Identification of singular samples in near infrared spectrum correction set by using Monte Carlo cross validation

Aixiao Zou [1, a], Yaxin Qu[2, b]

[1]College of electrical and control engineering of North China University of Technology,Beijing 100144, China;

[1]College of electrical and control engineering of North China University of Technology,Beijing 100144, China.

[a]15650725309@163.com, [b]qyx@ncut.edu.cn

Keywords: Near infrared spectroscopy, Monte Carlo cross validation, Partial least square, singular sample

Abstract. Identification of singular samples is the basis of the robustness of the calibration model for near infrared spectra. By using the Monte Carlo method of cross validation (MCCV) ,this experiment identifies the singular samples in the calibration set of starch samples. By using the method of partial least squares (PLS) ,modeling of starch samples coming from the before and after eliminating singular samples of calibration set , which is validated by test set, finally ,comparing the stability and prediction accuracy of the model before and after eliminating abnormal samples by using root mean square error of cross validation (RMSECV), correlation coefficient (R) and root mean square error of prediction (RMSEP) as evaluation index. The results show that the near infrared singular sample recognition algorithm of MCCV can effectively eliminate the singular sample,which can make the correction model have lower RMSECV, RMSEP and higher R, and significantly improve the stability and prediction ability of the model.

1. Introduction

Near infrared (NIR) is the electromagnetic wave between the visible spectrometry (VIS) and the mid infrared (MIR).The material testing association of the United States is defined the spectral region of the wavelength 780~2526nm as the near infrared spectral region[1].The information contented in the near infrared spectral region is very rich, and the near infrared spectroscopy technology has the advantages of no pollution, no damage, on-line detection and simultaneous determination of multiple components[1,2].At present,a large number of studies have demonstrated that the NIR spectroscopy can be used for the quantitative and qualitative analysis of non destructive testing in agricultural products.

Near infrared spectroscopy as a technique to analyze the physical and chemical properties by building models, the accuracy of the spectral data judged quickly and accurately is the premise of the reliable analysis[3].However, in the actual acquisition process of near infrared spectroscopy, according to the influence of the measuring instrument and the environment (light, temperature and humidity) and other objective factors,also the difference between different operators' loading level, it will inevitably lead to some singular samples[4]. The existence of these singular samples will affect seriously or even change the distribution of the overall data, thus affecting the accuracy of the calibration model.Therefore, it is the base of establishing robust model to judge accurately and eliminate the singular sample.

The singular sample is also called outlier, irregular or out of bounds , generally refers to those who fall outside the sample vector in general[5].At present, the identification method of the singular samples can be divided into two categories: the classical identification method and the robust identification method.Classic identification methods include residual method [6,7], Mahalanobis distance[8], leverage value [7] and the principal component plot[9], etc;while robust identification methods include ellipsoidal multivariate trimming [10,11], minimum volume estimator[12], minimum covariance determinant [13], etc. These diagnosis methods are suitable for the identification of a single singular sample, but when the correction is concentrated in the presence of multiple singular samples, the recognition effect is not reached usually[14].

In view of the above problems, this paper establishes a method of identifying the singular samples in the partial least squares model of near infrared spectroscopy - Monte Carlo cross validation method.This method uses Monte Carlo stochastic (MCS) sampling method , selecting randomly for 80% of the calibration samples to establish partial least squares model, while the remaining 20% starch samples were used as a test set to validate the model. After 1000 cycles, the mean (MEAN) and variance (STD) of a set of prediction residuals are obtained, and then we make the mean- variance distribution map.It is the singular samples most likely that those located in the region of high mean and high standard deviation.

2. Monte Carlo Cross Validation Method

The MCCV method used in this paper includes five steps: The first step is to pre process the spectrum of the starch sample;secondly,using the cross-validation method to determine the best number of principal components;then using MSC to establish a large number of PLS models and get the prediction residual of each sample;the following step is to calculate the the mean (MEAN) and standard deviation (STD) of each sample prediction residual , and making the mean-variance distribution graph for all samples;finally, identifying the singular samples by observing the area of the sample .Specific methods and steps are as follows:

1.For the spectral data of the collected samples, using normalization and multiplicative scatter correction to pre process respectively.

