High dimensional data reduction in risk and survival data analysis

Catherine Huber

University Paris Descartes
E-mail: catherine.huber.carol@gmail.com

Abstract: Risk analysis is a topic of increasing importance in multiple fields like environment, technology and biomedicine.

In survival data analysis and reliability, one is interested in all risk factors that may accelerate or decelerate the life length of individuals or machines.

Now, as immense data bases (big data) are available, several types of methods are needed to deal with the resulting curse of dimensionality: on one hand, methods that reduce the dimension while maximizing the infor-mation left in the reduced data, and then applying classical statistical models; on the other hand algorithms that apply directly to big data, i.e. artificial intelligence (machine learning), at the cost of a difficulty of interpretation in terms of the risk factors. Actually, those algorithms have a probabilistic interpretation. However, being often very good performers for prediction purposes, they lack explanatory interpretation.

We present here several methods for reducing the dimensionality of the data while maximizing the information relevant for the objective of the study, still present in the reduced data.