Software implementation of the Hard and Soft Partial Least Squares Discriminant Analysis

Y.V. Zontov\(^a\), S.V. Kucheryavskiy\(^b\), O.Ye. Rodionova\(^c,d\), A.L. Pomerantsev\(^c,d\)

\(^a\)National Research University Higher School of Economics (HSE), Moscow, Russia
\(^b\)Department of Chemistry and Bioscience, Aalborg University, Esbjerg, Denmark
\(^c\)Semenov Institute of Chemical Physics RAS, Moscow, Russia
\(^d\)Branch of Institute of Natural and Technical Systems RAS, Sochi, Russia

We present a software implementation of hard and soft approaches to Partial Least Squares Discriminant Analysis (PLS-DA) [1]. The Soft PLS-DA is based on Quadratic Discriminant Analysis applied to the super-score matrix T. It can simultaneously attribute a sample to several classes. It also allows to detect samples, which are not members of any training classes, and, therefore, reduce a number of false positives e.g. in the presence of outliers. The Hard PLS-DA represents mostly the conventional PLS-DA classification. Thus the software can be used for both multi-class as well as two-class classification.

Both methods are implemented as MATLAB toolbox, PLS-DA Tool, using object oriented programming. The toolbox provides instruments for data pre-processing as well as for interpretation, validation and visualization of classification models. The main class, \textit{PLSDAModel}, is responsible for the logic and contains implementation of both methods and auxiliary algorithms. The instance of this class has fields, which represent the actual model, and methods for data visualization and statistics. The \textit{PLSDAGUI} class provides graphical user interface, where a user can create and manipulate datasets, calibrate, validate and explore models interactively. This class encapsulates all the necessary data such as training or validation sets and labels of the samples and variables, and uses these elements for the model development and validation. The PLS-DA Tool has its own implementation of all necessary statistical functions and does not require the MATLAB Statistics Toolbox.

References