Two-stage variable selection methods for predicting the survival time of lung cancer

Seungyeoun Lee[1] and Youngchul Kim[2]

[1]Department of Applied Mathematics, Sejong University, Seoul 143-747, Korea, [2]Department of Statistics, Seoul National University, Seoul 151-747, Korea

In this paper, we consider the two-stage variable selection methods in the Cox model when a large number of gene expression levels are involved with survival time. Deciding which genes are associated with survival time has been a challenging problem because of the large number of genes and relatively small sample size (n << p). Several methods for variable selection have been proposed in the Cox model. Among those, we consider least absolute shrinkage and selection operator (LASSO), threshold gradient descent regularization (TGDR), and two different clustering threshold gradient descent regularization (CTGDR)—the *K*-means CTGDR and the hierarchical CTGDR, which have been proposed by Ma and Huang (2007). These methods are evaluated by the approach of Ma and Huang (2007). When we apply these methods to a data set of lung cancer by Bhattacharjee et al. (2001), none of the methods shows satisfactory performance in separating the two risk groups using the log-rank statistic based on the risk scores calculated from the selected genes. For identifying susceptible genes, we propose the two-stage variable selection procedure. In the first stage, we apply the shrinkage methods or regularization methods for selecting genes which have non-zero coefficients but need to be tested for their significance. In the second stage, the Cox's model is fitted with the selected genes in the first stage and the susceptible genes are selected by testing their significance from the fitted Cox's model. The risk score is calculated from the genes that are shown to be significant in the Cox model and the performance of log-rank statistics shows that the two risk groups are well separated. The proposed method is illustrated with a dataset of lung cancer and four different methods from the first stage are compared by the performance of the log rank tests.