World Academy of Science, Engineering and Technology
International Journal of Mathematical and Computational Sciences
Vol:18, No:06, 2024

Partial Least Square Regression for High-Dimensional and High-Correlated Data

Authors : Mohammed Abdullah Alshahrani

Abstract : This research focuses on the investigation of partial least squares (PLS) methodology to deal with high-dimensional correlated data. Current developments in technology have enabled experiments to produce data that are characterized by, first, the number of variables that far exceeds the number of observations and, second, variables that are substantially correlated between them. These types of data are commonly found in, first, chemometrics, where absorbance levels of chemical samples are recorded across hundreds of wavelengths in a calibration of a near-infrared (NIR) spectrometer. Second, they are also common to be found in genomics where copy number alterations (CNA) are recorded across thousands of genomic regions from cancer patients. In our study, we investigated key areas to address these challenges. Firstly, we tackled the issue of three main PLS algorithms having potentially different interpretations of relevant quantities. We unified these interpretations by identifying scenarios where all three algorithms yield the same estimates. Secondly, we explored the phenomenon of unusual negative shrinkage factors encountered during PLS model fitting. Unlike ridge regression or principal component regression, where shrinkage factors range between zero and one, PLS can exhibit factors greater than one or even negative, hence more aptly termed 'filter factors' rather than 'shrinkage factors'. This characteristic allows PLS to effectively handle high-dimensional data by applying shrinkage to estimates. To our knowledge, there has been no previous meaningful investigation on the negative filter factors (NFF) in PLS. In this research we present a novel result whereby we identify the condition for NFF to happen and investigate characteristics of the data that are associated with NFF to get an insight. Lastly, the main challenge of the application of PLS is in the interpretation of weights associated with the predictors. With hundreds and thousands of predictors, each and every predictor variable has non-zero weight. However, we expect that only some predictor variables are contributing to the association with the outcome variable. We, therefore, resort to the sparse estimation of predictor weights where some weights are zero estimated and the other weights are non-zero. A (standard) lasso estimation has a weakness in dealing with correlated variables as it picks up one variable within a correlation block without knowing the reason. A novel approach is needed to consider the dependencies between predictor variables in estimating the weights. We propose a new method where a new penalty function is introduced in the likelihood function associated with the estimation of weights. The penalty function is a combination of a lasso penalty that imposes sparsity and a penalty based on Cauchy distribution with a smoother matrix to take into account dependencies between genomic regions. The results show that the estimates of the weights are sparse: many weights are zero estimated, and those non-zero estimates are grouped and exhibit smoothness within them. The interpretation of genomic regions becomes easy, and the identification of important regions for each component can be done simultaneously with prediction in a single modeling framework. We investigate the relation between PLS and graphical modeling using the information in the weights to construct the graph with unsuccessful results. High-dimensional data where the number of predictors (p) exceeds the number of observations (n) are widely used in many applications of regression analysis. Ordinary least squares regression (OLS), which is the most well-known method for regression problems, has less performance with high-dimensional and highly- correlated data. Previous studies have shown that there is an association between copy number alterations (CNA) in some key genes and disease phenotypes. Moreover, it is very important in high-dimensional data to classify the samples into groups, such as tumor types, of gene expression data in bioinformatics and biology. However, the standard regression of classification methods will fail in these cases because the predictors matrix is singular and so, cannot be inverted. Hence, regularised methods are needed such as shrinkage methods and dimension reduction methods. One of the most suggested methods in the literature is partial least squares regression (PLS) for linear regression and classification.

Keywords : negative filter factors, partial least square regression, high-dimensional data, biostatistics, bioinformatics
Conference Title : ICAEM 2024 : International Conference on Applied and Engineering Mathematics
Conference Location : Venice, Italy
Conference Dates : June 20-21, 2024