Semi-Markov Models with Phase-Type Sojourn Distributions by A.C. Titman, L.D. Sharples, 2010

Rahul Biswas

Department of Statistics, University of Washington Seattle, WA, 98195, USA

1 Introduction

Categorical response panel data, observed at unevenly spaced discrete time points are often encountered in practice, especially in the context of disease progression (Guihenneuc-Jouyaux et al., 2000; Mandel, 2010). Time-homogeneous continuous time Markov chains (CTMCs) are the simplest and most tractable in such scenario (Kalbfleisch and Lawless, 1985). However homogeneous CTMCs have limiting restrictions of transition intensities constant over time, and, sojourn distributions being exponential, which are often unrealistic.

Inhomogeneous CTMCs (Kay, 1986; Hubbard, Inoue, and Fann, 2008; Titman, 2011) extend the setup to have transition intensities vary with respect to time since the process origin. But for diseases, often transition intensities may depend on the time spent in the current state (sojourn time), not just external time. Semi-Markov models have such a property and are considered in this paper (Cox & Miller 1965; McGilchrist & Hills 1991).

Although semi-Markov models are appealing, there are computational hurdles documented in fitting them to data. The likelihood is less tractable for observed panel data unless the model is assumed to be progressive, where a subject cannot reenter a state once exited (Joly and Commenges, 1999; Foucher et al., 2010). In the presence of reversible transitions, semi-Markov models are tractable only under stringent restrictions of an evenly spaced two-state recurrent model (Rosychuk and Thompson; 2001), or, if at least one state has exponential sojourn distribution (Kang, Lagakos; 2007).

Crespi et al. (2005) recorded computational advantages in using a latent birth-death process of symptom recurrences to model a two state health-diseased process. The resulting two-state health-diseased process was semi-Markov. But the latent structure allowed the likelihood can be expressed to have the same form as a hidden Markov model (HMM), thereby enabling usage of well-developed computational techniques for HMM. Titman & Sharples in the current paper developed a general methodology for semi-Markov modelling and inference using such latent CTMC.

In contrast to the restriction of Exponential sojourn distributions in homogeneous CTMCs, the sojourn-time of observable states in such latent CTMC setup has a phase type distribution, defined as the distribution of absorption time of a homogeneous CTMC with finite transient states and one absorbing state. An advantage is generality, as phase type distributions are dense in the class of all distributions with non-negative support, so any distribution with non-negative support can be approximated by a phase-type distribution (Neuts, 1974). Analytic tractability is also ensured with density, cumulative distribution function and failure rate being matrix exponentials. One disadvantage is that the model parameters may not be identifiable (Asmussen et al., 1996), which is a difficulty in frequentist estimation, but, typical scientifically meaningful functionals of sojourn distribution parameters are identifiable (Bladt et. al, 2003). The latent CTMC parameters in this paper has been constrained to yield a subclass of phase type distributions called Coxian phase-type distribution (See Figure 1) for sojourn time. The Coxian subclass is often opted for, as it has been recorded to provide similar approximations to distributions compared to the general phase-type class in many experiments, while being the faster one for computation (Asmussen etl al, 1996).

In this paper, the authors discuss a general approach to fitting a Semi-Markov model with a latent CTMC and Coxian phase-type sojourn distribution to panel observed categorical response data. The model is extended to incorporate misclassification error. Methods for inference of parameters while addressing non-identifiability concerns are discussed. The methods are applied to assess development of bronchilitis obliterans syndrome in post-lung-transpantation patients, making comparison with the standard popular method based on HMM. The quantities of scientific interest studied are the rate of disease onset, survival rates of patients before and after disease onset given survived for certain years after onset, and, extent of misclassification, which are one-dimensional functionals of the model parameters.

2 Methods

2.1 Likelihood for Continuous-time processes with panel data

Let $X(t)$ denote the continous-time discrete state stochastic process on a finite state space $\{1, \ldots, R\}$. The observed data for an individual subject consists of observed states $x_0, \ldots, x_n$ at times (can be irregular) $t_0, \ldots, t_n$ where time points and number of times n can be subject specific. The data is panel observed at discrete time points with no information about the trajectory of the process between observations. It is assumed throughout, that, the sampling mechanism from the continuous process is non-informative (Gruger, Kay, Schumacher, 1991).

General finite state continuous-time models may be defined according to the transition intensities from states r to s

$$q_{rs}(t, \mathcal{F}_t) = \lim_{\delta t \downarrow 0} \frac{P(X(t+\delta t) = s | X(t) = r, \mathcal{F}_t)}{\delta t}$$

where $\mathcal{F}_t$ is the filtration or past history of the process up to time t.

Under a Markov assumption (Kalbfleisch and Lawless, 1985), the transition intensities are only a function of t and the likelihood for an individual is

$$L(\theta) = \prod_{i=1}^{n} p_{x_{i-1},x_i}(t_{i-1}, t_i; \theta)$$

where $p_{x_{i-1},x_i}(t_{i-1}, t_i; \theta) = P(X(t_i) = x_i | X(t_{i-1}) = x_{i-1})$ are the transition probabilities. The transition probabilities can be expressed in terms of the transition intensities by solving the Kolmogorov forward equations (Cox & Miller; 1965). If time homogeneous, they are related through a matrix exponential.

