Applying Distributed Sequential Estimation Method to Analyze Huge Data

Yuan-chin Ivan Chang

Academia Sinica

Abstract

When analyzing a data set with huge sample size and lengthy variables, the computational issue is not ignorable. Computer scientists may want to resolve such kinds of issues from algorithm and hardware perspectives, which usually require a complicated software setup and/or modern computational facility. In this paper, we adopt a divide-and-conquer idea and propose a parallel sequential method for analyzing data sets with huge sizes. In addition, we adopt an adaptive sample selection, with a criterion from statistical experimental design, and an adaptive shrinkage estimation method to simultaneously accelerate estimating procedure and identify the effective variables. We then apply the proposed method to three real examples including energy usage of appliances and the particulate matter (PM2.5) concentration from two major cities in China.

Co-author: Zhanfeng Wang, Department of Statistics and Finance, University of Science and Technology of China, Hefei, China