Multivariate quality control solved by one-class partial least squares regression: Identification of adulterated peanut oils by mid-infrared spectroscopy

ARTICLE *in* JOURNAL OF CHEMOMETRICS · OCTOBER 2011

Impact Factor: 1.5 · DOI: 10.1002/cem.1402

3 AUTHORS, INCLUDING:

Lu Xu
Tongren University
51 PUBLICATIONS **342** CITATIONS
SEE PROFILE

Chen-Bo Cai
Chuxiong Normal University
32 PUBLICATIONS **220** CITATIONS
SEE PROFILE

Research Article

JOURNAL OF CHEMOMETRICS

Received: 22 October 2010, Revised: 15 May 2011, Accepted: 22 May 2011, Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/cem.1402

Multivariate quality control solved by one-class partial least squares regression: identification of adulterated peanut oils by mid-infrared spectroscopy

Lu Xu[a]*, Chen-Bo Cai[b] and De-Hua Deng[a]**

The Partial least squares class model (PLSCM) was recently proposed for multivariate quality control based on a partial least squares (PLS) regression procedure. This paper presents a case study of quality control of peanut oils based on mid-infrared (MIR) spectroscopy and class models, focusing mainly on the following aspects: (i) to explain the meanings of PLSCM components and make comparisons between PLSCM and soft independent modeling of class analogy (SIMCA); (ii) to correct the estimation of the original PLSCM confidence interval by considering a nonzero intercept term for center estimation; (iii) to investigate the potential of MIR spectroscopy combined with class models for identifying peanut oils with low doping concentrations of other edible oils.

It is demonstrated that PLSCM is actually different from the ordinary PLS procedure, but it estimates the class center and class dispersion in the framework of a latent variable projection model. While SIMCA projects the original variables onto a few dimensions explaining most of the data variances, PLSCM components consider simultaneously the explained variances and the compactness of samples belonging to the same class. The analysis results indicate PLSCM is an intuitive and easy-to-use tool to tackle one-class problems and has comparable performance with SIMCA. The advantages of PLSCM might be attributed to the great success and well-established foundations of PLS. For PLSCM, the optimization of model complexity and estimation of decision region can be performed as in multivariate calibration routines. Copyright © 2011 John Wiley & Sons, Ltd.

Keywords: peanut oils; quality control; SIMCA; partial least squares class model; mid-infrared spectrometry

1. INTRODUCTION

The Peanut (*Arachis hypogaea* L.) ranks fourth among oilseed crops in the world and is grown widely in tropical, subtropical, and temperate climates [1]. Peanut seeds contain 45–55% oil, and more than half of the global crop is used as an oilseed. China is one of the most important producers of peanuts, where the production of peanuts exceeds 14 000 000 tons per year, just second to the annual production of soybeans [2]. Pure peanut oil is pale yellow and nondrying with high contents of arachidonic, oleic, linoleic, palmitic, and stearic acids, as well as low concentrations of behenic and lignoceric acids [3]. In China, pure and authentic peanut oil is traditionally thought as a high-quality vegetable oil and serves as a major source of edible and cooking oil. Unfortunately, in the domestic market, pure peanut oil is sometimes adulterated with certain cheaper vegetable oils, such as rapeseed oil, soybean oil, corn oil, palm oil, salad oil, and so on. Therefore, it is necessary to develop reliable and quick analytical methods for discriminating pure peanut oils from various adulterated products.

For food control, the combination of spectroscopy (such as near infrared (NIR) [4–6] and mid-infrared (MIR) [7–9] spectroscopy) and chemometric methods provides a promising alternative approach to the traditional methods based on chemical analysis and sensory analysis [10–12]. In such investigations, chemical compositions of the samples are characterized by the measured multivariate spectra, and then multivariate calibration and/or pattern recognition methods are used to extract information concerning food quality. Some advantages of spectroscopy analysis are as follows: (i) it requires little or no sample preparations; (ii) the analysis time is significantly shortened compared with chemical analysis, so it is very suitable to analyze batch samples; (iii) it is a nondestructive analysis method and can be used for online analysis. Among various spectroscopic methods, NIR spectroscopy might be the most frequently used technique for noninvasive and quick analysis of food products, but recently, MIR spectroscopy has been increasingly used for the same purposes [7,8]. MIR covers the region between 4000 and 400 cm^{-1} and can be broadly segmented into four regions [9]: 4000–2500 cm^{-1} (X-H stretching region), 2500–2000 cm^{-1} (triple bond region), 2000–1500 cm^{-1}

