High-dimensional Categorical Data Clustering with Applications to Reliability Analysis

Jen Tang

Krannert School of Management

Purdue University

Abstract

In this paper, we propose a clustering algorithm for a class of high-dimensional data matrices with categorical attributes (columns) and a cluster structure in subjects (rows). Subjects are in the same cluster if they have the same categorical values for all categorical attributes. The entries in the observed data matrix, which our algorithm uses to cluster subjects, are assumed to be noisy due to random flipping and erasure (missing). We show that, by observing only a small fraction of noisy entries in the given data matrix, the proposed algorithm exactly recovers the underlying subject clusters with high probability; i.e., with probability one asymptotically. We also study the complexity and the finite-sample performance of the algorithm, based on simulated and real data. Application to a real dataset in reliability is also given.

This is joint work with Zhiyi Tian and Jiaming Xu.