2.2 Semi-markov models

In a semi-Markov model the transition intensities depend on the time spent in the current state

$$q_{rs}(t, \mathcal{F}_t) = q_{rs}(u) = \lim_{\delta t \downarrow 0} \frac{P(X(t+\delta t) = s | X_i(t) = r, T^* = t - u)}{\delta t}$$

where T^* denotes the time since entry into the current state. In semi-Markov models the sojourn times in each state are not constrained to have exponential distributions in contrast to time homogeneous Markov models. Titman & Sharples specify the sojourn distribution of $X(t)$ to be of Coxian phase-type (See Section 2.3).

Computation of the likelihood for panel observed states is more difficult compared to the Markov case. In the case of progressive models, where the process cannot reenter a state once exited, there is a finite number of possible paths that an individual can take conditional on their observed states. Computation of the likelihood requires considering each path and integrating over the possible sojourn times in each state of the path (Foucher et al. 2010). Numerical quadrature methods can be applied to compute the likelihood but become unattractive for models with more than 3 or 4 states because of the increasing dimension of the integrals.

For more general models with reverse transitions, direct integration is not possible because the number of possible state visits is unbounded. Computation of the transition probabilities defined as $p_{r,s}(u,t) = P(X(t) = s|X(u) = r, T^* = u)$ requires solution to a system of integral equations (Howard 1964; De Dominics and Manca 1984). However, limitation for panel observed states is that the likelihood cannot be expressed as the product of transition probabilities, as the observation times will not correspond to the entry time into the observed state. Under restrictions of evenly-spaced two-state recurrent model (Rosychuk & Thompson, 2001), or, having one state with exponential sojourn distributions (Kang & Lagakos, 2007), computation could be made tractable.

2.3 Coxian Phase-type Models with misclassification error

Recall that $X(t)$ denotes a continuous time stochastic process with state space $\{1, \dots, R\}$, where t is time since process origin. Let R be the only absorbing state. Underlying $X(t)$ there is the latent homogeneous CTMC $X^*(t)$ with latent state space $\{1_1, \dots 1_{s_1}\} \cup \dots \cup \{(R-1)_1, \dots (R-1)_{s_{R-1}} \cup R\}$. $X(t) = r \Leftrightarrow X^*(t) \in \{r_1, \dots r_{s_r}\}$ for $r < R$ and $X(t) = R \Leftrightarrow X^*(t) = R$. The sojourn distribution of each non-absorbing state r of $X(t)$ is assumed to be a k-phase Coxian phase-type distribution, defined to be the distribution of absorption time of a CTMC with transition diagram in Figure 1. Thus for observable state r, the latent process $X^*(t)$ can have the states $r_1, \dots, r_k$, with parameters λ_{r_j} = the intensity for movement from r_j to r_{j+1}, for $j = 1, \dots, k-1$, with the assumption that these latent phases

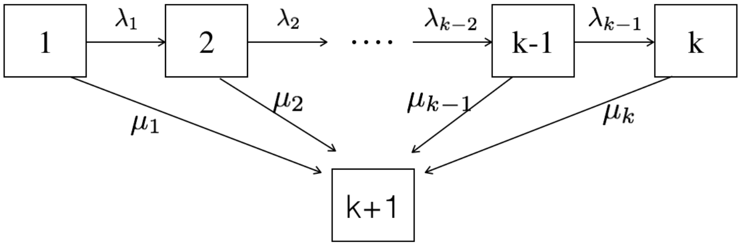
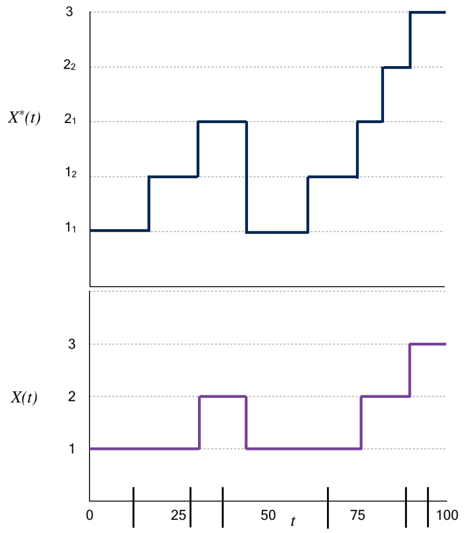


Figure 1: This is a transition diagram of a CTMC with $k+1$ states, $k+1$ being the only absorbing state and the remaining transient. A k-phase Coxian phase-type distribution is defined as the distribution of absorption time of a CTMC with above transition diagram.

for each observable state is progressive thereby all other transitions between r_j's have 0 intensity, and $\mu_{r_js} =$, the intensity for movement out of a latent phase of r_j to state s with the assumption that any state s is entered only at latent phase s_1, thereby all other intensities for transition from latent phases of r to latent phases of s are 0. Figure 1 summarizes the transition diagram for transitions between latent phases of observable state r and transition out to observable state s. For greater estimability and interpretation of moving out of a state to another it is assumed that $\mu_{r_js} = \tau_{r_j}\mu_{r_1s}\forall s$, $\tau_{r_1} = 1$. That is rates of exiting from latent phase r_j relative to r_1 change by the same factor irrespective of destination. It is noted that the resulting $X(t)$ is a semi-Markov process.

