Outlier Detection in High Dimensional Data
In the context of Clustering

By

Jenniebie C. Salagubang

A thesis submitted in partial fulfillment

of the requirements for the degree of

Master of Science in Statistics

School of Statistics

University of the Philippines

Diliman, Quezon City

March 2011

Abstract

As the storage size of databases increases, datasets we extract are usually confounded with two common problems: high dimensionality and outliers. We proposed an algorithm based on the forward search algorithm in both the principal components analysis and cluster analysis to identify outliers in a high dimensional data. The simulation study confirmed the viability of identifying outliers in high dimensional data thru the complementation of Principal Components Analysis and Cluster Analysis implemented via the forward search algorithm.

Keywords: high dimensional data, outliers, forward search algorithm, principal components analysis, cluster analysis.