* *Correspondence to: Lu Xu, College of Chemistry and Chemical Engineering, Anyang Normal University, Anyang 455000, Henan, China. E-mail: lxchemo@163.com*

** *Correspondence to: De-Hua Deng, College of Chemistry and Chemical Engineering, Anyang Normal University, Anyang 455000, Henan, China. E-mail: ddh@aynu.edu.cn*

a *L. Xu, D.-H. Deng*
College of Chemistry and Chemical Engineering, Anyang Normal University, Anyang 455000, Henan, China

b *C.-B. Cai*
Department of Chemistry and Life Science, Chuxiong Normal University Chuxiong 675000, China

(double bond region), and 1500–400 cm^{-1} (fingerprint region). Characteristic absorption bands are associated with major components of food, which is the basis of MIR analysis.

To tackle the task of identifying doped peanut oils, class modeling techniques (CMTs) [13–16] are required to answer a general question of whether a new object should be accepted or rejected by a class of interest (e.g., pure and authentic oils). This question is typical of many practical problems, such as the traceability of protected denomination of origin foods and multivariate quality control in an industrial process [11,12]. For such problems, only class A is studied, and the samples rejected by class A are generally not well defined, or in other words, one cannot expect the rejected samples come from a predetermined class (for instance, class B or C). As pointed out in Ref. [15], classification methods that classify samples into two or more categories known beforehand are often improperly used in food quality control. Therefore, this paper will focus on CMTs. Some of the most commonly used CMTs include [15] the following: (i) soft independent modeling of class analogy (SIMCA) [17] based on principal component analysis (PCA); (ii) unequal dispersed classes [18] based on the hypothesis of multivariate normal distribution and the Hotelling T^2 statistics; (iii) those methods based on potential functions estimating the multivariate probability distribution of training samples [19–21].

In a recent paper [22], a partial least squares (PLS) regression [23,24] was proposed to build a partial least squares class model (PLSCM), where the CMT problems are shown to be easily solved by multivariate calibration routines. In this paper, PLSCM is applied to the identification of doped peanut oils by MIR spectrometry. The focuses are fixed on the following aspect (i) to explain the meanings of PLSCM components and make comparisons between PLSCM and SIMCA; (ii) to correct the estimation of the original PLSCM confidence interval by considering a nonzero intercept term for center estimation; (iii) to improve the original PLSCM in determining model complexity by performing an *F*-test [25,26] of the prediction error sum of squares (PRESS) obtained by Monte Carlo cross-validation (MCCV) [27,28], in which this procedure can reduce the risk of selecting too many components and including uncorrelated data variances when characterizing a class; and (iv) to investigate the potential of MIR spectroscopy combined with class models for identifying peanut oils with low doping concentrations of other edible oils. More details of the methods and results will be presented in the following parts of the paper.

2. METHODS

2.1. Soft independent modeling of class analogy

Soft independent modeling of class analogy [17] is by far the most well-known method for describing the class structure of a data set in chemometrics. In SIMCA, PCA is performed for different classes, and the significant principal components (PCs) are used to describe each class. A new sample is target tested by the class models and is then accepted or rejected according to the estimated confidence intervals. SIMCA has also been used for identification of an objective class (one-class problems), where it works as an outlier test [29].

