Abstract
Methods based on statistics for quantitative modelling of near infrared spectra

X. Shao,* H. Xu, Z. Liu and W. Cai
College of Chemistry, Nankai University, Tianjin 300071 P.R. China.
E-mail: xshao@nankai.edu.cn

Introduction

Partial least squares (PLS) models for quantitative analysis of near infrared (NIR) spectral spectroscopy are often not satisfactory, due to the presence of outlier and the redundant variables. As useful tools for signal analysis, statistics can be used to test the significance of objects in a way that accounts for randomness and uncertainty in the observations. The result derived from these data seems more reliable. In this work, in order to construct high quality PLS models, methods based on statistics are proposed for outlier detection and variable selection in NIR analysis, respectively.

Theory and algorithm

The underlying philosophy of the statistic-based methods is that, by building random models, the probability of objects such as outliers or important variables in a normal model should be significantly different from that in random models. The method for outlier detection builds PLS models by using random test cross-validation, the models are sorted by the prediction residual error sum of squares (*PRESS*), and the outliers are recognised according to the accumulative probability of each sample in the sorted models. On the other hand, two statistic-based methods are proposed for variable selection. One method is named as MC-UVE, which is an integration of the Monte Carlo (MC) technique and uninformative variable elimination (UVE). The method evaluates the reliability of each variable through calculating the stability of each variable, and the stability is calculated with coefficients of random models built with MC technique. Another method i.e., RT-PLS is based on the randomisation test, in which significance testing of each variable is performed. With the statistics obtained by the significance tests, important variables can be detected.

Results and discussion

Figure 1 shows an example of the method for outlier detection.

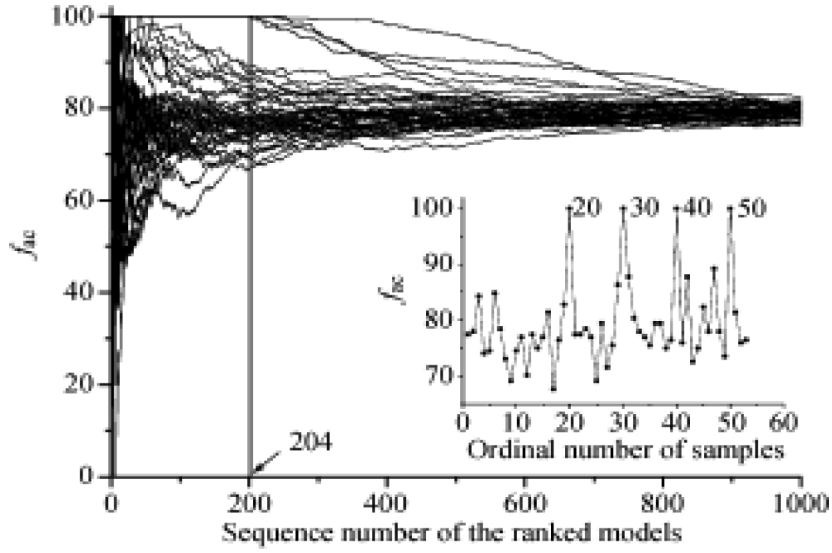

Figure 1. A statistical plot of 1000 models obtained with the proposed outlier detection method.

The four artificial outliers were all detected by the proposed method, however, only outlier No.30 was detected by the leave one out cross-validation (LOOCV) method. The proposed method was more sensitive to the outliers than the LOOCV method.

Table 1 summarizes the mean *RMSEP* with standard deviation (σ) over 100 runs of the MC-UVE-PLS model for prediction of the oil content in corn samples.

The *RMSEP* obtained by MC-UVE-PLS was obviously better than that by PLS and UVE-PLS. On the other hand, fewer variables were used in the MC-UVE-PLS model. Figure 2 shows the statistic values obtained by RT-PLS for the variables in the NIR spectra of plant samples.

The variables with statistic values below the dotted line were considered as important ones and retained. With the retained variables, the mean *RMSEP* with standard deviation were 0.0927 and 0.0015, respectively. The results indicate that statistic-based methods like MC-UVE-PLS and RT-PLS can build effective models and improve the predictive accuracy of models.

Table 1. A comparison of the results obtained by PLS, UVE-PLS and MC-UVE-PLS.

Model	Number of variables	Component number	*RMSEP*(σ*)
PLS	700	4	0.1215
UVE-PLS	131	4	0.1003
MC-UVE-PLS	110	4	0.0974(0.0031)

*σ is the standard deviation of 100 *RMSEP*s.

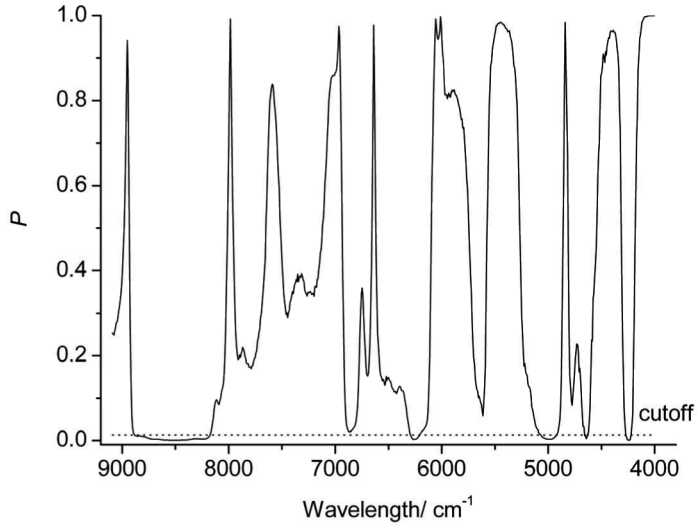

Figure 2. Statistic values of the variables in NIR spectra obtained by RT-PLS.