2.In this paper,a cross-validation method is used to determine the best number of principal components,using RMSECV as a criterion.Specific working means: Firstly, establishing the principal component is d (f=1,2,... d),For one of the principal component ,m samples were selected from the calibration set of n samples for prediction,usually using the leave-one-out cross validation (LOOCV), that is, the value of m is 1.One samples were taken each time, and the rest of the (N-1) samples were used to establish the calibration model to predict the one sample.Then, selecting the other one from the n sample as a predictor, repeating this process.After repeating the process of modeling and prediction, until the n samples are predicted once and only once, then getting the predicted residual square sum (PRESS) which is corresponding to the principal component.

$$PRESS = \sum_{i=1}^{\mathrm{n}} (y_i - \hat{y}_i) \tag{1}$$

The relationship between RMSECV and PRESS is:

$$RMSECV = \sqrt{\frac{PRESS}{\mathrm{n}-1}} \tag{2}$$

The smaller the RMSECV value, the better the predictive ability of the model.Therefore, selecting the smallest principal component corresponding to the minimum value of RMSECV.

3.Using MCCV to establish the PLS model[16,17].In this paper,MSC is used to divide all samples into two parts, 80% of the samples are used as the calibration set to establish partial least squares model, and the remaining 20% of the samples are used as the test set.In order to ensure that each sample is predicted, 1000 PLS models are established in this paper, which can get the prediction residuals of each sample.The basic principle of establishing the PLS model [18] is as follows:

Firstly, the concentration matrix Y (n samples, m components) and the absorbance matrix X (n samples, p wavelength points) were decomposed into the form of feature vector:

$$Y_{\mathrm{n}\times m} = U_{n\times d} \times Q_{d\times m} + F_{n\times m} \tag{3}$$

$$X_{n\times p} = T_{\mathrm{n}\times d} \times P_{d\times p} + E_{\mathrm{n}\times p} \tag{4}$$

The U and T are the concentration characteristic factor matrix and the absorbance characteristic factor matrix, Q is the concentration loading matrix, P is the absorbance loading matrix, F and E respectively are the residual matrix of concentration and the residual matrix of absorbance.

Establishing the regression model between U and T:

$$U = T \times B + E_{\mathrm{d}} \tag{5}$$

In the formula, E is a random error matrix, and B is a diagonal linear regression coefficient matrix. For the samples to be measured, if the absorbance vector is x, then the concentration of y can be solved by the formula:

$$\mathrm{y} = x \times (U \times X)' \times B \times Q \tag{6}$$

4.Calculating the mean and standard deviation of each sample, and making the mean-variance distribution map.According to the observation, we can know that the samples which are located in the region of high average or high standard deviation are most likely to be singular samples.

According to the above algorithm, building the models by two ways.One is the suspected singular samples identified existing in the calibration set and the other is the suspected singular samples identified not existing in the calibration set .By using R, RMSECV and RMSEP to evaluate the suspected samples before and after eliminating singular model ,so as to test whether the identified sample is the singular sample.

3. Datasets

3.1 Sources And Collection Of Data

The experimental material is a sample of the starch provided by the laboratory of a starch processing factory,a total of 700 starch samples,and randomly selecting 600 of them as a calibration set, the remaining 100 starch samples for the test set. The moisture content of starch samples has been measured accurately by the national standard method, and the measured value is used as the reference value in the modeling process.

This experiment using Fourier transform near infrared spectrometer VER-TEX 70, places the starch sample in the sample cup of sample stage of diffuse reflection, to collect the near infrared spectra and sample the large sample cup rotary , the environmental temperature is 23~25℃, scanning times are 64 , the scanning range is 950~1650nm, the spectral resolution is 5nm, and making near infrared diffuse reflection scanning to 700 samples of starch. In order to eliminate the influences of the uniformity of the sample, the every sample was repeated for three times, and then the average value was calculated as the sample spectrum. Near infrared spectra of starch samples are shown in figure 1.

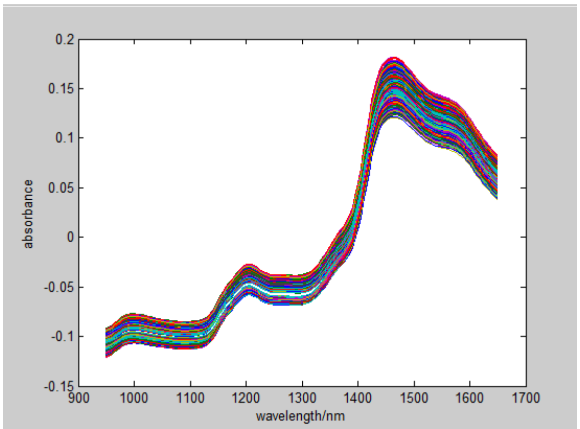

Fig.1 Near infrared spectra of starch samples

3.2 Spectral Data Preprocessing

In addition to containing the dominant spectral information of the tested samples, various interference information also be contained, such as high frequency noise, baseline drift and random error of the instrument. Therefore, it is necessary to preprocess the original spectrum before the correction model is set up. In this experiment, two kinds of methods, the standard and the multiple scatter correction, are used to deal with the spectrum pretreatment.

