Article 1.1

Multi-State Markov Models in Survival Analysis

Guillermo Marshall

University of Colorado Health Sciences Center, Denver, U.S.A.

Abstract. This work discusses a general k-state Markov model when the process is observed at irregular intervals and the exact transition times are not available. This represents an extension of previous work by [13]. The Markov model is a generalization of parametric models in survival analysis and direct relations between the transition probabilities and transition intensities with survival functions are derived. Of particular interest is the application of this model to survival studies. An exact likelihood function is proposed to replace Kay's approximation when the exact transition time to the absorbing state is observed. A model with covariables and its likelihood function is provided in detail. Natural extensions to non-homogenous Markov models are discussed, and a model for time dependent covariables is proposed. As an example, a Markov model is applied to longitudinal data from a study of young patients with diabetes in order to describe the natural course of diabetic retinopathy.

Keywords: Multi-state models; Markov processes; Survival analysis; Longitudinal data; Time-dependent covariables.

Reference: This article was written based on the main results of my Doctoral Dissertation under the supervision of Richard H. Jones from University of Colorado, 1990. Although this article was never published in the present form, some of the most important results were published as an entry in the *Encyclopedia of Biostatistics* with the title Predictive Modelling for Prognosis, Article 1.4 of this book and in Article 1.5

1 Introduction

In recent years, Markov and semi-Markov models have become important tools to describe and help understand the progression and regression of multi-state diseases such as cancer, HIV infection, leukemia, diabetes, and many others. These models have been used by many authors in the area of biomedical sciences to find possible markers for the transition from stable states to the accelerated phase and/or the irreversible (absorbing) state of a disease, and also to describe the natural course of these diseases. Klein, Klotz and Grever [14] use a three-state Semi-Markov Model in a study of patients with chronic myelogenous leukemia to analyze the effect of elevated blood levels of adenosine deaminase as a marker for transition from stable disease to blast crisis, and then to death. Kay [13] proposed a methodology to fit a general k disease state Markov model in continuous time with application to the analysis of cancer markers in survival studies. Longini et al. [15] use a sub-model to describe the distribution of the incubation period for AIDS patients.

Important contributions in the area of continuous-time Markov models have been due to work in areas such as social sciences, demography and engineering. Schoen and Kenneth (1979) use Markov models to estimate increment-decrement life tables with applications to marital-status. Kalbfleisch and Lawless (1985) introduce a continuous-time Markov model to analyze panel data, and Kalbfleisch, Lawless and Vollmer (1983) propose methods to estimate the parameters of this model from aggregate data. Madsen, Spliid and Thyregod (1985) applied discrete and continuous-time Markov models to describe the variation of cloud cover at an airport.

This paper discusses a general k-state Markov model in which the exact transition times are not observed, and represents an extension of the work of Kay [13]. Of particular relevance is the extension of the relation between continuous-time Markov models and survival analysis functions. Multi-state Markov models represent a generalization of parametric models in survival analysis to the analysis of data concerning multiple events.

Multi-state Markov models are extended to include covariables in the transition intensities as in proportional hazard models [3] , and a general model based on a non-homogeneous Markov process is explored. Models for time- dependent covariables are also proposed. An important application of these models is discussed and analyzed. Data from a longitudinal study in young patients with diabetes from the Barbara Davis Center for Childhood Diabetes, University of Colorado Health Sciences Center, are used to determine factors affecting the natural course of diabetic retinopathy.

This paper is organized in four sections. Section 2 presents a brief review of the most relevant aspects of Markov processes in continuous time related to the methodology of the multi-state Markov models. Section 3 introduces the multi-state Markov model with covariables for a partially observed time-homogeneous Markov process. An exact likelihood function is proposed to replace Kay's approximation when the transition to the absorbing state is observed, and a complete relation between the multi-state model and survival analysis is derived. Finally, in Section 4, a non-homogeneous model is introduced.

2 Continuous-time Markov Processes

Introduction

Suppose we observe a continuous-time stochastic process $\{X(t), t \geq 0\}$ with a finite number of values in $E = \{1, 2, \ldots, k\}$ called states. We say that $\{X(t)\}$

is a continuous-time Markov process if for all times $t > s > u > 0$, and for any states i, j and $h \in E$,

$$pr\{X(t) = j|X(s) = i, X(u) = h\} = pr\{X(t) = j|X(s) = i\}. \tag{1}$$

This conditional probability represents the probability of a transition from the state i at time s to the state j at time t, and it is denotated as $p_{ij}(s,t)$. These transition probabilities have the basic properties, $0 \leq p_{ij}(s,t) \leq 1$, $p_{ii}(t,t) = 1$, $p_{ij}(t,t) = 0$ if $j \neq i$ and

$$\sum_{j=1}^{k} p_{ij}(s,t) = 1.$$

For any time τ in the interval (s,t) the transition probability $p_{ij}(s,t)$ can be written using the Chapman- Kolmogorov equation as

$$p_{ij}(s,t) = \sum_{v=1}^{k} p_{iv}(s,\tau)p_{vj}(\tau,t) \ .$$

This equation can be written in matrix notation as $\mathbf{P}(s,t) \ = \ \mathbf{P}(s,\tau)\mathbf{P}(\tau,t)$, where $\mathbf{P}(s,t)$ is the transition probability matrix of dimension $k \times k$ with elements $p_{ij}(s,t)$.

The Markov process $X(t)$ can also be characterized in terms of the transition intensities,

$$q_{ij}(t) = \lim_{dt \to 0} \frac{Pr\{\ X(t+dt) = j \ | \ X(t) = i\ \}}{dt} \ , \quad i \neq j$$

which represent instantaneous transition rates between the different states. For mathematical convenience, we define

$$q_{ii}(t) = -\sum_{j \neq i}^{k} q_{ij}(t).$$

The transition probability $p_{ij}(s,t)$ satisfies the Kolmogorov forward differential equations

$$\frac{dp_{ij}(s,t)}{dt} = \sum_{v=1}^{k} p_{iv}(s,t)q_{vj}(t),$$

or in matrix notation

$$\frac{d\mathbf{P}(s,t)}{dt} = \mathbf{P}(s,t)\mathbf{Q}(t) \tag{2}$$

with the initial condition $\mathbf{P}(t,t) = \mathbf{I}$, where $\mathbf{I}$ is the $k \times k$ identity matrix.

