Model-based Outlier Detection in Multivariate Data with Applications to Detecting Cheating in Tests

Yunxiao Chen

London School of Economics and Political Science
E-mail: y.chen186@lse.ac.uk

Abstract: In this talk, we introduce a statistical framework for the detection of outliers in multivariate data, where outliers are defined as observations (rows) and manifest variables (columns) which deviate from a pre-specified factor model. The outliers are modeled by a factor model component with sparse structures in both the observations and the manifest variables. This problem is motivated by an application to cheating detection in education, where an observation is an examinee and a manifest variable is a test item. The outliers correspond to the cheating examinees and leaked test items, where the cheating examinees have cheated in the exam on the leaked items.

A constrained joint likelihood estimator is proposed that detects row- and column outliers by group truncated L1 constraints. Consistency and rate of convergence are established for this estimator. A DC (difference of convex functions) programming algorithm is developed for the computation of this estimator. The proposed method is applied to an educational testing data set and successfully recovers the cheating examinees and leaked items that have been flagged by the testing program. (This is a joint work with Dr. Xiaoou Li and Mr. Haoran Zhang.)