Ordered Predictors Selection: an intuitive method to find the most relevant variables in multivariate regression

Reinaldo Francisco Teófilo, João Paulo Ataide Martins*, Márcia Miguel Castro Ferreira

Laboratório de Quimiometria Teórica e Aplicada – Instituto de Química – Universidade Estadual de Campinas

Keywords: predictor selection, multivariate calibration, chemometrics

Multivariate regression techniques are widely used to model chemical, physical, sensory data, besides quantitative structure activity and property relationships, i.e. QSAR and QSPR, respectively. The quality of a multivariate calibration model depends, among others, on the quality of the data (objects and also variables). Several strategies are available to evaluate the predictive ability of a regression model by measuring the error on objects that were not used when building the regression model. In the original Partial Least Squares (PLS) and Principal Component Regression (PCR) methods, all variables were used and hence they were denominated full-spectrum methods[1]. However, it is well known that better results could be obtained when only the most important variables were selected and applied. Variable selection is crucial, especially in QSAR/QSPR studies.

This work introduces a simple and intuitive method for feature selection. In this method, the variables are sorted in a decreasing order with respect to its importance to the model. The ordered variables are then evaluated using increments over a window previously defined. The root mean square errors of cross-validation (RMSECV) and the correlation coefficient of cross-validation (r_{cv}) values are stored to each analyzed window. The best variables are indicated by lower RMSECV and higher r_{cv}. As a consequence, the algorithm was named *Ordered Predictors Selection* (OPS). Several vectors or their combinations can be used to order the variables, as the correlation vector, regression vector, loadings vectors combination, modeling power vector, the difference between the spectra relative to the higher and lower concentration values, and so on. The choice will depend on the data set under study. The advantages of this method are: (i) applicability to highly correlated data set (spectroscopy, voltammetry, process control) and to data that present lower correlation among the variables (QSAR, QSPR, mass spectrometry, nuclear magnetic resonance); (ii) objectivity on selecting variables that present the relevant chemical information, since the vectors chosen for variable sorting are selected too; (iii) requirement of few input parameters for the selection method, being necessary only the independent variable matrix and the dependent variable vector.

The algorithm was written in Matlab code and its performance was evaluated on three data set, i.e., QSPR data (Set A), Mid-infrared data (Set B) and Voltammetric data (Set C). The table presents the results for the full data set and OPS data. The regression method used was the PLS. It is concluded that the selection of informative variables improved the predictive ability, besides making the models more interpretative and parsimonious, since a lower number of variables was selected.

		Full data					OPS data				
	Factors	**nVars**	**rmsecv**	**r_{cv}**	**rmsep**	**r_p**	**nVars**	**rmsecv**	**r_{cv}**	**rmsep**	**r_p**
Set A	2	677	0.199	0.42	0.183	0.79	5	0.151	0.73	0.083	0.99
Set B	4	3351	34.68	0.59	22.770	0.88	30	25.03	0.81	6.7	0.99
Set C	4	353	0.027	0.95	0.016	0.97	30	0.009	0.99	0.0035	0.99

Acknowledgments. To CNPq and FAPESP for their financial support.

References

[1] Martens H.; Naes T. Multivariate Calibration (2nd edn), vol.1. Wiley: Chichester, UK, 1989, 35-70.