Time-homogeneous Markov processes

Important mathematical simplifications are obtained by assuming that the process $\{X(t), t \geq 0\}$ is homogeneous in time. The consequences of this assumption are that the transition intensities between the different states $q_{ij}(t)$ are constant over time, and the transition probabilities $p_{ij}(s,t)$ depend only on the time differences $t-s$. Equation (2) reduces to a system of differential equations with constant coefficients,

$$\frac{d\mathbf{P}(t-s)}{dt} = \mathbf{P}(t-s)\mathbf{Q},$$

for which the closed form solution is

$$\mathbf{P}(t-s) = e^{\mathbf{Q}(t-s)} = \sum_{n=0}^{\infty} \frac{\{\mathbf{Q}(t-s)\}^n}{n!}.$$

If $\mathbf{Q}$ has distinct eigenvalues, $\rho_1, \rho_2, \ldots, \rho_k$, and $\mathbf{A}$ is a square matrix where the jth column is the eigenvector associated with ρ_j, then we can calculate $\mathbf{P}(t-s)$ as

$$\mathbf{P}(t-s) \;=\; \mathbf{A}\mathrm{diag}\{\, e^{\rho_1(t-s)}, e^{\rho_2(t-s)}, \ldots, e^{\rho_k(t-s)} \,\}\mathbf{A^{-1}} \,. \tag{3}$$

Individual transition probabilities can be calculated, for any value of $t-s$, as

$$p_{ij}(t-s) = \sum_{v=1}^{k} a_{iv} e^{\rho_v(t-s)} a^{vj}, \tag{4}$$

where a_{ij} and a^{ij} represent the elements (i,j) of the matrices $\mathbf{A}$ and $\mathbf{A^{-1}}$. For more details about continuous-time Markov processes see Cox and Miller [2] and Chiang [1].

Non-homogeneous Markov processes

Non-homogeneous Markov processes are natural generalizations when transition rates change over time. Although this extension is important, the complexities of the mathematics and the computational difficulties are important obstacles in practical applications. Although a solution of the system of the differential equations (2) exists for a general form of Q(t), a closed form solution exists only for particular cases. If $\mathbf{Q}(t)$ is triangular, for example, processes with unidirectional transition like HIV infection, then a solution of (2) can be found by sequential integration. A good discussion of this method can be found in Raman and Chiang (1973) and Davies (1985).

An alternative method for finding a solution of (2) for a general $\mathbf{Q}(t)$ is to approximate the intensity matrix $\mathbf{Q}(t)$ by dividing the period of follow-up into K intervals $[\tau_1, \tau_2), [\tau_2, \tau_3), \ldots, [\tau_K, \infty)$ and assuming that the intensity matrix is constant during each interval. Hence, the intensity matrix $\mathbf{Q}(t)$ is approximated by a step function

$$\mathbf{Q}(t) = \mathbf{Q}_l, \quad t \in [\tau_l, \tau_{l+1}),$$

for $l = 1, 2, \ldots, K$. Local solutions of the Kolmogorov differential equations are

$$\mathbf{P}(\tau_l, t) = e^{\mathbf{Q}_l(t-\tau_l)},$$

where $t \in [\tau_l, \tau_{l+1})$, subject to the condition $\mathbf{P}(\tau_l, \tau_l) = \mathbf{I}$, for $l = 1, 2, \ldots, K$.

Using the Chapman-Kolmogorov equation and the above local solution, we can find a global solution as follows. For any two times s and t with $s \leq t$

$$\begin{aligned}\mathbf{P}(s,t) &= \mathbf{P}(s, \tau_{l+1})\mathbf{P}(\tau_{l+1}, \tau_{l+2}) \cdots \mathbf{P}(\tau_m, t) \\ &= e^{\mathbf{Q}_l(\tau_{l+1}-s)} e^{\mathbf{Q}_{l+1}(\tau_{l+2}-\tau_{l+1})} \ldots e^{\mathbf{Q}_m(t-\tau_m)},\end{aligned} \tag{5}$$

where $\tau_l < s \leq \tau_{l+1}$, $\tau_m < t \leq \tau_{m+1}$.

In the special case when the transition intensity matrices $\mathbf{Q}_1, \ldots, \mathbf{Q}_K$ permute, $\mathbf{Q}_i\mathbf{Q}_j = \mathbf{Q}_j\mathbf{Q}_i$, the expression (5) can be reduced to the exponent of the sum of the $\mathbf{Q}_i$ matrices. In general, if for any two times s and t, $\mathbf{Q}(s)\mathbf{Q}(t) = \mathbf{Q}(t)\mathbf{Q}(s)$, (2) has a closed form solution,

$$\mathbf{P}(s,t) = e^{\int_s^t \mathbf{Q}(u)du} \ . \tag{6}$$

A basic example of a transition intensity matrix that satisfies this is $\mathbf{Q}(t) = \mathbf{Q}h(t)$, where $h(t)$ is any positive function of t. In this case the closed form solution (6) reduces to

$$\mathbf{P}(s,t) = e^{\mathbf{Q}\int_s^t h(u)du} \ .$$

If h(t) = 1 we have the solution of the time-homogeneous Markov process discussed in the previous section. For more details about the system of equations (2) and methods to obtain its solutions see Gantmacher [7] and Hochstadt [9].

