Lights: a generalized joint model for high-dimensional multivariate longitudinal data and censored durations

Simon Bussy*[1], Antoine Barbieri[2], Sarah Zohar[1], and Anne-Sophie Jannot[1,3]

[1] *INSERM, UMRS 1138, Centre de Recherche des Cordeliers, Paris, France*
[2] *INSERM, UMR 1219, Bordeaux Population Health Research Center, Univ. Bordeaux, France*
[3] *Biomedical Informatics and Public Health Department, EGPH, APHP, Paris, France*

Abstract

In many clinical studies, it has become increasingly common to record the values of longitudinal covariates until the occurrence of an event of interest. With the development of electronic health records and an increased focus on personalized medicine, the need to implement multivariate models that account for a large number of longitudinal ~~outcomes~~ simultaneously is critical. Despite this, the state-of-the-art methods in this context have predominantly focused on univariate data, or on problems where the number of longitudinal ~~outcomes~~ are very low. Our paper introduces a prognostic method called lights (generaLized joInt hiGH-dimensional longiTudinal Survival) to deal with the problem of joint modeling of longitudinal data and censored durations, in a high-dimensional context. The later introduces a latent variable modeling the heterogeneity within the patient population, with subgroups ~~of~~ different risk, and supposes a group-specific marker trajectory with a generalized linear mixed model for each longitudinal marker given the subgroup, and a group-specific Cox risk of event with multiple shared association defined through a known functional family. Inference is achieved using a novel fast stochastic approximation of a quasi-newton EM algorithm by minimizing the negative log-likelihood penalized with elastic-net or group lasso regularization on the different parameter vectors of the model, depending on the desired interpretability power. The estimated latent class membership posterior probabilities are used as discriminative marker rule in the cross-validation procedure for selecting the best regularization hyper-parameters. The statistical performance of the method is examined on an extensive Monte Carlo simulation study, and finally illustrated on a publicly available dataset. On this high-dimensional dataset, our proposed method outperforms the state-of-the-art models regarding risk prediction in terms of C-index, with a computing time orders of magnitude faster. In addition, it provides powerful interpretability by automatically pinpointing significant covariates being relevant from a clinical perspective. Thus, we propose a powerfull tool for personalized medicine, with the ability of automatically determining significant prognostic longitudinal biomarkers, which is of increasing importance in many areas of medicine.

*Corresponding author: simon.bussy@gmail.com