Robust Partial Least Squares Logistic Regression For High Dimensional Data

Nedret Billor and Asuman Seda Turkmen
Department of Mathematics and Statistics
221 Parker Hall, Auburn University, AL 36849

Abstract

The large number of genes (p) compared to the tissue sample size (n) causes to be many of the statistical modeling approaches inappropriate and therefore efficient methods for dimension reduction and information extraction are of great interests. Partial Least Squares (PLS) is a highly efficient statistical regression technique that is well suited for the analysis of such high-dimensional data. This method searches for a set of components (latent vectors) that performs a simultaneous decomposition of X (an $n \times p$ design matrix) and Y (the response variable vector) with the constraint that these components explain as much as possible of the covariance between X and Y. This is an advantage of PLS over principal component regression (PCR) that derives the latent vectors without reference to the response variable, Y. Originally PLS methods are proposed for continuous response and homogenous data (i.e., free of outliers). In this paper we investigate the problems that arise in the PLS regression technique when the response variable Y is binary and the problems of high dimensionality and outliers are present. In addition we propose a robust PLS algorithm that (i)reduces dimension by a robust procedure, and (ii)performs robust logistic regression (RLR) of response on the latent vectors obtained at the dimension reduction step. Real and simulated data sets are employed to compare performances of the methods based on PCR and PLS.

Key words and phrases: Dimension reduction; High dimensional data; Logistic regression; Partial least squares; Robustness.