CHINESE JOURNAL OF ANALYTICAL CHEMISTRY
Volume 44, Issue 2, February 2016
Online English edition of the Chinese language journal

Available online at www.sciencedirect.com
ScienceDirect

Cite this article as: Chin J Anal Chem, 2016, 44(2), 305–309.

RESEARCH PAPER

Outlier Detection for Multivariate Calibration in Near Infrared Spectroscopic Analysis by Model Diagnostics

LI Zheng-Feng[1], XU Guang-Jin[1], WANG Jia-Jun[1], DU Guo-Rong[2], CAI Wen-Sheng[2], SHAO Xue-Guang[2,3,*]

[1] R&D Center, China Tobacco Yunnan Industrial Co. Ltd., Kunming 650231, China
[2] Research Center for Analytical Sciences, College of Chemistry, Nankai University, Tianjin 300071, China
[3] College of Chemistry and Environmental Science, Kashgar University, Kashgar 844000, China

Abstract: Outlier detection is an important task in multivariate calibration because the quality of a calibration model is determined by that of the calibration data. An outlier detection method was proposed for near infrared (NIR) spectral analysis. The method was based on the definition of outlier and the principle of partial least squares (PLS) regression, i.e., an outlier in a dataset behaved differently from the rest, and the prediction result of a PLS model was an accumulation of several independent latent variables. Therefore, the proposed method built a PLS model with a calibration dataset, and then the contribution of each latent variable was investigated. Outliers were detected by comparing these contributions. An NIR spectral dataset of orange juice samples was adopted for testing the method. Six outliers were detected in the calibration set. The root mean squared error of cross validation (RMSECV) was reduced from 16.870 to 4.809 and the root mean squared error of prediction (RMSEP) was reduced from 3.688 to 3.332 after the removal of the outliers. Compared with a robust regression method, the result of the proposed method seemed more reasonable.

Key Words: Multivariate calibration; Outlier detection; Partial least squares (PLS); Near infrared spectroscopy; Quantitative analysis

1 Introduction

Near infrared (NIR) spectroscopy is a powerful tool widely used in measuring chemical and physical properties, and multivariate calibration is the key technique in the quantitative analysis with NIR spectroscopy. For building quantitative models, various strategies were proposed, such as multivariate linear regression (MLR), principal component regression (PCR), partial least squares (PLS) regression[1,2], and support vector machines (SVM)[3–5], etc. For improving the applicability, the methods like nonlinear modeling, local regression, and consensus strategy were developed[6]. On the other hand, tremendous works were done for building a high quality model, including spectral pretreatment and variable selection techniques, such as multiplicative scattering correction (MSC), orthogonal signal correction (OSC)[7], wavelet transformation (WT)[8], interval PLS (iPLS)[9], uninformative variables elimination (UVE)[10,11], competitive adaptive reweighted sampling (CARS)[12], successive projections algorithm (SPA)[13], and randomization test (RT) method[14].

One of the important factors to determine the quality of a calibration model is the calibration set. Outlier detection is a difficult problem in multivariate calibration because outliers contained in the calibration set may have a significant effect on the quality of the model[15]. Therefore, a large number of methods were proposed to detect outliers in different kinds of dataset[16–19]. These methods were proved useful when the

Received 11 October 2015; accepted 28 October 2015
*Corresponding author. Email: xshao@nankai.edu.cn
This work is supported by the National Natural Science Foundation of China (No. 21475068) and the Major Project of China National Tobacco Corporation (No. Ts-03-20110020).

DOI: 10.1016/S1872-2040(16)60907-6

outliers did not interfere with each other. When multiple outliers existed in a dataset, masking and swamping phenomena occurred and may make these methods inefficient[18,19]. Robust modeling was a better strategy for dealing with the contaminated datasets because it reduced the pernicious effects of outliers automatically. Robust simple partial least squares (RSIMPLS)[20] was successfully used to detect outliers and improve the calibration model. In the method, a regression diagnostic plot was constructed to visualize and classify the outliers, i.e., good leverage, bad leverage and vertical outlier. The two leverages were extreme objects in the calibration spectra, and the good one followed the regression model obtained with the majority of the observations while the bad one did not. A vertical outlier was defined as an observation that had a large concentration residual in the calibration.