For one-class problems, SIMCA starts by determining the number of PCs to describe the structure of the training samples. The number of PCs is often selected by cross-validation [30,31]. For samples from class A, the column-centered training data matrix $\mathbf{X}_A(n\times p)$ containing n samples characterized by p features can be explained by the r primary PCs:

$$\overline{\mathbf{X}}_A = \overline{\mathbf{U}}\,\overline{\mathbf{S}}\,\overline{\mathbf{V}}^T \tag{1}$$

where $\overline{\mathbf{X}}_A$ is the reconstructed training data and $\overline{\mathbf{U}}$, $\overline{\mathbf{S}}$, and $\overline{\mathbf{V}}$ are the same as in common PCA, only with the secondary PCs truncated. The superscript "T" means transpose of a matrix. The unexplained residuals of the training data $\mathbf{X}_A$ are assumed to have a normal distribution, and its standard deviation (s_A) can be calculated as follows:

$$s_A = \sqrt{\sum_{k=1}^{n}\sum_{j=1}^{p} e_{kj}^2/[(l-r)(n-r-1)]} \tag{2}$$

where e_{kj} is the difference between the elements in the kth row and jth column of $\mathbf{X}_A$ and $\overline{\mathbf{X}}_A$, respectively; l is the minimum of $n-1$ and p; and s_A can also be calculated using PCs:

$$s_A = \sqrt{\sum_{k=1}^{n}\sum_{j=r+1}^{l} t_{kj}^2/[(l-r)(n-r-1)]} \tag{3}$$

where t_{kj} is the jth PC score of kth training sample. Geometrically, s_A can be seen as a measure of the Euclidean distances from the training samples in class A to the space spanned by r significant PCs.

Based on the estimated s_A, the confidence limit for class A can be derived by introducing a critical value of the Euclidean distance to the mentioned PC space, which can be expressed as follows:

$$s_{crit} = \sqrt{F_{crit}s_A^2} \tag{4}$$

where F_{crit} is the one-sided value of *F*-test with the degrees of freedom $l-r$ and $(l-r)(n-r-1)$.

For a new sample, the distance (s_{un}) of the unknown sample from the objective class A can be calculated as follows:

$$s_{un} = \sqrt{\sum_{j=r+1}^{l} t_{un,j}^2/(l-r)} \tag{5}$$

where $t_{un,j}$ is the jth PC score for the unknown sample. If s_{un} is less than s_{crit}, the unknown sample is accepted by class A; otherwise, it is rejected.

It is known that SIMCA can lead to a large number of objects that are wrongly rejected (a large α-error) [29]. Some authors [32–35] propose different strategies to overcome the given problems, but the different degrees of freedom and correction procedures suggest the problem is still to be solved. Moreover, the estimation of model complexity to describe a class is not straightforward, and the decision results are largely influenced by it.

2.2. Partial least squares class model

Partial least square has been widely used in various fields of chemometrics. As a key method in chemometrics, its statistical properties have been extensively studied and its foundations are well established. In a recent paper [22], a PLS procedure is proposed to develop a class model. With training data $\mathbf{X}$ $(n\times p)$ containing n representative objects with p features

characterizing a class, PLSCM performs the following latent variable regression procedure:

$$\mathbf{1} = \mathbf{Xb} + \mathbf{e} \tag{6}$$

where **1** is an $n \times 1$ response vector with all the elements being ones, **b** is the vector of regression coefficients, and **e** the vector of model errors. It should be highlighted that in PLSCM, **X** should not be column centered; otherwise, all the columns will be orthogonal to the response vector.

As suggested in the original paper, **b** is deduced as in a usual latent variable model:

$$\begin{aligned} \mathbf{T} &= \mathbf{XW} \\ \mathbf{1} &= \mathbf{Tq} + \mathbf{e} \\ \mathbf{b} &= \mathbf{Wq} \end{aligned} \tag{7}$$

where **T** is a matrix with columns containing *A* orthogonal latent variables, **W** is a matrix with columns containing loadings, and **q** is a vector of regression coefficients relating **T** and the response vector **1**.

The first latent variable $\mathbf{t}_1 = \mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{1}$, the *i*th ($i = 2 \sim A$) latent variable $\mathbf{t}_i$ can be computed as $X_i X_i^{\mathrm{T}}\mathbf{1}$, where $\mathbf{X}_i = (\mathbf{I} - \mathbf{T}_{i-1}\mathbf{T}_{i-1}^{+})\mathbf{X}$ is the projection of X onto the complementary space spanned by the first *i-1* latent variables.

