Matias Quiroz and Mattias Villani
Division of Statistics and Machine Learning
Dept. of Computer and Information Science
Linköping University

November 13, 2016

Master's thesis proposal

Modeling the cause of censoring in discrete-time survival data

Censored data, i.e. when measurements of subjects are only partially known, are widely occurring in Machine Learning. When the measurement is the time until the occurrence of an event of interest (e.g. death, cure, marriage), censored data are often called survival data. In many applications time is measured discretely (e.g. days, months, weeks) and thus discrete-time survival models are suitable (Singer and Willett, 1993). The likelihood for such models is expressed through a censoring variable, which identifies the subjects we only have partial information for. Traditional models assume that the censoring variable is independent of the survival time (given relevant covariates).

This project aims at developing flexible models for the censoring variable as a function of covariates and, moreover, model its dependence on the survival time. This is often called informative censoring, see e.g. Scharfstein et al. (2001). Modeling the censoring variable as a function of covariates gives useful insights on the characteristics of censored subjects: for example, what factors are determining that only partial information is available? Moreover, by taking the informative censoring into account we omit possible bias in the parameter estimates.

Examples of flexible models that can be of interest include (but are not limited to): finite mixture models, hierarchical models, non-linear spline models or even non-parametric models such as Gaussian process or Dirichlet process mixtures. The models will be estimated via state-of-the-art Markov chain Monte Carlo with Bayesian variable selection (Villani et al., 2012; Quiroz and Villani, 2013). The variable selection gives insights about importance of covariates in different parts of the model and can thus be useful to determine the characteristics of censored subjects.

This project will give you excellent training in writing programming code for efficient estimation of flexible models. You will learn a great deal about generic MCMC samplers, which will allow you to estimate complex statistical models.

References

Quiroz, M. and Villani, M. (2013). Dynamic mixture-of-experts models for longitudinal and discrete-time survival data. *Riksbank Research Paper Series*, (99).

Scharfstein, D., Robins, J. M., Eddings, W., and Rotnitzky, A. (2001). Inference in randomized studies with informative censoring and discrete time-to-event endpoints. *Biometrics*, 57(2):404–413.

Singer, J. D. and Willett, J. B. (1993). It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational and Behavioral Statistics*, 18(2):155–195.

Villani, M., Kohn, R., and Nott, D. J. (2012). Generalized smooth finite mixtures. *Journal of Econometrics*, 171(2):121–133.