Thesis of Comparison of Sequential Lasso Method and Grouped Lasso Method For Feature Selection in Interaction Models

Many methods have been developed for feature selection in the literature such as penalized likelihood approaches and sequential methods, etc. However, the focus of these methods is mainly on additive models without interaction terms. Feature selection with interaction models raises additional issues which cannot be tackled directly applying the methods for additive models.

Two different methods have been recently proposed for feature selection in interaction models. One is extended Lasso approach with group penalties on the coefficients of the main effect and interaction terms such that a certain hierarchical structure of the effect is imposed, which is referred to as Extended GLasso approach in this paper. Another one is an extended sequential Lasso approach which treats main effect and interaction terms differently by the EBIC for interactive models, which is referred to as Extended SLasso approach in this paper. The aim of this paper is to study their practical advantages and disadvantages under high dimensional settings in interaction models.

In this paper, there is a brief summary of the literature on high dimensional feature selection procedures and selection criteria. A description of the algorithms of the Extended SLasso method and Extended GLasso method are included in detail. Their practical advantages and disadvantages under high dimensional settings in interaction models are studied via simulation in R. R code of the Extended SLasso method and different data settings using two types of models are developed. The comparison of their performance in terms of their PDR and FDR are made with analyses. Comprehensive simulation of different scenarios are carried out. The simulation results indicate better overall performance of the Extended SLasso due to its significantly larger PDR and acceptably small FDR. The Extended SLasso is considered having significant advantage over Extended GLasso as far as selection consistency and computation difficulty are concerned.