High-dimensional estimation of counting process intensities

Sarah Lemler

Abstract

We consider the problem of obtaining a prognostic on the survival time adjusted on high-dimensional covariates. Towards this end, we consider two different approaches.

First, we propose the construction of an estimator of the general conditional intensity function. We estimate it by the best Cox proportional hazards model given two dictionaries of functions. The first dictionary is used to construct an approximation of the logarithm of the baseline hazard function and the second to approximate the relative risk. As we are in high-dimension, we consider the Lasso procedure to estimate the unknown parameters of the best Cox model approximating the conditional intensity. We provide non-asymptotic oracle inequalities for the Lasso estimato|r of the conditional intensity. Our results rely on an empirical Bernstein's inequality for martingales with jumps.

Then, in a second part, we consider an intensity that rely on the Cox model. In this Cox model, two parameters are unknown : the baseline function, function of the time and the regression parameter connected to the covariates. We propose a two-step procedure to estimate the intensity : a Lasso procedure to estimate the regress|ion parameter in high-dimension and then, a model selection procedure to estimate the baseline function. We establish a non-asymptotic oracle inequality on the baseline function.