

ParLeS v2.0 - freeware to pre-process spectra and perform Partial Least Squares (PLS) regression with delete-one-jackknife cross validation

R.A. VISCARRA ROSSEL

Australian Centre for Precision Agriculture, Faculty of Agriculture, Food & Natural Resources, McMillan Building A05, The University of Sydney, NSW 2006, Australia

Partial Least Squares (PLS) regression (also known as PLSR or PLSR1) is a popular modelling technique in chemometrics, econometrics and in industrial applications. It is a technique that is also commonly used in spectral quantitative analysis. Research in science often involves using variables that are easily (or cheaply) measured to explain or predict the behaviour of response variables that are often much more difficult (or expensive) to acquire. When the factors are few in number and are not significantly redundant (or collinear) and have a well understood relationship to the responses, then multiple linear regression (MLR) can be a useful way to turn data into information. However, if these conditions break down, then MLR will not be efficient or appropriate. PLS is a method used to construct predictive models when factors are many and highly collinear, e.g. in reflectance spectroscopy. The emphasis of PLSR is on predicting the response. However, when used interactively with proper graphics and validation, it also allows the user to attain a good causal insight into the underlying relationships between the variables.

PLSR is closely related to Principal Component Regression (PCR). However, PLSR is performed in a slightly different manner. Take the case where we want to use spectral reflectance data to model and then estimate the value/concentration of a soil property: Instead of first decomposing the spectra into a set of eigenvectors and scores, and regressing them against the soil values as a separate step, PLSR actually uses the soil information during the decomposition process (the decomposition of both the spectral and the soil data into their most common variations is performed simultaneously). PLSR takes advantage of the correlation that exists between the spectra and the soil values. So, the resulting spectral vectors are directly related to the soil values/concentrations.

Advantages of PLSR:

- handles multicollinearity
 - robust in terms of data noise and missing values
 - balances the two objectives of explaining response and predictor variation thus predictions are more robust
 - calibrations generally more robust
 - single step decomposition and regression
 - can give good insight into underlying relationship between variables
-

Disadvantages of PLSR:

- calculations are slightly slower than more classical methods, although this is no longer much of an issue

ParLeS v2.0 (Fig. 1) is freeware available upon email request (r.rossel@agec.usyd.edu.au) that provides options for, and performs the following tasks: (i.) spectral data linearization through transformations to absorbance and/or Kubelka-Munk optical density units (ii.) spectral corrections, de-trending and smoothing using various methods including multiplicative signal correction (MSC), standard normal variate (SNV), wavelet transform, first and second derivatives, etc., (iii.) mean centre and or variance scale data (iii.) delete-one-jackknife cross validation to determine optimal number of bilinear factors for (iv.) calibration modelling and (v.) prediction of unknowns. The following data can be saved for external analysis and plotting: (i.) pre-processed spectra, (ii.) the cross validation predictions, (iii.) the scores (t), loadings (P), weights (W) and slopes (B) of the regression and the predictions of unknowns.

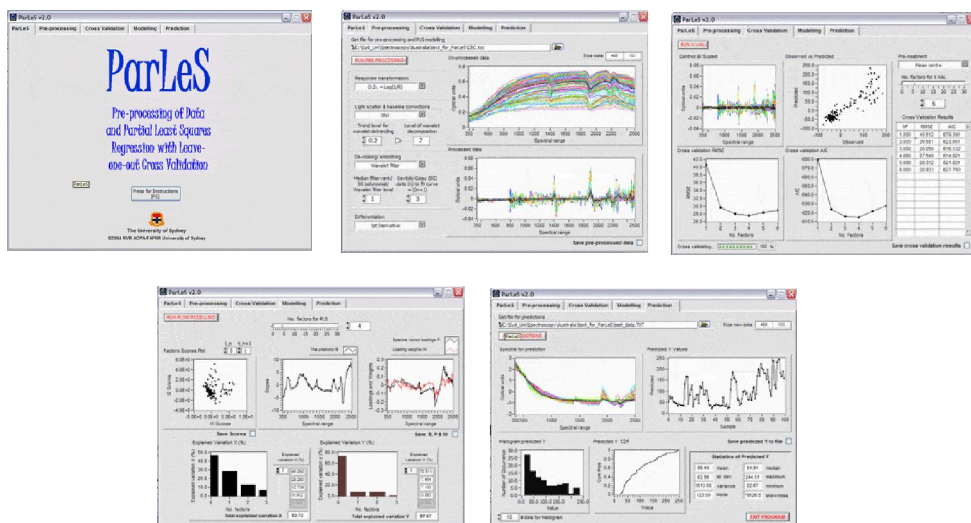


Fig 1. ParLeS v2.0 interface