High-dimensional variable selection via tilting

Haeran Cho

London School of Economics

16th November 2011

Abstract

This paper considers variable selection in linear regression models where the number of covariates is possibly much larger than the number of observations. High dimensionality of the data brings in many complications, such as (possibly spurious) high correlations among the variables, which result in marginal correlation being unreliable as a measure of association between the variables and the response.

We propose a new way of measuring the contribution of each variable to the response which takes into account high correlations among the variables in a data-driven way. The proposed tilting procedure provides an adaptive choice between the use of marginal correlation and tilted correlation for each variable, where the choice is made depending on the values of the hard-thresholded sample correlation of the design matrix. We study the conditions under which this measure can successfully discriminate between the relevant and the irrelevant variables and thus be used as a tool for variable selection. Finally, an iterative variable screening algorithm is constructed to exploit the theoretical properties of tilted correlation, and its good practical performance is demonstrated in a comparative simulation study.