For a class model, the variance or standard deviation of prediction errors **e** in equation (6) will be a measure of the sample discreteness in a class and can be used to reject or accept a new object. The original paper assumes **e** has a normal distribution with a mean of zero and an estimated standard deviation $\hat{\sigma}$. For a given significance level α, the interval of predicted response value for accepting a new sample is as follows:

$$1 - z_{1-\alpha/2} \cdot \hat{\sigma} < \hat{y}_{\mathrm{un}} < 1 + z_{1-\alpha/2} \cdot \hat{\sigma} \tag{8}$$

where $z_{1-\alpha/2}$ is the critical value of standard normal distribution and $\hat{y}_{\mathrm{un}}$ the predicted response of a new sample. Although the original paper assumes the prediction error has a mean of zero, in this paper, we propose to estimate the mean of **e** by MCCV and correct the original estimation of class confidence internal as follows:

$$1 - \hat{\mu}_e - z_{1-\alpha/2} \cdot \hat{\sigma} < \hat{y}_{\mathrm{un}} < 1 - \hat{\mu}_e + z_{1-\alpha/2} \cdot \hat{\sigma} \tag{9}$$

where $\hat{\mu}_e$ can be estimated form MCCV:

$$\hat{\mu}_e = \mathrm{mean}(\mathbf{e}_{\mathrm{MCCV}}) \tag{10}$$

where $\mathbf{e}_{\mathrm{MCCV}}$ is a vector containing the prediction errors of all the left-out samples during MCCV.

A problem is how to estimate $\hat{\sigma}$. Seen from equation (6), the value of root mean square error of calibration (RMSEC) can be used to estimate $\hat{\sigma}$. However, because the training samples have been already used in training the class model, RMSEC value tends to underestimate the prediction errors and wrongly reject more samples [29]. Therefore, the prediction errors obtained by MCCV are used to estimate $\hat{\sigma}$:

$$\hat{\sigma} = \sqrt{\sum_{i=1}^{N} \left(1 - \hat{y}_i - \hat{\mu}_e\right)^2 / N - 1} \tag{11}$$

where *N* is the total number of left-out samples in multiple resampling process of MCCV and $\hat{y}_i$ is predicted response of a left-out sample.

Another crucial problem is the selection of PLS components in PLSCM. Selecting too few latent variables will fail to characterize the class sufficiently, whereas models with too many latent variables will include the class-uncorrelated data variances and have a bad prediction performance. Therefore, the predicted residual sum of squares (PRESS) obtained by MCCV are subject to a well-established *F*-test to select the proper number of components [25,26]. This procedure selects the fewest PLS components that have a PRESS value not significantly larger than the minimum PRESS value.

The vector **1** used as a response vector means all the objects in the same class should be distributed as close to each other as possible. Intuitively, PLSCM projects the high-dimensional data onto a subspace where all the training samples belonging to the same class are compact. By the mentioned PLS regression, the regression coefficients are shrunk to stabilize the model variance. Some authors have also compared the shrinkage by kernel PLS regression with regularization by support vector machines with kernel functions, and the two approaches are found to be equally effective [36,37]. So, because both PLSCM and support vector data description (SVDD) [38] optimize the closeness of a class, in a sense, PLSCM is also similar to SVDD devoted to one-class problems. However, this paper will focus on comparing the performances of SIMCA and PLSCM, so the comparison of PLSCM and SVDD is beyond the scope of this paper and can be discussed elsewhere. While SIMCA projects the data onto a few PCs explaining most of the data variances, PLSCM considers both the explained variances and compactness of a class. As mentioned previously, without normalization, the loading weights for a PLSCM latent variable are as follows:

$$\mathbf{w}_i = \mathbf{X}_i^{\mathrm{T}}\mathbf{1} \tag{12}$$

Actually, the weights for the PLSCM latent variables are the same as the mean sample spectrum except for a scaling factor. The scores in $\mathbf{t}_i = \mathbf{X}_i\mathbf{w}_i$ can be generally seen as the projection lengths of the training samples onto the mean sample as shown in Figure 1. Because the within-class samples are similar to the mean sample, the angle between a mean spectrum and a training spectrum is limited, which means large projection lengths or a considerable amount of explained variances of **X** will be achieved by PLSCM projection. Meanwhile, the training samples in the same class are similar to each other, so their projections onto the mean sample should be distributed compactly. Essentially, PLSCM seems to make a reasonable compromise between explained data variances and within-class closeness.

