A Geometric Algorithm for Contrastive Principal Component Analysis in High Dimension

Shao-Hsuan Wang

Graduate Institute of Statistics, National Central University

Abstract: Principal component analysis (PCA) has been widely used in exploratory data analysis. Contrastive PCA (Abid et al., 2018), a generalized method of PCA, is a new tool used to capture features of a target dataset relative to a background dataset while preserving the maximum amount of information contained in the data. With high dimensional data, contrastive PCA becomes impractical due to its high computational requirement of forming the contrastive covariance matrix and associated eigenvalue decomposition for extracting leading components. In this work, we propose a geometric curvilinear-search method to solve this problem and provide a convergence analysis. Our approach offers significant computational efficiencies. Specifically, it reduces the time complexity from O((n∨m)p^2) to a more manageable O((n∨m)pr), where n, m are the sample sizes of the target data and background data, respectively, p is the data dimension and r is the number of leading components. Additionally, we streamline the space complexity from O(p^2), necessary for storing the contrastive covariance matrix, to a more economical O((n∨m)p), sufficient for storing the data alone. Numerical examples are presented to show the merits of the proposed algorithm.