In this study, an outlier detection method was proposed from a completely different aspect by comparing the behavior of the observations in a PLS model. The basic assumption of the method was that an outlier in a dataset behaved differently from the rest. In a PLS model, the prediction result was an accumulation of several independent principal components, each component representing a factor of the data variance. If the weights of an observation in different components were significantly different from the others, it could be an outlier. The essence of the proposed method was to investigate the behavior of each sample in each principal factor of the model. Therefore, the method was named as "model diagnostics".

2 Theory and calculations

Generally, an outlier is defined as a value in a dataset that does not fit with the rest. In PLS modeling, the definition can be extended as an observation (including the reference concentration and spectrum) that does not fit with the model built with the others. Therefore, in some cases, 'influential observation' is a better description of the outliers, because outliers may have bad or good influence on a model[17,21]. In this study, the outlier was defined as the observation that behaves differently from the others in a dataset.

Several principal components (PCs) or latent variables (LVs) were included in a PLS model, and the prediction result was an accumulation of several independent principal components or latent variables (LVs) in a PLS model. For majority of the observations, the weights were distributed in a reasonable range, but for the outliers, their weights might significantly different from the others. Therefore, if the weights of an observation in the LVs were distributed in a different way, it would be considered as an outlier.

To show the principle of the proposed method, a simulated dataset was generated by using Gaussian function. A total of 60 spectra with 101 data points were simulated. Each spectrum was generated with the combination of the six peaks of different intensity (concentration), which were generated by random numbers. Furthermore, 1.0% (in intensity of the signal) random noise was added to the spectra. Taking the 4th component as the predicting target, the weights of the observations in each LV of the PLS model is plotted in Fig.1a. The weight means the contribution to the predicted result. Because both the spectra and the concentrations were centralized in the modeling, the weights were distributed around zero. As shown in Fig.1a, the first six LVs had a significant contribution to the predicted results. This evidently corresponded to the correct number of the components, which were discussed in our previous work[22]. On the other hand, the Fig.1a also shows that the weights in each LV was evenly distributed, implying no outlier in the dataset.

For simulating the outliers in the dataset, artificial errors were added to the observation No. 10, 20, 30, 40 and 50. In the concentration vector, three times of the standard deviation of the concentrations were added to the observation No. 10, 20 and 30, and the extra spectral information was added to the spectrum No. 40 and 50, respectively. Figure 1b displays the weights of the components in the PLS model. Compared with Fig.1a, there were two apparent differences, i.e., (1) extra LVs were needed due to the outliers, and (2) outliers (No. 10, 20, 30, 40 and 50 as labeled in the figure) were clearly observed from the weights of 6th and 7th LVs.

As shown in Fig.1b, the outliers were found out more easily, local outlier factor (LOF)[23] was adopted. LOF was based on a concept of a local density, where locality was given by *k* nearest neighbors, whose distance was used to estimate the density. By comparing the local density of an observation to the local densities of its neighbors, the points that had a substantially

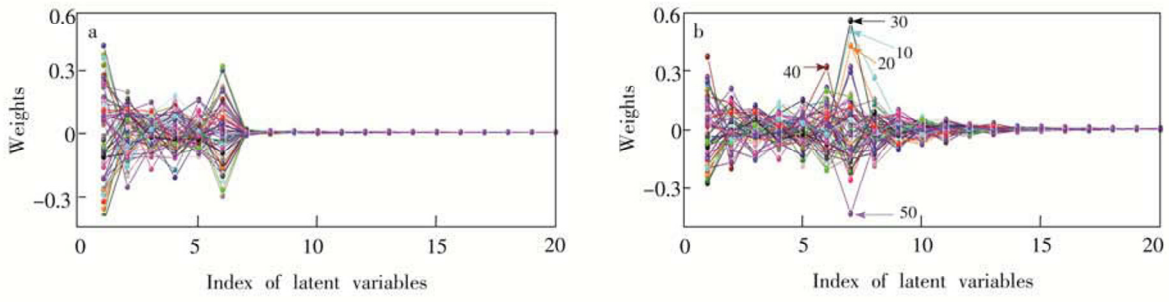

Fig.1 Weights of each sample in each factor in PLS model of simulated spectra
(a) Without outlier; (b) with outliers