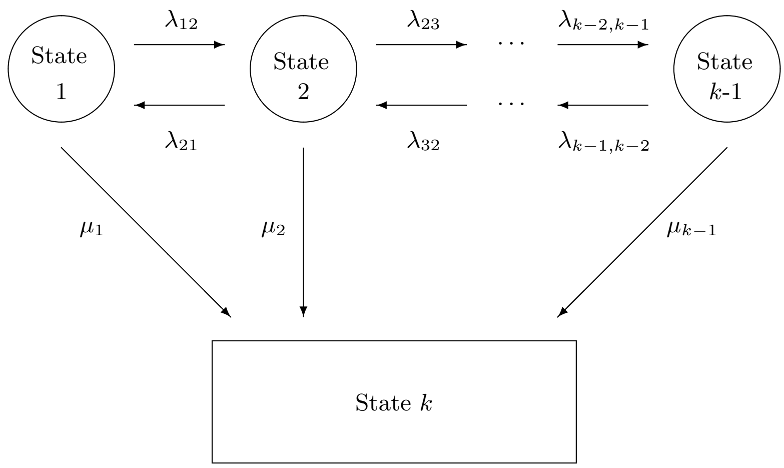
Computation of the transition probability matrix $\mathbf{P}(s,t)$ using equation (5) can be obtained using canonical decompositions for each of the terms $\exp\{\mathbf{Q}_l(\tau_{l+1} - \tau_l)\}$ as in (3) for the time- homogeneous case. The element (i, j) of the resulting product of matrices is $p_{ij}(s,t)$. The computing cost of calculating $p_{ij}(s,t)$ will depend on how small the intervals (τ_l, τ_{l+1}) are chosen, and the length of the intervals between observations of the process.

3 Time-Homogeneous Markov Models

The Basic Model

Suppose we have a model with $k-1$ transient disease states $j = 1, \ldots, k-1$ and a single absorbing state $j = k$, see Figure 1. The transient states are assumed to be ordered according to j and instantaneous transitions can take place from state j to the adjoining states $j-1$ or $j+1$. Transitions can also take place from any of the transient states to the absorbing state k.

Figure 1. A multi-state model with $k-1$ transient states and one absorbing state. The model has a total of $3k-5$ parameters, $2k-4$ λ's and $k-1$ μ's.



Submodels can be obtained by eliminating some of the parameters when some of the transitions are theoretically impossible or are unlikely to be observed during the study time. Longini et al. [15] use a submodel to describe the incubation period of AIDS with only progression transitions to the adjoining states.

For the model in Figure 1 the transition intensity matrix $\mathbf{Q}$ can be written as

$$\mathbf{Q} = \begin{pmatrix} -(\mu_1+\lambda_{12}) & \lambda_{12} & \cdots & 0 & \mu_1 \\ \lambda_{21} & -(\mu_2+\lambda_{21}+\lambda_{23}) & \cdots & 0 & \mu_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -(\mu_{k-1}+\lambda_{k-1,k-2}) & \mu_{k-1} \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix} .$$

When equally spaced observations are available, and where a discrete time Markov model can be considered, a continuous- time Markov model will be more parsimonious. This model has $3k-5$ different parameters in contrast with a discrete time model with $k(k-1)$ independent parameters.

Given the form of the transition intensity matrix $\mathbf{Q}$, since all the λ's and μ's are non–negative, the eigenvalues of $\mathbf{Q}$ have negative real parts (Cox and Miller, 1965), except that $\rho_k = 0$. A necessary condition for all the eigenvalues of $\mathbf{Q}$ to be real is that

$$\lambda_{i,i+1}\lambda_{i+1,i} \geq 0 \quad , \quad i = 1, \ldots, k-1 \ .$$

The proof of this follows from the fact that the eigenvalue associated with the kth row of $\mathbf{Q}$ is always equal to zero, and that the $(k-1)\times(k-1)$ upper left sub-matrix of $\mathbf{Q}$ is a tri–diagonal matrix. If a tri- diagonal matrix satisfies the above restrictions then it is said to be symmetrizable and have a real eigensystem [19].

Let $\mathbf{R}$ be the $(k-1)\times(k-1)$ upper left sub-matrix of $\mathbf{Q}$. $\mathbf{R}$ is symmetrizable by a similarity transformation with diagonal matrix $\mathbf{D}$. If we define $\mathbf{D}$ as $d_{11} = 1$ and $d_{ii} = (\lambda_{21}\lambda_{32}\cdots\lambda_{i,i-1})/(\lambda_{12}\lambda_{23}\cdots\lambda_{i-1,i})$ for $i = 2,\ldots,k-1$, then

$$\mathbf{T} = \mathbf{D}^{-\frac{1}{2}}\mathbf{R}\mathbf{D}^{\frac{1}{2}}$$

is a tri-diagonal symmetric matrix with elements $t_{ii} = q_{ii}$ and $t_{i,i+1} = t_{i+1,i} = (\lambda_{i,i+1}\lambda_{i+1,i})^{\frac{1}{2}}$.

If a $\lambda_{i,i+1}$ or $\lambda_{i+1,i}$ is zero, the eigenvalues of $\mathbf{R}$ are the eigenvalues of a number of smaller tri-diagonal matrices, so that this case does not cause difficulties. Once we transform $\mathbf{R}$ to $\mathbf{T}$, we can use standard routines for finding the eigensystem in tri-diagonal symmetric matrices [10, 19]

Regression Models

An extended model can be developed by introducing covariables in the basic model discussed in the previous section. As in the proportional hazard model (Cox, 1972), a proportional transition intensity model can be used for each possible transition of the process.

Suppose that each individual under study has an associated vector of covariables $\mathbf{z}' = (z_1, z_2, z_3, ..., z_p)$, then for a given $\mathbf{z}$, we assume that the Markov process is homogeneous with an intensity matrix $\mathbf{Q}(\mathbf{z})$ with elements

$$\lambda_{ij}(\mathbf{z}) = \lambda_{ij}\psi(\mathbf{z};\boldsymbol{\beta}_{ij}) \quad \text{and} \quad \mu_i(\mathbf{z}) = \mu_i\psi(\mathbf{z};\boldsymbol{\beta}_{ij}),$$

where λ_{ij} and μ_i represent the baseline transition rates, and $\boldsymbol{\beta}_{ij}$ represents the vector of regression coefficients associated with $\mathbf{z}$.

Different parametric forms of ψ can be considered such as the linear form $\psi(\mathbf{z};\boldsymbol{\beta}_{ij}) = 1 + \boldsymbol{\beta}'_{ij}\mathbf{z}$, the logistic, $\psi(\mathbf{z};\boldsymbol{\beta}_{ij}) = \log(1 + e^{\boldsymbol{\beta}'_{ij}\mathbf{z}})$, or the log linear form $\psi(\mathbf{z};\boldsymbol{\beta}_{ij}) = e^{\boldsymbol{\beta}'_{ij}\mathbf{z}}$. When referring to this model we will assume a log linear form since it is easy to apply and is always positive. For a further discussion of these parametric forms see Cox and Oakes (1984).

