Modified Gene Shaving Algorithm – A Dimension Reduction and Clustering Method

Donna Mae S. Raymundo

A thesis submitted as partial fulfilment of

the requirements for the degree of

MASTER OF SCIENCE IN STATISTICS

School of Statistics

University of the Philippines

Diliman, Quezon City

May 2017

Abstract

High dimensional data exist in digital images, financial time series and gene expression microarrays. Dealing with high dimensionality has become a challenge, where the difficulty lies on how to visualize and explore the high dimensional function or data set. Gene shaving is a statistical method which is based on Principal Component Analysis (PCA) that has proven its worth in visualization and exploration of microarray data. In this paper, the gene shaving algorithm was modified using PCA and Sparse Principal Component Analysis (SPCA), and the modified algorithms were explored in terms dimension reduction and clustering of variables in a more general (not necessarily microarray) high dimensional data setting. The proposed algorithms were evaluated through a simulation study. Simulation results suggest that the modified algorithms identify a singly cluster of variables that may already best explain the variations in the entire data and/or that already are the most informative. Also, the algorithms may produce overlapping clusters, whose variables in the succeeding clusters (other than the first cluster) are those that may provide information not inherent to the first cluster. The modified algorithms are thus potential and useful for exploration and identification of a group of variables worth for further investigation, as well as clustering/groups of variables for understanding variable structures/relationships.

Keywords: High dimensional data, Cluster of Variables, Gene Shaving, Principal Component Analysis, Sparse Principal Component Analysis, Dimension Reduction