Ranking-Based Variable Selection for high-dimensional data

Abstract

We propose Ranking-Based Variable Selection (RBVS), a technique aiming to identify important variables influencing the response in high-dimensional data. The RBVS algorithm uses subsampling to identify the set of covariates which non-spuriously appears at the top of a chosen variable ranking. We study the conditions under which such set is unique and show that it can be successfully recovered from the data by our procedure. Unlike many existing high-dimensional variable selection techniques, within all the relevant variables, RBVS distinguishes between the important and unimportant variables, and aims to recover only the important ones. Moreover, RBVS does not require any model restrictions on the relationship between the response and covariates, it is therefore widely applicable, both in a parametric and non-parametric context. We illustrate its good practical performance in a comparative simulation study. The RBVS algorithm is implemented in the publicly available R package rbvs.