4. Results And Discussion

In order to verify the effectiveness of the proposed algorithm,two methods were used to pre process the 600 samples of the calibration set, then the PLS algorithm is used to reduce the dimension of the calibration set data and determine the number of principal components, and then using the MCCV method to eliminate singular samples in the calibration set, while building PLS models for water content by using starch samples in the correction set before and after eliminating singular samples,finally ,comparing the stability and forecasting precision of the model before and after eliminating singular samples by using RMSECV, R and RMSEP as evaluation index . Figure 2 and figure 3 respectively show the distribution of the mean and the variance of the residuals in the two pretreatment methods. Figure 3, figure 2 shows that the 538th sample is obviously located in the high standard deviation area, which can be temporarily determined as a suspected singular sample. Then, taking RMSECV, RMSEP and R as evaluation indexes, to compare the stability and predictive ability of the model before and after eliminating the singular samples, so as to determine whether the sample is a singular sample. Results are shown in table 1:

Table 1 Cross validation results of PLS calibration model before and after eliminating sample 538

Pretreatment method	Singular sample	Before removing the singular sample		After removing the singular sample	
		R	RMSECV	R	RMSECV
Standardization	538	0.7069	1.3707	0.8509	1.2675
multiplicative scatter correction	538	0.7332	1.2713	0.8925	1.1633

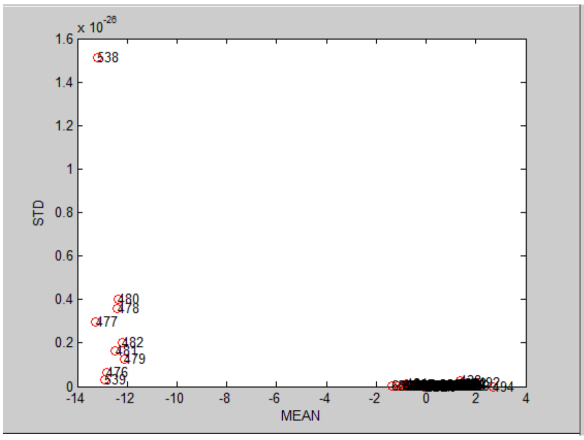

Fig.2 MCCV is used to identify the singular samples in the case of standardization

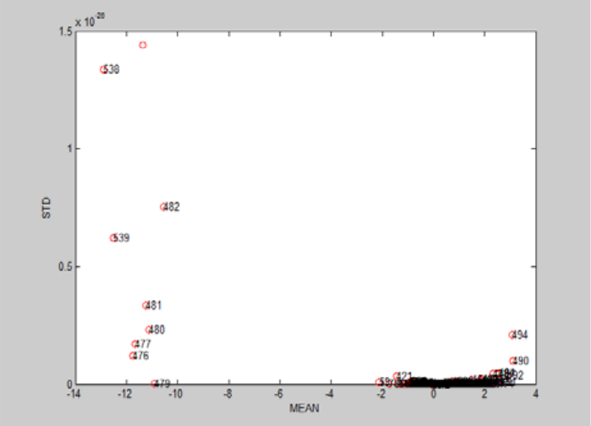

Fig.3 MCCV is used to identify the singular samples in the case of MSC

We can see from table 1, under the condition of using different pretreatment methods, the value of RMSECV of correction model after removing No. 538 sample compared with before eliminating, is significantly decreased, R is significantly increased, which indicates that the removal of No. 538 samples do improve the stability of the model.

In order to verify whether the prediction accuracy of the model after removing No. 538 sample is improved, the 100 samples of the test set are predicted by using the PLS model, and the results are shown in Table 2:

Table 2 Prediction results of PLS calibration model before and after eliminating singular samples

Pretreatment method	Singular sample	Before removing the singular sample		After removing the singular sample	
		R	RMSEP	R	RMSEP
Standardization	538	0.8501	0.7069	0.9013	0.6509
multiplicative scatter correction	538	0.8032	1.2550	0.9128	0.5026

We can see from table 2, under the condition of using different spectral pretreatment methods, the value of RMSECV of correction model after removing No. 538 sample compared with before eliminating, is significantly decreased, R is significantly increased, which indicates that the removal of No. 538 samples effectively enhance the prediction ability of the model.

