Variable Selection in Partial Least Squares Methods: overview and recent developments

Laura Trinchera, Edith Le Floch, Arthur Tenenhaus

Variable Selection in Partial Least Squares Methods: overview and recent developments

Laura Trinchera[1], Edith Le Floch[2], Arthur Tenenhaus[1]

[1]SUPELEC, Department of Signal Processing and Electronic Systems, Gif-sur-Yvette, France
laura.trinchera@supelec.fr arthur.tenenhaus@supelec.fr

[2]CEA, Neurospin, LNAO, Gif-sur-Yvette, France edith.lefloch@gmail.com

Keywords: Partial Least Squares, landscape matrix, sparsity

Abstract. Recent developments in technology enable collecting a large amount of data from various sources. Moreover, many real world applications require studying relations among several groups of variables. The analysis of landscape matrices, i.e. matrices having more columns (variables, p) than rows (observations, n), is a challenging task in several domains.

Two different kinds of problems arise when dealing with high dimensional data sets characterized by landscape matrices. The first refers to computational and numerical problems. The second deals with the difficulty in assessing and understanding the results. Dimension reduction seems to be a solution to solve both problems. We should distinguish between feature selection and feature extraction. The first refers to variable selection, while feature extraction aims to transform the data from high-dimensional space to low-dimensional space.

Partial Least Squares (PLS) methods are classical feature extraction tools that work in the case of high-dimensional data sets. Since PLS methods do not require matrices inversion or diagonalization, they allow us to solve computational problems. However, results interpretation is still a hard problem when facing with very high-dimensional data sets. Moreover, recently Chun & Keles (2010) showed that asymptotic consistency of PLS regression estimator for the univariate case does not hold with the very large p and small n paradigm.

Nowadays interest is increasing in developing new PLS methods able to be, at the same time, a feature extraction tool and a feature selection method. The first attempt to perform variable selection in univariate PLS Regression framework was presented by Bastien et al. in 2005. More recently Le Cao et al. (2008) and Chun & Keles (2010) proposed two different approaches to include variable selection in PLS Regression, based on L1 penalization (Tibshirani, 1996).

In our work, we will investigate all these approaches and discuss the pros and cons. Moreover, a new version of PLS Path Modeling algorithm including variable selection will be presented.

Main references.

Bastien P., Esposito Vinzi V., Tenenhaus M. (2005). PLS generalised linear regression. Computational Statistics and Data Analysis, 48: 17-46.

Chun H., Keles S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. Journal of the Royal Statistical Society - Serie B, 72 (1): 3-25.

Le Cao K.A., Rossouw D., Robert-Granié L.C., Besse P. (2008). A sparse PLS for variable selection when integrating omics data. Statistical Application in Genetics and Molecular Biology, 7 (1).

Tibshirani R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B, 58(1): 267–288.