R. (1964), "An Analysis of Transformations" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 26, 211–252.

Chambers, J. M. (1977), *Computational Methods for Data Analysis*, New York: John Wiley.

Cox, D. R., and Miller, H. D. (1965), *The Theory of Stochastic Processes*, London: Methuen (chap. 4).

Frydman, H. (1980), "The Embedding Problem for Markov Chains With Three States," *Mathematical Proceedings of the Philosophical Society*, 87, 285–294.

Hinkley, David V., and Runger, G. (1984), "The Analysis of Transformed Data," *Journal of the American Statistical Association*, 79, 302–309.

Jennrich, Robert I., and Bright, Peter B. (1976), "Fitting Systems of Linear Differential Equations Using Computer Generated Exact Derivatives," *Technometrics*, 18, 385–392.

Kingman, J. F. C. (1962), "The Imbedding Problem for Finite Markov Chains," *Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 1, 14–24.

Kingman, J. F. C., and Williams, D. (1973), "The Combinatorial Structure of Nonhomogeneous Markov Chains," *Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 26, 77–86.

Rao, C. R. (1973), *Linear Statistical Inference and Its Applications* (2nd ed.), New York: John Wiley.

Singer, B., and Spilerman, S. (1976a), "The Representation of Social Processes by Markov Models," *American Journal of Sociology*, 82, 1–54.

——— (1976b), "Some Methodological Issues in the Analysis of Longitudinal Surveys," *Annals of Economic and Sociological Measurement*, 5, 447–474.

Tuma, N. B., Hannan, M. T., and Groeneveld, L. P. (1979), "Dynamic Analysis of Event Histories," *American Journal of Sociology*, 84, 820–854.

Tuma, N. B., and Robins, P. K. (1980), "A Dynamic Model of Employment Behavior: An Application to the Seattle and Denver Income Maintenance Experiments," *Econometrica*, 48, 1031–1052.

Wasserman, S. (1980), "Analyzing Social Networks as Stochastic Processes," *Journal of the American Statistical Association*, 75, 280–294.