libPLS: An Integrated Library for Partial Least Squares Regression and Discriminant Analysis

Hong-Dong Li[1]*, Qing-Song Xu[2] and Yi-Zeng Liang[1*]

1 College Of Chemistry and Chemical Engineering, Central South University, Changsha 410083, P.R. China
2 School of Mathematics and Statistics, Central South University, Changsha 410083, P.R. China

Correspondence: lhdcsu@gmail.com or yizeng_liang@263.net

Abstract

Partial least squares (PLS) have gained wide applications especially in chemometrics, metabolomics/metabonomics as well as bioinformatics. To our knowledge, an integrated PLS library that include not only basic PLS modeling algorithms but also advanced and/or recently developed methods on model assessment, outlier detection and variable selection is in lack. Here we present libPLS which provides an integrated platform for developing PLS regression and/or discriminant analysis (PLS-DA) models. This library is written in MATLAB and freely available at www.libpls.net.

Keywords: Partial least squares, outlier detection, variable selection, model population analysis

Introduction

Partial least squares (PLS) are the cornerstone method in chemometrics [1-3] and have been widely used in other fields such as metabolomics/metabonomics [4,5], bioinformatics [6]. Software for developing PLS regression or discriminant analysis (PLS-DA) models are available, such as SIMCA-P, the PLS toolbox and the PLS R-package. Building a PLS model usually involves several steps such data pretreatment, cross validation, model development and validation. If a model performs poor, outlier detection and variable selection might help a lot. Outlier detection/removal can help improve the quality of data [7-12]. With variable selection, we are able to single out a sub-set of informative variables which may lessen overfitting, greatly improve a model's performance and allows for an easy-to-explain model especially in the situation of "large *p*, small *n*" setting where a large number of irrelevant or interfering variables may exist [13-23]. To our knowledge, many developed especially recently developed methods for outlier detection and variable selection scatter in different individually owned in-house codes or software. And there is a lack of software that integrates data pretreatment, outlier detection, variable selection, model assessment and PLS modeling together so as to facilitate the modeling procedure.

In the present work, we presented libPLS which provides an integrated environment for PLS regression [3] or discriminant analysis [24,25]. Except for PLS algorithms, it contains a set of useful modeling-related methods including data pretreatment, the

Kennard-Stone method for sample partition [26], the Monte Carlo method for outlier detection [12], uninformative variable elimination (UVE) [15,21,27] and Competitive Adaptive Reweighted Sampling (CARS) for variable selection [18], and Monte Carlo cross validation [28] for model assessment [29] and so on. Specifically, this library is featured in a set of model population analysis (MPA)-based methods [16,19,20,30], which are a new type of data analysis algorithms developed based on the statistical analysis of user-interested outputs of a large number of sub-models built with the help of resampling. MPA approaches are expected to give more reliable results than those methods based on a single model [30]. With this library, it is expected that users can develop their PLS or PLS-DA models easily.

The algorithms provided in this library are not aimed to be comprehensive, but are expected to be helpful. We provided detailed document on how to build and assess PLS models at www.libpls.net. And we will update this library when necessary.

Methods

There are different algorithms for implementing PLS, including the SIMPLS [2], the Nonlinear Iterative Partial Least Squares (NIPALS) algorithm [31] and the eigenvector decomposition based method. The NIPALS algorithm is used in this library.

In the current version (ver1.6) of libPLS, we implemented a series of algorithms covering different aspects of PLS modeling, which are categorized and listed in **Table 1.** We will not detail each method here but with references provided. To guide the use of these methods, reader are recommended to refer to the two demo scripts named demo_PLS_Regression.m and demo_PLS_Discriminant_Analysis.m. Simply by running these domos, you will know how to run a PLS modeling procedure. Alternatively, the online documentation would be also helpful.

Table 1. Implemented algorithms in the current version (ver1.6) of libPLS.

Category	Algorithms
Model building	Partial least squares (PLS)
	Linear discriminant analysis (LDA)
Data pretreatment	Mean-centering
	autoscaling
Sample partition	Kennard-Stone algorithm [26]
Model assessment	leave-one-out cross validation(LOOCV)
	K-fold cross validation
	double cross validation (DCV)
	Monte Carlo cross validation (MCCV) [28]
	repeated double cross validation (RDCV) [32]
	Using an independent test set
Outlier detection	The Monte Carlo method [12]
Variable selection	Variable importance in projection(VIP)
	Target Projection (TP) [33,34]
	Uninformative Variable Elimination (UVE, also MC-UVE) [15,21,27]
	Competitive Adaptive Reweighted Sampling (CARS-PLS, CARS-PLSDA) [18]
	Random Frog (coupled with PLS or PLS-DA) [35]
	Subwindow Permutation Analysis (coupled with PLS-DA) [19]
	Moving Window Partial Least Squares(MWPLS) [22]

Conclusions

We have developed the libPLS library which is aimed to facilitate the procedure of building PLS or PLS-DA models as well as performing related data analysis such as outlier detection and variable selection. It contains a set of algorithms covering different aspects of PLS modeling. This library is open source and freely available.

