Global Criteria for Sparse Penalized Partial Least Squares

Tzu-Yu Liu
University of Michigan, Ann Arbor, United States joyliu@umich.edu

Laura Trinchera
Rouen Business School, Mont-Saint-Aignan, France laura.trinchera@rouenbs.fr

Arthur Tenenhaus
Supélec, Gif-sur-Yvette, France Arthur.Tenenhaus@supelec.fr

Dennis Wei
University of Michigan, Ann Arbor, United States dlwei@eecs.umich.edu

Alfred O. Hero*
University of Michigan, Ann Arbor, United States hero@eecs.umich.edu

With advancing technology comes the need to extract information from increasingly high-dimensional data, whereas the number of samples is often limited. Dimension reduction techniques and models with sparsity become important problems. Partial least squares (PLS) regression combines dimensionality reduction and prediction using a latent variable model. It was first developed for regression analysis in chemometrics, and has been successfully applied to many different areas, including sensory science and more recently genetics. Moreover, PLS algorithm is designed precisely to operate with high dimensional data but the resulting PLS model tend to overfits when the number of predictors increases while the number of samples is limited. Therefore, variable selection becomes essential for PLS to be applied to high-dimensional sample-limited problems. It not only avoids over-fitting, but also provides more accurate predictors and yields more interpretable estimates. In this work we propose a global criterion for PLS that changes the sequential optimization for a K component model in Statistically Inspired Modification of PLS (SIMPLS) into a unified optimization formulation, which we refer to as global SIMPLS. This enables us to perform global variable selection, which penalizes the total number of variables across all PLS components. We formulate PLS with global sparsity as a variational optimization problem with the objective function equal to the global SIMPLS criterion plus a mixed norm sparsity penalty on the weight matrix. The mixed norm sparsity penalty is the L1 norm of the L2 norm on the subsets of variables used by each PLS component. The proposed global penalty guarantees that the selected variables are shared among ll the K PLS components. A novel augmented Lagrangian method is proposed to solve the optimization problem, which enables us to obtain the global SIMPLS components and to perform joint variable selection simultaneously. A greedy algorithm is proposed to overcome the computation difficulties in the iterations, and soft thresholding for sparsity occurs naturally as part of the iterative solution. Experimental results show that our approach to PLS regression attains better performance (lower mean squared error, MSE) with many fewer selected predictor variables. These experiments include a chemometric wine data set, a human viral challenge study dataset, in addition to numerical simulations.

Key Words: PLS, variable selection, dimension reduction, augmented Lagrangian optimization.