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, particularly in the context of disease processes (Guihenneuc-Jouyaux et al., 2000; Mandel, 2010). Continuous time stochastic processes form a lucid model in such scenario. Homogeneous continuous time Markov chain (CTMC) are simple and tractable for panel data (Kalbfleisch and Lawless, 1985), but they have limiting restrictions of transition intensities constant over time, and, sojourn distributions are 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 as models, there are computational hurdles recorded in fitting them. The likelihood is recorded as less tractable for panel observed 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, it is shown to be tractable 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 homogeneous CTMC in a two-state healthy-diseased recurrent model. With state space $\{0, 1, 2, \ldots\}$, a subject was considered to be healthy if in state 0 and ill otherwise. The resulting model is semi-Markov with transition intensities depending on time since entry into a state. 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. This method of modelling with latent states has a long history (Cox, 1955a).

In contrast to the restriction of Exponential sojourn distribution in homogeneous CTMCs, the sojourn-time for such latent CTMC has a phase type distribution. An advantage is generality, as phase type distributions are dense in the class of all distributions with non-negative support. Analytic tractability is also ensured with density, cumulative distribution function and failure rate being matrix exponentials (Neuts, 1974). One disadvantage is that the model parameters may not be identifiable (Asmussen etl 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 distribution called Coxian phase-type distribution 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 of 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.