References

1. Geladi P, Kowalski BR (1986) Partial least-squares regression: a tutorial. Anal Chim Acta 185: 1-17.
2. De Jong S (1993) SIMPLS: An alternative approach to partial least squares regression. Chemometr Intell Lab 18: 251-263.
3. Wold S, Sjöström M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics. Chemometr Intell Lab 58: 109-130.
4. Yi L-Z, He J, Liang Y-Z, Yuan D-L, Chau F-T (2006) Plasma fatty acid metabolic profiling and biomarkers of type 2 diabetes mellitus based on GC/MS and PLS-LDA. FEBS Letters 580: 6837-6845.
5. Trygg J, Holmes E, Lundstedt Tr (2006) Chemometrics in Metabonomics. Journal of Proteome Research 6: 469-479.
6. Nguyen D, Rocke DM (2002) Tumor classification by partial least squares using microarray gene expression data. Bioinformatics 18: 39 - 50.
7. Barnett V, Lewis T (1994) Outliers in Statistical Data. New York: John Wiley &Sons.
8. Walczak B (1995) Outlier detection in multivariate calibration. Chemometr Intell Lab 28: 259-272.
9. Egan WJ, Morgan SL (1998) Outlier Detection in Multivariate Analytical Chemical Data. Anal Chem 70: 2372-2379.
10. Hubert M, Vanden Branden K (2003) Robust methods for partial least squares regression. J Chemometr 17: 537-549.
11. Verboven S, Hubert M (2005) LIBRA: a MATLAB library for robust analysis. Chemometr Intell Lab 75: 127-136.
12. Cao DS, Liang YZ, Xu QS, Li HD, Chen X (2010) A New Strategy of Outlier Detection for QSAR/QSPR. J Comput Chem 31: 592-602.
13. Ugulino Araújo MC, Saldanha TCB, Galvão RKH, Yoneyama T, Chame HC, et al. (2001) The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. Chemometr Intell Lab 57: 65-73.
14. Lee KE, Sha N, Dougherty ER, Vannucci M, Mallick BK (2003) Gene selection: a Bayesian variable selection approach. Bioinformatics 19: 90-97.
15. Cai W, Li Y, Shao X (2008) A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra. Chemometr Intell Lab 90: 188-194.
16. Li H-D, Liang Y-Z, Xu Q-S, Cao D-S (2009) Model population analysis for variable selection. J Chemometr 24: 418-423
17. Rajalahti T, Arneberg R, Berven FS, Myhr K-M, Ulvik RJ, et al. (2009) Biomarker discovery in mass spectral profiles by means of selectivity ratio plot. Chemometr Intell Lab 95: 35-48.
18. Li H-D, Liang Y-Z, Xu Q-S, Cao D-S (2009) Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. Anal Chim Acta 648: 77-84.

19. Li H-D, Zeng M-M, Tan B-B, Liang Y-Z, Xu Q-S, et al. (2010) Recipe for revealing informative metabolites based on model population analysis. Metabolomics 6: 353-361.
20. Li H-D, Liang Y-Z, Xu Q-S, Cao D-S, Tan B-B, et al. (2011) Recipe for Uncovering Predictive Genes using Support Vector Machines based on Model Population Analysis. IEEE/ACM T Comput Bi 8: 1633-1641.
21. Centner V, Massart D-L, de Noord OE, de Jong S, Vandeginste BM, et al. (1996) Elimination of Uninformative Variables for Multivariate Calibration. Anal Chem 68: 3851-3858.
22. Jiang J-H, Berry RJ, Siesler HW, Ozaki Y (2002) Wavelength Interval Selection in Multicomponent Spectral Analysis by Moving Window Partial Least-Squares Regression with Applications to Mid-Infrared and Near-Infrared Spectroscopic Data. Anal Chem 74: 3555-3565.
23. Xu H, Liu Z, Cai W, Shao X (2009) A wavelength selection method based on randomization test for near-infrared spectral analysis. Chemometr Intell Lab 97: 189-193.
24. Barker M, Rayens W (2003) Partial least squares for discrimination. J Chemometr 17: 166-173.
25. Westerhuis JA, Hoefsloot HCJ, Smit S, Vis DJ, Smilde AK, et al. (2008) Assessment of PLSDA cross validation. Metabolomics 4: 81-89.
26. Kennard RW, Stone LA (1969) Computer aided design of experiments. Technometrics 11: 137-148.
27. Han Q-J, Wu H-L, Cai C-B, Xu L, Yu R-Q (2008) An ensemble of Monte Carlo uninformative variable elimination for wavelength selection. Anal Chim Acta 612: 121-125.
28. Xu Q-S, Liang Y-Z (2001) Monte Carlo cross validation. Chemometr Intell Lab 56: 1-11.
29. Stone M (1974) Cross-validatory choice and assessment of statistical predictions. J R Stat Soc B 36: 111-147.
30. Li H-D, Liang Y-Z, Xu Q-S, Cao D-S (2012) Model population analysis and its applications in chemical and biological modeling. TrAC 38: 154-162.
31. Wold H (1975) Path models with latent variables: The NIPALS approach. In H. B. et al. (Ed.),Path models with latent variables: The NIPALS approach. In H. B. et al. (Ed.),ing (pp. 307–357). Academic Press.
32. Filzmoser P, Liebmann B, Varmuza K (2009) Repeated double cross validation. J Chemometr 23: 160-171.
33. Arneberg R, Rajalahti T, Flikka K, Berven FS, Kroksveen AC, et al. (2007) Pretreatment of Mass Spectral Profiles: Application to Proteomic Data. Anal Chem 79: 7014-7026.
34. Rajalahti T, Arneberg R, Kroksveen AC, Berle M, Myhr K-M, et al. (2009) Discriminating Variable Test and Selectivity Ratio Plot: Quantitative Tools for Interpretation and Variable (Biomarker) Selection in Complex Spectral or Chromatographic Profiles. Anal Chem 81: 2581-2590.
35. Li H-D, Xu Q-S, Liang Y-Z (2012) Random Frog: an efficient reversible jump Markov Chain Monte Carlo-like approach for gene selection and disease classification. Anal Chim Acta 740: 20-26.