To incorporate that the states in the process are observed with misclassification error, it is considered that observed states are $O(t)$ which are related to $X(t)$ by taking a value with misclassification probability $e_{rs} = P(O(t) = s|X(t) = r) = P(O(t) = s|X^*(t) = r_j)$ for $r, s = 1, \ldots, R$, $j = 1, \ldots, k$. It is noted that the process $(O(t), X^*(t))$ is a Hidden Markov Model (HMM). Thereby computational methods tailored for HMM can be used for inference. Figure 2 shows an example of a disease process with latent trajectory $X^*(t)$ and the corresponding disease trajectory $X(t)$ and misclassified trajectory $O(t)$ for a two-state reversible disease model.



Observation times	t_1	t_2	t_3	t_4	t_5	t_6
$X^*(t)$	1_1	1_2	2_1	1_2	2_2	3
$X(t)$	1	1	2	1	2	3
$O(t)$	1	1	1	2	2	3

Figure 2: This is an illustration of a disease process with latent trajectory $X^*(t)$, corresponding disease trajectory $X(t)$, and misclassified trajectory $O(t)$. The three states of the disease process $X(t)$: {1=healthy, 2= BOS, 3=death} with 3 the only absorbing state. $X(t)$ has underlying latent homogeneous CTMC $X^*(t)$ with state space $S^* = \{1_1, 1_2, 2_1, 2_2, 3\}$ with $\{3\}$ an absorbing state and rest transient. $O(t)$ is the observed process which is the misclassified version of $X(t)$. States which are misclassified in this example are marked in red.

2.4 Inference and Identifiability

Denoting $o_0, \ldots, o_n$ to be the sampled observations at times $t_0, \ldots, t_n$ for an individual, the latent HMM structure of the model enables to write the likelihood contribution for the individual based on the forward-backward algorithm, as follows.

$$L = \pi \boldsymbol{M}_1^* \boldsymbol{M}_2^* \ldots \boldsymbol{M}_n^* \mathbf{1} \tag{1}$$

where $\boldsymbol{M}_i^*$ is a $\{k(R-1)+1\} \times \{k(R-1)+1\}$ matrix with (r^*, s^*) entry $e^*_{s^* o_i} p^*_{r^* s^*}(t_i - t_{i-1})$ for $r^*, s^* \in S^*$, with $P^*(t) = ((p^*_{r^* s^*})) = e^{\boldsymbol{\Lambda} t}$, and, π is the initial state distribution.

The presence of misclassification and HMM setup increases possibility of non-identifiability (MacDonald & Zucchini, 1997). In fact for a HMM with two states and balanced pbservation times, it is not possible to simultaneously identify the misclassification probabilities and transition intensities or probabilities without putting constraints, for instance that misclassification probabilities are each < 0.5 (Rosychuk & Thompson, 2003).

In addition, the Coxian phase-type structure induces non-identifiability. For example, the sojourn time distribution for state r is exponential if $\tau_{r_j} = 1 \forall j = 1, \ldots, k$. The parameters, $\lambda_{r_j}, j = 1, ..., k - 1$, are redundant and unidentifiable in this situation.

As ensuring identifiability is not straightforward, the authors suggest inspection of the likelihood function by evaluating the Hessian matrix to ensure the estimated parameters are at a maxima and performing optimization procedures from a wide range of starting values to ensure a strict global maxima has been attained, or, checking for unimodality of the likelihood. If non-identifiability is evident, reduction of the complexity of the model can be attempted. Else, the parameters can be constrained. The authors suggest performing the optimization by first maximizing the profile likelihood on a grid of possible fixed values of the λ_{r_j}. Other possible constraints suggested are having some λ_{r_j} as known constants, or constraining some of the sojourn distributions to be exponential.

Due to identifiability problems for exponential sojourn distributions in the current model, the likelihood ratio test for Markov versus Coxian phase-type semi-Markov model may not have an asymptotic χ^2 distribution. Chen et. al (2007) proposed a penalized likelihood ratio tests for homogeneity in finite mixture models. Authors considered the same method but with an appropriate penalty for the current situation so that the λ_{r_j}'s are away from 0 and ∞ and identifiable irrespective of value of τ.

Titman & Sharples (2008) proposed a goodness of fit test applicable for Markov and Hidden Markov Models. The same is used for testing gooness of fit in current model which boils down to a HMM.

3 Description of data

4 Results

4.1 Incorporating covariates