The solution of the Kolmogorov forward system of equations for this model is $\mathbf{P}(t-s;\mathbf{z}) = \exp\{\mathbf{Q}(\mathbf{z})(t-s)\}$, and a spectral decomposition of $\mathbf{Q}(\mathbf{z})$ can be used to calculate $\mathbf{P}(t-s;\mathbf{z})$ for any value of $t-s$ from equation (3). Individual transition probabilities can be evaluated for any value of $t-s$, using an expression similar to (4),

$$p_{ij}(t-s;\mathbf{z}) = \sum_{v=1}^{k} a_{iv}(\mathbf{z}) e^{\rho_v(\mathbf{z})(t-s)} a^{vj}(\mathbf{z}), \tag{7}$$

where $a_{ij}(\mathbf{z})$ and $a^{ij}(\mathbf{z})$ represent the elements (i,j) of the matrices $\mathbf{A}(\mathbf{z})$ and $\mathbf{A}(\mathbf{z})^{-1}$, and where $\rho_1(\mathbf{z}), \rho_2(\mathbf{z}), \ldots, \rho_k(\mathbf{z})$ and $\mathbf{A}(\mathbf{z})$ are the eigenvalues and the matrix of eigenvectors of the transition intensity matrix $\mathbf{Q}(\mathbf{z})$.

Survival Analysis

A point of major interest in practical applications is the relationship between this Markov model and survival analysis functions, including the survival function, the hazard function, the median lifetime, the mean lifetime and the residual mean lifetime. Let T be a random variable which represents the lifetime of individuals in a homogeneous population. The survival function, given that the process is in state i at time $s=0$, $S_i(t) = pr\{T > t|X(0) = i\}$, for a subject with covariables $\mathbf{z}$, is

$$S_i(t|\mathbf{z}) = 1 - p_{ik}(t;\mathbf{z}),$$

where $p_{ik}(t;\mathbf{z})$ is the element (i,k) of the transition probability matrix $\mathbf{P}(t;\mathbf{z})$. The density function, expressed in terms of the survival function, $f_i(t) = -dS_i(t)/dt$, for a subject with covariables $\mathbf{z}$, is

$$f_i(t|\mathbf{z}) = \sum_{j=1}^{k-1} p_{ij}(t;\mathbf{z})\mu_j(\mathbf{z}) \ .$$

The hazard function $h_i(t|\mathbf{z}) = f_i(t|\mathbf{z})/S_i(t|\mathbf{z})$ as a function of the transition probabilities and intensities is

$$h_i(t|\mathbf{z}) = \sum_{j=1}^{k-1} \frac{p_{ij}(t;\mathbf{z})}{\sum_{v=1}^{k-1} p_{iv}(t;\mathbf{z})} \mu_j(\mathbf{z}),$$

which represents a weighted mean of the transition rates from the transient states to the absorbing state k.

The median lifetime from the transient state i to the absorbing state k is defined as the value of $t = \xi_i$ that satisfies $p_{ik}(\xi_i;\mathbf{z}) = 0.5$. The mean lifetime, $E_i = E\{T|X(0) = i\}$, is also a parameter of interest. Again, in terms of the transition probabilities, the mean lifetime is

$$E_i(\mathbf{z}) = \int_0^\infty S_i(t|\mathbf{z})\ dt \ = \sum_{j=1}^{k-1}\sum_{v=1}^{k-1} a_{iv}(\mathbf{z})(-\frac{1}{\rho_v(\mathbf{z})})a^{vj}(\mathbf{z}),$$

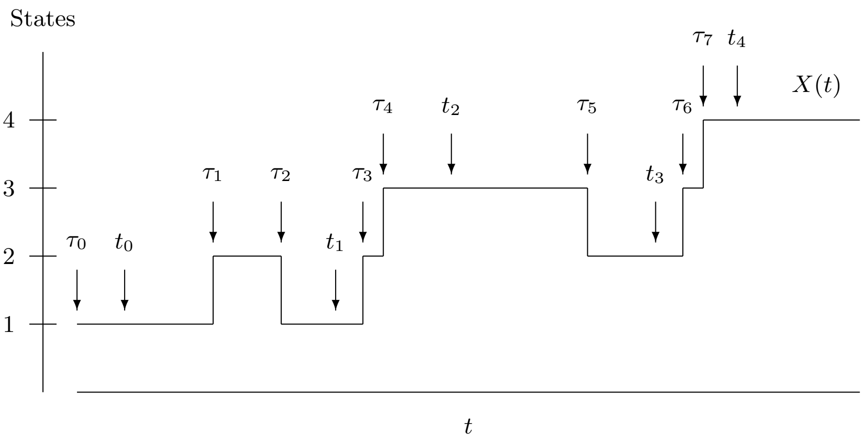
provided $\rho_v(\mathbf{z}) < 0,\ foreverystatev < k$. Finally, the residual mean lifetime, $m_i(t) = E\{T - t|X(0) = i\}$, can be expressed in terms of the Markov model as

$$m_i(t|\mathbf{z}) = \frac{\int_t^\infty S_i(u|\mathbf{z})\ du}{S_i(t|\mathbf{z})} = \frac{\sum_{j=1}^{k-1}\sum_{v=1}^{k-1} \frac{1}{\rho_v(\mathbf{z})} a_{iv}(\mathbf{z}) e^{\rho_v(\mathbf{z})t} a^{vj}(\mathbf{z})}{1 - p_{ik}(t;\mathbf{z})} \ .$$

The Data

The type of data collected will be different in each application and is directly dependent on the nature of the process. When the exact transition times of the process $\tau_1, \cdots, \tau_{m'}$ are available, see Figure 2, the statistical methods for estimating the parameters of the multi-state model are straightforward. Closed form solutions for the maximum likelihood estimates can be derived for the basic model, and an approach similar to fitting exponential regression models can be used for the model with covariables.

Figure 2. An example of a process with 4 states where the τ_i are the actual transition times and where the t_i are the observation times of the process.



