Detection of Influential Observations in a Regression Analysis with High Dimensional Data

by

Maria Regina C. Garcia

A thesis submitted in partial fulfillment of the
requirements for the degree of

Master of Science (Statistics)

School of Statistics
University of the Philippines
Diliman, Quezon City

October 2010

ABSTRACT

There is a growing availability of high dimensional data that is often characterized by heterogeneity with some extreme values. High dimensionality often yield serious repercussions in fitting a regression model for one, the number of independent variables can exceed the sample size. Also, when there is a cluster of influential observations most methods of identification of influential observations suffer from swamping effect and masking effect.

We propose a method that first extracts principal components of the predictors to reduce dimensionality. Then the regression model with principal components as predictors is estimated through the forward search algorithm to simultaneously identify influential observations. The simulation study yield evidence that the algorithm can identify influential observations and is not affected by swamping and masking effects usually encountered in high dimensional data.

Keywords: high dimension data, principal component analysis, influential observation, forward search algorithm.