Principal Component Analysis for Large Distributed and Continuously Updating Data Sets

Zheng-Jian Bai* Raymond H. Chan* Franklin T. Luk†

Abstract

Identifying the patterns or highlighting the similarities and differences of large data sets is often needed in data mining. Principal component analysis (PCA) is a common technique for finding patterns in data of high dimension. It computes a set of eigenvectors corresponding to the dominant eigenvalues of the covariance matrix generated by the data.

Clustering algorithms based on PCA can be used effectively when data sets are centrally located. However, in practice, large data sets are distributed over a network clusters or on a data grid at different locations, and it is expensive or eventually impossible to centralize too large data sets. Here we introduce a new method for computing PCA of large data sets distributed in different locations without moving the data sets to a central location. Continuous update can also be accommodated without centralizing them first.

*(zjbai, rchan@math.cuhk.edu.hk) Department of Mathematics, Chinese University of Hong Kong, Shatin, NT, Hong Kong. The research of the second author was partially supported by the Hong Kong Research Grant Council Grant CUHK4243/01P and CUHK DAG 2060220.

†(luk@cs.rpi.edu) Department of Computer Science, Rensselaer Polytechnic Institute, Troy, New York 12180, USA.