In clinical studies in which each realization of the process is a different patient, it is very unusual to observe the exact transition times. The typical information available are the visits of the patients to the clinic, $t_0, t_1, \cdots, t_{m+1}$, as shown also in Figure 2. We assume that the data obtained on each subject are unequally spaced in time, and that the exact transition times are not available. For a model of $k = 4$ states, the following data correspond to weeks from the date of diagnosis and the state of the disease of the patient at that specific date [13].

Patient	*Data*
1	$(0,2)$ $(41,2)$ $(78,1)$ $(95,3)$ $(104,4)$
2	$(0,1)$ $(17,1)$ $(52,4)$
3	$(0,3)$ $(23,2)$ $(58,3)$ $(72,2)$

At the date of diagnosis, patient 2 was in state 1, and seventeen weeks later patient 2 was in state 1. This patient could have remained in state 1 for the whole 17 weeks, or could have transferred out of state 1 and back in again.

Thirty-five weeks later, at week 52, the patient died. Survival times for these patients are 104 weeks for patient 1, 52 weeks for patient 2 and more than 72 weeks for patient 3. The data of patient 3 represent a censored observation.

The Likelihood Function

Suppose that the observation times of the process for a subject are $t_0 < t_1 < \cdots < t_m < t_{m+1}$, and that $x(t_0) = i_0, x(t_1) = i_1, \ldots, x(t_{m+1}) = i_{m+1}$ represent the observed states of the process at these particular times. Then the joint distribution of $X(t_1), X(t_2), \ldots, X(t_{m+1})$ given $X(t_0)$ and the vector of covariables $\mathbf{z}$ can be represented, using the Markov property (1) and conditional probabilities, as

$$\prod_{j=1}^{m+1} p_{i_{j-1},i_j}(\Delta t_j; \mathbf{z}), \tag{8}$$

where $\Delta t_j = t_j - t_{j-1}$.

The above expression represents the contribution to the likelihood function for this subject if the arrival time at the absorbing state is interval-censored, in other words, if t_{m+1} is not the known time of transition to the absorbing state. In survival studies, t_{m+1} may represent the exact arrival time at the absorbing state k, which is the lifetime, or it may represent the end of the study for this subject, which is the censuring time.

Let $T = \tau$ be the exact arrival time at k, and c be the censoring time for this subject. Then

$$t_{m+1} = \min(\tau, c) \quad \text{and } \delta = \begin{cases} 1 \text{ if } \tau \leq c \\ 0 \text{ if } \tau > c \,. \end{cases}$$

If $\delta = 1$, the contribution of this last transition to the likelihood is

$$f_{i_m}(\Delta t_{m+1}|\mathbf{z}) = \sum_{j=1}^{k-1} p_{i_m,j}(\Delta t_{m+1}; \mathbf{z}) \mu_j(\mathbf{z}) \;,$$

and if $\delta = 0$ the contribution is $S_{i_m}(\Delta t_{m+1}|\mathbf{z}) = 1 - p_{i_m,k}(\Delta t_{m+1}; \mathbf{z})$. The above expression for $\delta = 1$ is a continuous time version of Kay's likelihood contribution [13].

The likelihood function for this subject can be then written as

$$L_h(\boldsymbol{\theta}) = \prod_{j=1}^{m} p_{i_{j-1},i_j}(\Delta t_j; \mathbf{z}) \{f_{i_m}(\Delta t_{m+1}|\mathbf{z})\}^{\delta} \{S_{i_m}(\Delta t_{m+1}|\mathbf{z})\}^{1-\delta}. \tag{9}$$

The full likelihood can be obtained from the product of the individual contributions. The subject subscript h has been omitted for m, i_j, t_j, $\mathbf{z}$ and δ in the expression (9) and in the rest of this paper for simplicity.

Without the first term in expression (9) this likelihood function is equal to the likelihood for parametric models in survival analysis with censored observations. For a model with two states $k = 2$, and with constant transition intensities, this Markov model reduces to a survival analysis model using the exponential distribution. In particular $p_{11}(t|\mathbf{z}) = \exp\{-\mu(\mathbf{z})t\}$ and $p_{12}(t|\mathbf{z}) = 1 - \exp\{-\mu(\mathbf{z}t\}$, therefore the above contribution to the likelihood is $\{f(t_{m+1}|\mathbf{z})\}^{\delta}\{S(t_{m+1}|\mathbf{z})\}^{1-\delta}$.

The likelihood function (9) can be extended to the case of time-dependent covariables $\mathbf{z}(t)$ by replacing $\mathbf{z}$ by $\mathbf{z}_{j-1}$, where the covariables are asummed to be constant between two obsertations

$$\mathbf{z}(t) = \mathbf{z}_{j-1} \quad \text{for} \quad t_{j-1} \leq t < t_j.$$

Parameter Estimation

Let $\boldsymbol{\theta}_{ij} = (\log \lambda_{ij}, \beta_{ij1}, \ldots, \beta_{ijp})$ be the parameters associated with the transition between states i to j, and $\boldsymbol{\theta}$ be a vector made up of the $\boldsymbol{\theta}_{ij}$ vectors. A log transformation is used to prevent the baseline transition intensities from taking negative values during the iterative process of estimation. Let $\eta_{ij} = \log q_{ij}(\mathbf{z})$ be the log transition intensity for a subject with covariables $\mathbf{z}$

$$\eta_{ij} = \log \lambda_{ij} + \beta_{ij1}z_1 + \cdots + \beta_{ijp}z_p.$$

Maximum likelihood estimates of $\boldsymbol{\theta}$ can be found by maximizing the log-likelihood function $l(\boldsymbol{\theta}) = \sum l_h(\boldsymbol{\theta})$ where

$$l_h(\boldsymbol{\theta}) = \sum_{j=1}^{m} \log\{p_{i_{j-1},i_j}(\Delta t_j; \mathbf{z})\} + \delta \log\{f_{i_m}(\Delta t_{m+1}|\mathbf{z})\} + (1-\delta)\log\{S_{i_m}(\Delta t_{m+1}|\mathbf{z})\}.$$

