

Sparse Cox Regression for High Dimensional Data

Dr. L. Kaderali

University of Heidelberg

Viroquant Research Group Modeling

It has previously been demonstrated that gene expression data correlate with event-free and overall survival in several cancers. A number of methods exists that assign patients to different risk classes based on expression profiles of their tumor. However, predictions of actual survival times in years for the individual patient, together with confidence intervals on the predictions made, would provide a far more detailed view, and could aid the clinician considerably in evaluating different treatment options. Furthermore, such a method could be analyzed to infer knowledge about the relevant disease genes, hinting at potential disease pathways and pointing to relevant targets for drug design. Here too, confidences on the relevance values for the individual genes would be useful to have.

To tackle these questions, we combine a Cox regression model with a Bayesian learning approach. Using a strong sparsity prior, the method can efficiently deal with the low sample number, high dimensionality setting characteristic of microarray datasets. Using MCMC sampling, the approach taken allows us to make statements such as "with probability 0.6, the patient will survive between 3 and 4 years". Similarly, relevance values with confidence intervals for the individual genes are computed.

We evaluated the method on simulated data and applied the algorithm to several publicly available and proprietary datasets. In our presentation, we will present the theoretic framework underlying the approach, show selected results, and discuss open questions and further work.