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

Authors : Mohammed Abdullah Alshahrani

Abstract : The research focuses on investigating the use of partial least squares (PLS) methodology for addressing challenges associated with high-dimensional correlated data. Recent technological advancements have led to experiments producing data characterized by a large number of variables compared to observations, with substantial inter-variable correlations. Such data patterns are common in chemometrics, where near-infrared (NIR) spectrometer calibrations record chemical absorbance levels across hundreds of wavelengths, and in genomics, where thousands of genomic regions' copy number alterations (CNA) are recorded from cancer patients. PLS serves as a widely used method for analyzing high-dimensional data, functioning as a regression tool in chemometrics and a classification method in genomics. It handles data complexity by creating latent variables (components) from original variables. However, applying PLS can present challenges. The study investigates key areas to address these challenges, including unifying interpretations across three main PLS algorithms and exploring unusual negative shrinkage factors encountered during model fitting. The research presents an alternative approach to addressing the interpretation challenge of predictor weights associated with PLS. Sparse estimation of predictor weights is employed using a penalty function combining a lasso penalty for sparsity and a Cauchy distribution-based penalty to account for variable dependencies. The results demonstrate sparse and grouped weight estimates, aiding interpretation and prediction tasks in genomic data analysis. High-dimensional data scenarios, where predictors outnumber observations, are common in regression analysis applications. Ordinary least squares regression (OLS), the standard method, performs inadequately with high-dimensional and highly correlated data. Copy number alterations (CNA) in key genes have been linked to disease phenotypes, highlighting the importance of accurate classification of gene expression data in bioinformatics and biology using regularized methods like PLS for regression and classification.

Keywords : partial least square regression, genetics data, negative filter factors, high dimensional data, high correlated data
Conference Title : ICCAM 2024 : International Conference on Computational and Applied Mathematics
Conference Location : Amsterdam, Netherlands
Conference Dates : August 05-06, 2024