The first derivative of the log-likelihood function with respect to θ_{uvl} is

$$\frac{dl_h(\boldsymbol{\theta})}{d\theta_{uvl}} = \left\{ \sum_{j=1}^{m} \frac{1}{p_{i_{j-1},i_j}(\Delta t_j; \mathbf{z})} \frac{dp_{i_{j-1},i_j}(\Delta t_j; \mathbf{z})}{d\eta_{uv}} + \frac{\delta}{f_{i_m}(\Delta t_{m+1}|\mathbf{z})} \frac{df_{i_m}(\Delta t_{m+1}; \mathbf{z})}{d\eta_{uv}} \right.$$
$$\left. + \frac{1-\delta}{S_{i_m}(\Delta t_{m+1}|\mathbf{z})} \frac{dS_{i_m}(\Delta t_{m+1}; \mathbf{z})}{d\eta_{uv}} \right\} \frac{d\eta_{uv}}{d\theta_{uvl}},$$

where the derivative of $f_i(t|\mathbf{z})$ with respect to η_{uv} is

$$\frac{df_i(t|\mathbf{z})}{d\eta_{uv}} = \sum_{j=1}^{k-1} \{ \frac{dp_{ij}(t; \mathbf{z})}{d\eta_{uv}} e^{\eta_{jk}} + p_{ij}(t; \mathbf{z}) \frac{de^{\eta_{jk}}}{d\eta_{uv}} \},$$

and where

$$\frac{dS_i(t|\mathbf{z})}{d\eta_{uv}} = -\frac{dp_{ik}(t; \mathbf{z})}{d\eta_{uv}}.$$

The derivative $dp_{ij}(t;\mathbf{z})/d\eta_{uv}$ in the three expressions above is

$$\frac{dp_{ij}(t;\mathbf{z})}{d\eta_{uv}} = \sum_{r=1}^{k}\sum_{s=1}^{k} a_{ir}(\mathbf{z})w_{rs}^{uv}(t;\mathbf{z})a^{sj}(\mathbf{z}) \tag{10}$$

where

$$w_{rs}^{uv}(t;\mathbf{z}) = \begin{cases} g_{rs}^{uv}(\mathbf{z})(e^{\rho_r(\mathbf{z})t} - e^{\rho_s(\mathbf{z})t})/(\rho_r(\mathbf{z}) - \rho_s(\mathbf{z})) & \text{if } r \neq s \\ g_{rr}^{uv}(\mathbf{z})te^{\rho_r(\mathbf{z})t} & \text{if } r = s \ , \end{cases}$$

and $g_{rs}^{uv}(\mathbf{z})$ is the (r,s) entry in

$$\mathbf{G}^{uv}(\mathbf{z}) = \mathbf{A}(\mathbf{z})^{-1}\frac{d\mathbf{Q}(\mathbf{z})}{d\eta_{uv}}\mathbf{A}(\mathbf{z}) \ .$$

The derivative of $d\eta_{uv}/d\theta_{uvl}$ is

$$\frac{d\eta_{uv}}{d\theta_{uvl}} = \begin{cases} 1 & \text{for } \ l = 1 \\ z_{l-1} & \text{for } \ l = 2, \ldots, p+1. \end{cases}$$

The subscripts u and v refer to the transition between states u to v.

Quasi-Newton algorithms can be used to minimize $-2l(\boldsymbol{\theta})$ using only the likelihood function and finite differences to obtain numerical approximations of the derivatives, or by using the likelihood function and explicit expressions for the derivatives. A discussion of these two approaches can be found in Dennis and Schnabel [6] as well as a modular system of algorithms for unconstrained minimization. Dennis and Schnabel [6] also provide algorithms for computing numerical approximations of the second derivatives of the log–likelihood using finite differences of the original function or the gradients if they are available.

Once we have the maximum likelihood estimates of the parameters of the transition intensity matrix $\mathbf{Q}(\mathbf{z})$, we also have estimates of the transition probability matrix $\mathbf{P}(t;\mathbf{z};\boldsymbol{\theta})$. In particular, an estimates of $p_{ij}(t;\mathbf{z};\boldsymbol{\theta})$ can be obtained as $p_{ij}(t;\mathbf{z};\hat{\boldsymbol{\theta}})$.

An estimate of the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$ is obtained by inverting the negative of the empirical information matrix,

$$\widehat{\mathbf{V}}(\hat{\boldsymbol{\theta}}) = -\left\{\frac{d^2L(\boldsymbol{\theta})}{d\boldsymbol{\theta}d\boldsymbol{\theta}'}\right\}_{\theta=\hat{\theta}}^{-1} \ .$$

The estimate of the asymptotic variance of $p_{ij}(t;\mathbf{z};\hat{\boldsymbol{\theta}})$ can be found using the δ method as

$$\widehat{V}\{p_{ij}(t;\mathbf{z};\hat{\boldsymbol{\theta}})\} = \left\{\frac{dp_{ij}(t;\mathbf{z};\boldsymbol{\theta})}{d\boldsymbol{\theta}}\right\}_{\theta=\hat{\theta}}' \widehat{\mathbf{V}}(\hat{\boldsymbol{\theta}}) \left\{\frac{dp_{ij}(t;\mathbf{z};\boldsymbol{\theta})}{d\boldsymbol{\theta}}\right\}_{\theta=\hat{\theta}}$$

where $dp_{ij}(t;\mathbf{z};\boldsymbol{\theta})/d\boldsymbol{\theta}$ can be evaluated using expression (10). An estimate of the survival function can be obtained directly from the estimate of the transition probability matrix as $S_i(t|\mathbf{z}) = 1 - p_{ik}(t;\mathbf{z};\hat{\boldsymbol{\theta}})$. An approximate variance for $\widehat{S}(t|\mathbf{z})$ is obtained as

$$\widehat{V}\{\hat{S}(t|\mathbf{z})\} = \widehat{V}\{p_{ik}(t;\mathbf{z};\hat{\boldsymbol{\theta}})\}.$$

The Natural Course of Diabetic Retinopathy

Diabetic retinopathy currently is the leading cause of new cases of blindness in people aged 20 to 74 years in the United States, and is considered a progressive disease among people with insulin-dependent (type I) diabetes mellitus (IDDM).

Improvement of early stages of retinopathy as part of the natural course has been poorly understood. In the past, physicians believed diabetic retinopathy was a strictly progressive disease. Using a basic multi-state Markov model Garg, Marshall, Chase, et al. (1990) have shown that the natural course of early diabetic retinopathy involves both progression and regression.

