# 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, 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 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.

# References

Asmussen, S., Nerman, O., and Olsson, M. (1996). Fitting phase-type distributions via the em algorithm. *Scandinavian Journal of Statistics*, pages 419–441.

Bladt, M., Gonzalez, A., and Lauritzen, S. L. (2003). The estimation of phase-type related functionals using markov chain monte carlo methods. *Scandinavian Actuarial Journal*, 2003(4):280–300.

Cox, D. and Miller, H. (1965). The theory of stochastic processes, methuen & co. *Ltd, London, UK*.

Cox, D. R. (1955). The analysis of non-markovian stochastic processes by the inclusion of supplementary variables. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 51, pages 433–441. Cambridge Univ Press.

Crespi, C. M., Cumberland, W. G., and Blower, S. (2005). A queueing model for chronic recurrent conditions under panel observation. *Biometrics*, 61(1):193–198.

Foucher, Y., Giral, M., Soulillou, J., and Daures, J. (2010). A flexible semi-markov model for interval-censored data and goodness-of-fit testing. *Statistical methods in medical research*, 19(2):127–145.

Guihenneuc-Jouyaux, C., Richardson, S., and Longini, I. M. (2000). Modeling markers of disease progression by a hidden markov process: application to characterizing cd4 cell decline. *Biometrics*, 56(3):733–741.

Hubbard, R. A., Inoue, L., and Fann, J. R. (2008). Modeling nonhomogeneous markov processes via time transformation. *Biometrics*, 64(3):843–850.

Joly, P. and Commenges, D. (1999). A penalized likelihood approach for a progressive three-state model with censored and truncated data: Application to aids. *Biometrics*, 55(3):887–890.

Kalbfleisch, J. and Lawless, J. F. (1985). The analysis of panel data under a markov assumption. *Journal of the American Statistical Association*, 80(392):863–871.

Kang, M. and Lagakos, S. W. (2007). Statistical methods for panel data from a semi-markov process, with application to hpv. *Biostatistics*, 8(2):252–264.

Kay, R. (1986). A markov model for analysing cancer markers and disease states in survival studies. *Biometrics*, pages 855–865.

Mandel, M. (2010). Estimating disease progression using panel data. *Biostatistics*, page kxp057.

McGlLCHRIST, C. and HlLLS, L. (1991). A semi-markov model for ear infection. *Australian & New Zealand Journal of Statistics*, 33(1):5–16.

Neuts, M. F. (1974). *Probability distributions of phase type*. Purdue University. Department of Statistics.

Rosychuk, R. J. and Thompson, M. E. (2001). A semi-markov model for binary longitudinal responses subject to misclassification. *Canadian Journal of Statistics*, 29(3):395–404.

Titman, A. C. (2011). Flexible nonhomogeneous markov models for panel observed data. *Biometrics*, 67(3):780–787.