The experimental results show that: the No. 538 sample of starch samples of calibration set is singular sample, using the MCCV method can quickly and effectively identify the singular samples. The MCCV method is applied to the identification of the singular samples in the near infrared spectrum, which can effectively improve the stability and prediction accuracy of the model.

5. Summary

In this experiment, we use the Monte Carlo cross validation method to effectively eliminate the singular samples in a large number of starch samples, which can solve the problem of identifying the singular samples in the existing methods. The results show that the MCCV method can improve the stability and accuracy of the calibration model. Of course, the current research on the MCCV method is still in depth, the next step is to further optimize the work of the algorithm, in order to make the algorithm can be more accurate to determine the threshold of abnormal samples and the algorithm play the greatest potential in the near infrared spectrum of the quantitative modeling, analysis and application , so as to improve the stability and accuracy of the prediction model.

Reference

[1].Lu Wanzhen,Yuan Hongfu,Au Guangtong,et al.Modern Near Infrared Spectroscopy Analytical Technology[M]. Beijing: China Petrochemical Press, 2000. 146.

[2].Philip Williams, Karl Norris. Near Infrared Technology in the Agriculture and Food Industries. 2nded[M].,American Association of Cereal Chemists, Minnesota USA: AACC, 2001.

[3].Lu Rong, Chen Wenliang, Xu Kexin, et al. Application of rapid detection of singular points in near infrared spectroscopy measurement of milk composition[J].Spectroscopy and Spectral Analysis,2005,02:207-210.

[4].Chu Xiaoli.Molecular Spectroscopy Analytical Technology Combined with Chemometrics and Its Applications[J]. Beijing:Chemical Industry Press,2011,7.

[5].Gowen A , Downey G, Esquerre C, et al. Preventing over-fitting in PLS calibration models of near-infrared (NIR) spectroscopy data using regression coefficients[J]. Chemometr, 2010.

[6].Walczak B. Outlier detection in multivariate calibration[J]. Chemometr Intell Lab Syst, 1995, 28: 259-272.

[7].Paul S R, Fung K Y. A generalized extreme studentized residual multiple outlier detection procedure in linear regression[J]. Technometrics, 1991, 33: 339-348.

[8].Mark H. Use of Mahalanobis distances to evaluate sample preparation methods for near-infrared reflectance analysis[J]. Anal Chem, 1987, 59: 790-795.

[9].de Groot P J, Postma G J, Melssen W J, et al. Application of principal component analysis to detect outliers and spectral deviations in near-field surface-enhanced Raman spectra[J]. Anal Chim Acta, 2001, 446: 71-83.

[10].Gnanadesikan R, Kettenring J R. Robust estimates, residuals, and outlier detection with multiresponse data[J]. Biometrics, 1972, 28: 81-124.

[11].Walczak B, Massart D L. Robust principal components regression as a detection for outliers[J]. Chemometr Intell Lab Syst, 1995, 27: 41-54.

[12].Rousseeuw P J. Multivariate estimation with high breakdown point[J].In mathematical statistics and applications, 1985, pp. 283-297.

[13].Rousseeuw P J, Leroy A M. Robust regression and outlier detection[M]. John Wiley & Sons Inc., New York, 1987, 216-247.

[14].Liu Zhichao,Cai Wensheng,Shao Xueguang. The Monte Carlo cross validation is used for the identification of singular samples in the near infrared spectrum[J].Science China(B: Chemistry),2008,04,316-323.

[15]. Xu Q S, Liang Y Z. Monte Carlo cross validation[J]. Chemometr Intell Lab Syst, 2001, 56(1): 1-11.

[16]. Gourvenec S, Fernandez Pierna J A, Massart D L, et al. An evaluation of the PoLiSh smoothed regression and the Monte Carlo cross-validation for the determination of the complexity of a PLS model[J]. Chemometr Intell Lab Syst, 2003, 68(1): 41-51.

[17].Yan Yanlu, Chen Bin, Zhu Dazhou. Principle,technology and application of near infrared spectroscopy[M]. Beijing:China Light Industry Press,2013.