Variable selection and estimation in generalized linear models with the seamless L_0 penalty

Xihong Lin

Department of Biostatistics
Harvard University

Abstract

We propose variable selection and estimation in generalized linear models using the seamless (SELO) penalized likelihood approach. The SELO penalty is a smooth function that very closely resembles the discontinuous L0 penalty. We develop an efficient algorithm to fit the model, and show that the SELO-GLM procedure has the oracle property in the presence of a diverging number of variables. We propose a Bayesian Information Criterion (BIC) to select the tuning parameter. We show that under some regularity conditions, the proposed SELO-GLM/BIC procedure consistently selects the true model. We perform simulation studies to evaluate the finite sample performance of the proposed methods. Our simulation studies show that the proposed SELO-GLM procedure has a better finite sample performance than several existing methods, especially when the number of variables is large and the signals are weak. We apply the SELO-GLM to analyze a breast cancer genetic dataset to identify the SNPs that are associated with breast cancer risk.