Penalized likelihood approach to variable selection for Cox's regression model under nested case-control sampling

Jie-Huei Wang[1]* (王价輝), Chun-Hao Pan[2], I-Shou Chang[3], Chao Agnes Hsiung[3] and Yi-Ching Wang[4]

[1] Institute of Statistics, National Tsing Hua University

[2] Department of Mathematics, National Central University

[3] Population Health Sciences, National Health Research Institutes

[4] Department of Pharmacology, National Cheng Kung University

Abstract

Assuming Cox's regression model, we consider penalized likelihood approaches to conduct variable selection under nested case-control sampling or case-cohort sampling, which are useful in the current study of genomic epidemiology. Penalized non-parametric maximum likelihood estimate (PNPMLE) are characterized by self-consistency equations derived from score functions, which form the basis of the algorithm to compute PNPMLE. A cross-validation method is used to choose the tuning parameter within a family of penalty function. Simulation studies indicate that the numerical performance of PNPMLE is satisfactory and that LASSO performs best when cohort size is small and SCAD performs best when cohort size is large and may eventually perform as well as the oracle estimator, resembling the findings when *i.i.d.* sampling is considered. This method is also illustrated to look for DNA methylation biomarkers. For SCAD, consistency, asymptotic normality and oracle properties of the PNPMLE, the sparsity property of the penalty, and a consistent estimate of the asymptotic variance, based on observed profile likelihood, are established.

Keywords: nested case-control sampling, oracle property, penalized maximum likelihood estimate, profile likelihood, SCAD, variable selection