Information Synthesis and Variable Selection Using A Penalized Empirical Likelihood Approach

Ying Sheng

University of California at San Francisco
E-mail: ying.sheng@ucsf.edu

Abstract: The aggregate data from large databases has become increasingly available, producing auxiliary information that can be incorporated to regulate the analysis of smaller-scale studies which collect more comprehensive individual-level information. Properly incorporating such information is anticipated to improve accuracy and efficiency but is technically challenging. Auxiliary information is usually at aggregate level and can be available via different statistical quantities. In this paper, we propose a unified approach to synthesize auxiliary information in the logistic regression analysis of individual-level data with a diverging number of parameters. The proposed approach summarizes various types of auxiliary information via a system of nonlinear population estimating equations and incorporates it in variable selection and model estimation using a novel penalized empirical likelihood method. We further extend the approach to account for potential inconsistency between the auxiliary information and study subjects, and uncertainties in the auxiliary information. The resulting estimator possesses the oracle property and is asymptotically more efficient than the usual penalized maximum likelihood estimator.

Simulation studies show that the proposed approach performs better with higher efficiency for model estimation and higher accuracy for variable selection. We applied the proposed approach to a pediatric kidney transplant study analysis for illustration.