The natural course of early diabetic retinopathy in young subjects with type I diabetes was evaluated during 693 patient visits for 259 subjects over a mean of 2.4 years. All 259 subjects had direct ophthalmoscopy (with pupils dilated) by at least two examiners (one ophthalmologic and one pediatric), followed by color retinal photography, intravenous fluorescein angiography , and slit-lamp examinations. The retinal specialist graded retinal findings with a modified Airlie House classification of diabetic retinopathy. A grade of I indicates no retinopathy; grades II–III, microaneurysms or microaneurysms and one other finding; grades IV-V, advanced background changes with intra retinal microvascular abnormalities; and grade VI, proliferative retinopathy. The category assigned was that of the more severely involved eye.

Based on this classification, a 4–state Markov model was used considering grades I, II–III and IV–V as transient states and grade VI as an absorbing state, as shown in Figure 3. In this case the exact arrival times at the absorbing state were interval-censored and the likelihood function (8) was used to estimate the paramaters.

Figure 3. A model with covariables for diabetic retinopathy

$$\text{Grade I} \underset{\lambda_{21}(\mathbf{z})}{\overset{\lambda_{12}(\mathbf{z})}{\rightleftarrows}} \text{Grades II–III} \underset{\lambda_{32}(\mathbf{z})}{\overset{\lambda_{23}(\mathbf{z})}{\rightleftarrows}} \text{Grades IV–V} \xrightarrow{\mu_3} \text{Grade VI}$$

The influence of duration of diabetes, HbA1, age and sex on the transition intensities between various stages of diabetic retinopathy was evaluated. A model was fit for each of these four factors for all transitions except the transition from state 3 to the absorbing state, as shown in Figure 3. Duration of diabetes, HbA1 and age are time dependent covariables, but they were very stable between two observed times.

Table 1. Summary of models fit

N	Model	Parameters	$-2LogLike$	χ^2	P–value	Ref	AIC
1	Null	5	602.66	–	–	–	612.66
2	HbA1	9	592.85	9.81	0.044	1	610.85
3	Age	9	586.90	15.76	0.003	1	604.90
4	Duration	9	580.93	21.73	0.001	1	598.93
5	Sex	9	599.21	3.45	0.486	1	617.21
6	HbA1+			14.84	0.005	4	
	Duration	13	566.09	26.76	0.001	2	592.09*
7	Age+			5.82	0.213	4	
	Duration	13	575.11	11.79	0.019	3	601.11

* Best model

Table 1 shows a summary of the models and the significance of the effect of each factor. Age is a significant factor in the development of diabetic retinopathy, but its effect is partially due to the confounding effect of duration of diabetes. Model 7 shows that the effect of age is no longer significant when duration of diabetes is also in the model. Sex shows no significant effect on the process of early diabetic retinopathy. Duration and HbA1 are the two most important independent factors affecting the process of progression and regression of early diabetic retinopathy.

Table 2. Maximum likelihood estimates for the final model

Parameter	Estimate	Std. Error	Relative Risk
λ_{12}	0.0003	0.0004	–
λ_{21}	0.0173	0.0036	–
λ_{23}	0.0134	0.0036	–
λ_{32}	2.0540	2.0570	–
μ_3	0.0080	0.0047	–
$\beta_{12,HbA1}$	0.3044	0.1010	1.356
$\beta_{12,Duration}$	0.1386	0.0453	1.149
$\beta_{32,Duration}$	-0.2900	0.0825	0.748

Table 2 shows parameter estimates for a model which includes duration of diabetes and HbA1 as covariables. This model involves the only 3 regression parameters that are statistically significant in both the univariate and multivariate models 2, 4 and 6 for these two variables.

From this table it can be inferred that one unit of the absolute variation in HbA1 represents an increase of 35.6% on the transition rate from normal to grades II-III in early diabetes retinopathy. One additional year in duration of diabetes represents an increase of 15% in the transition rate from normal to grades II-III and a reduction of 25% for the chances of regression from grades IV–V to grades II-III.

4 Non-Homogeneous Markov Models

In the course of modelling a multi-state Markov process it is natural to raise the question of whether the transition intensities are constant over time. In some processes the transition parameters will be clearly time-dependent. The homogeneous model can be extended to the non-homogeneous case by assuming that the multi-state model has time-varying parameters.

Different parametric and semi-parametric functions of the time can be proposed to model the transition intensities. From the wide selection of parametric families of hazard functions one of the most popular is the Weibull hazard function, $\lambda(t) = p\lambda(\lambda t)^{p-1}$. This is often used in survival analysis, and the exponential hazard function ($p = 1$) is a special case. Splines and other smoothers are attractive semi-parametric approaches for describing the changes in the transition intensities over time. If the transition intensities are arbitrary functions of time, the Kolmogorov system of differential equations does not have a closed form solution and it becomes increasingly more difficult, if not impossible, to obtain a simple expression for the likelihood function in contrast with the homogeneous model. As was mentioned before, when the process has unidirectional transitions, a simple closed form solution for the transition probability matrix can be found by systematic integration.

A practical solution of this problem is to subdivide the period of follow-up into K intervals $[\tau_1, \tau_2), [\tau_2, \tau_3), \ldots, [\tau_K, \infty)$, and assume a constant intensity matrix in each interval. The intensity matrix can be written as $\mathbf{Q}(t) = \mathbf{Q}_l$, where $t \in [\tau_l, \tau_{l+1})$. A local solution of the Kolmogorov differential equations can then be found for each interval and a general solution can be obtained applying the Chapman-Kolmogorov equation as in expression (5).

Parametric forms for the transition intensities for this piecewise constant model are,

$$\lambda_{ijl} = \lambda_{ij} t_l^{\gamma_{ij}} \quad \text{and} \quad \mu_{il} = \mu_i t_l^{\gamma_{ik}}$$

or

$$\lambda_{ijl} = \lambda_{ij} e^{\gamma_{ij} t_l} \quad \text{and} \quad \mu_{il} = \mu_i e^{\gamma_{ik} t_l},$$

where t_l is the center point in the interval $[\tau_l, \tau_{l+1})$, and where λ_{ijl} and μ_{il} are the elements (i,j) and (i,k) of the intensity matrix $\mathbf{Q}_l$. The advantage of assuming a parametric form for the transition intensities is the dramatic reduction in the number of parameters involved in the model. In addition, a test for a constant transition intensity from i to j is equivalent to the test $H_0 : \gamma_{ij} = 0$. A global test for homogeneity of the process can be performed by doing a simultaneous test for all the γ's.

The joint distribution of $X(t_1), X(t_2), \dots, X(t_{m+1})$ given $X(t_0)$ can be represented as

$$\prod_{j=1}^{m+1} p_{i_{j-1},i_j}(t_{j-1}, t_j)$$

where $p_{ij}(s,t)$ is the element (i,j) of the resulting product of matrices in (5). Using the same argument as in section (3.5), this expression represents the contribution to the likelihood for this subject if the arrival time at the absorbing state is interval-censored. If t_{m+1} is the lifetime for this subject or the censuring time, then the contribution to the likelihood is equivalent to expression (9) and can be written as

$$L(\boldsymbol{\theta}) = \prod_{j=1}^{m} p_{i_{j-1},i_j}(t_{j-1}, t_j)\{f_{i_m}(t_m, t_{m+1})\}^{\delta}\{S_{i_m}(t_m, t_{m+1})\}^{1-\delta},$$

where

$$f_{i_m}(t_m, t_{m+1}) = \sum_{j=1}^{k-1} p_{i_m,j}(t_m, t_{m+1})\mu_j(t_{m+1})$$

and where $S_{i_m}(t_m, t_{m+1}) = 1 - p_{i_m,k}(t_m, t_{m+1})$. Quasi-Newton algorithms can be used to minimize $-2l(\boldsymbol{\theta})$ using only the likelihood function and finite differences to obtain approximations for the first two derivatives (Dennis and Schnabel [6]).

This model can be easily extended to include time-dependent covariables. The model for the transition intensities can be written as

$$\lambda_{ij}(t, \mathbf{z}(t)) = \lambda_{ij}(t) e^{\boldsymbol{\beta}_{ij}\mathbf{z}(t)} \text{ and } \mu_i(t, \mathbf{z}(t)) = \mu_i e^{\boldsymbol{\beta}_{ik}\mathbf{z}(t)}$$

where $\lambda_{ij}(t)$ is the baseline transition intensity from the state i to the state j, $\boldsymbol{\beta}_{ij}$ is a vector of regression coefficient, and $\mathbf{z}(t) = \mathbf{z}_l$ is the vector of covariables, constant for every value of t in $[\tau_l, \tau_{l+1})$. The baseline transition intensities $\lambda_{ij}(t)$ can be modelled using a parametric form. In that case the model can be written

$$\lambda_{ijl} = \lambda_{ij} t_l^{\gamma_{ij}} e^{\boldsymbol{\beta}_{ij}\mathbf{z}_l} \text{ and } \mu_{il} = \mu_i t_l^{\gamma_{ik}} e^{\boldsymbol{\beta}_{ik}\mathbf{z}_l}.$$

References

1. Chiang, C.L. (1968). Introduction to Stochastic Processes in Biostatistics. Wiley, New York.
2. Cox, D.R. and Miller, H.D. (1965). The Theory of Stochastic Processes. Methuen, London.
3. Cox, D.R. (1972). Regression models and life-tables (with discussion). Journal of the Royal Statistical Society, Series B **34**, 187-220.
4. Cox, D.R. and Oakes, D. (1984). Analysis of Survival Data. Chapman and Hall, London.
5. Davies, G.S. (1985) A Note on a Continuous-Time Markov Manpower Model. Journal of Applied Probability, **22**, 932–938.
6. Dennis, J.E., Jr., and Schnabel, R.B. (1983). Numerical Methods for Unconstrained Optimization and Nonlinear Equations. Prentice-Hall, Englewood Cliffs, New Jersey.
7. Gantmacher, F.R. (1959). The Theory of Matrices, vols 1–2, Chelsea New York.
8. Garg, S.K., Marshall, G., Chase, H.P., Jackson, W., Archer, P., Crews, M. (1990). The Use of the Markov Processes in Describing the Natural Course of Diabetic Retinopathy. Archives of Ophthalmology, **108**, 1245–1247.
9. Hochstadt, H. (1963). Differential Equations: A Modern Approach. Dover, New York.
10. Horn, R.A., Johnson, C.R. (1985). Matrix Analysis, Cambridge University Press, Cambridge.
11. Kalbfleisch, J.D., Lawless, J.F. and Vollmer, W.M. (1983). Estimation in Markov Models from Aggregate Data. Biometrics **39**, 907–919.
12. Kalbfleisch, J.D. and Lawless, J.F. (1985). The Analysis of Panel Data Under a Markov Assumption. Journal of the American Statistical Association **80**, 832–871.
13. Kay, R. (1986). A Markov Model for Analysing Cancer Markers and Disease States in Survival Studies. Biometrics **42**, 855–865.
14. Klein J.P., Klotz, J.H. and Grever, M.R. (1984). A Biological Marker Model for Predicting Disease Transitions. Biometrics **40** 927–936.
15. Longini Jr. I.M., Clark, W.S., Byers, R.H., Ward, J.W., Darrow, W.W., Lemp, G.F. and Hethcote, H.W. (1989). Statistical Analysis of the Stages of HIV Infection Using a Markov Model. Statistics in Medicine **8**, 831–843.
16. Madsen, H., Spliid, H., Thyregod, P. (1985). Markov Models in Discrete and Continuous Time for Hourly Observations of Cloud Cover. Journal of Climate and Applied Meteorology **24**, 629–639.
17. Raman, S. and Chiang, C.L. (1973). On a Solution of the Migration Process and the Application to a Problem in Epidemiology. Journal of Applied Probability **10**, 718–727.
18. Schoen, R. and Land, K.C. (1979). A General Algorithm for Estimating a Markov-Generated Increment-Decrement Life Table With Applications to Marital-Status Patterns. Journal of the American Statistical Association **74**, 761–776.
19. Wilkinson, J.H. (1965). The Algebraic Eigenvalue Problem. Clarendon Press, Oxford.