3. EXPERIMENTAL

3.1. Preparation of authentic and doped peanut oil samples

A set of 22 pure and authentic peanut oils of different batches manufactured by Shandong Luhua Group Co., Ltd, Yantai, China is purchased from domestic markets. The peanut oils are manufactured in some major producing areas of peanut, including Shandong (8), Henan (6), and Jiangsu (8). To represent and simulate the samples from other producing areas, we prepared another 68 pure peanut oil samples by sufficient blending of the mentioned 22 raw samples with different ratios. Therefore, a total of 90 samples of pure pressing peanut oil are prepared for MIR analysis. All the pure peanut oils are extracted by pressing and stored in a cool, dark area before spectrometry analysis.

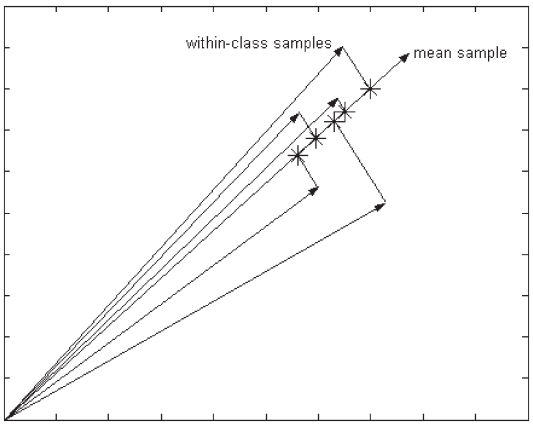

Figure 1. The geometric meaning of a partial least squares class model latent variable.

Eighteen samples of rapeseed oil (4), soybean oil (3), corn oil (2), palm oil (3), and salad oil (6) are collected from domestic markets. Then 110 doped peanut oil samples are prepared by mixing the pure peanut oils with different contents of the given oils ranging from 3% to 90%. A list of the adulterated peanut oil samples are shown in Table I. The doped samples are stored under the same conditions as the pure peanut oil samples.

3.2. Mid-infrared spectrometric analysis

The MIR transmission spectra are measured in the range of 4000 and 400 cm^{-1} on a Nicolet 380 infrared spectrophotometer with a DTGS KBr detector by Thermo Fisher Scientific Inc., Waltham, USA. No preprocessing of oil samples is performed, and the MIR spectra of all the oil samples are measured in a KBr demountable absorption cell without any solvents. The resolution is 4 cm^{-1}, and the scanning interval is 1.929 cm^{-1}. The scanning time is set to be 64, because a larger scanning time cannot significantly improve the quality of spectral data. Some of the raw MIR spectra are demonstrated in Figure 2.

4. RESULTS AND DISCUSSIONS

Seen from Figure 2, the absorption band in 3000–2800 cm^{-1} can be attributed to the stretching vibration of -CH, and the band in 1740–1680 cm^{-1} is likely to be caused by stretching vibration of C=O in carboxyl group and C=C in unsaturated fatty acids. The absorption bands in fingerprint region are more difficult to explain, the band around 1460 cm^{-1} might be the asymmetric bending of $-CH_2$, the band around 1200 cm^{-1} can be attributed to the vibration of carbon skeleton, and the band around 720 cm^{-1} can be caused by the rocking or wagging of $-CH_2-$ in a long carbon chain $-(CH_2)_n-$. Although the MIR spectra of pure peanut oils are obviously different from those of other vegetable oils, the difference between the spectra of pure and

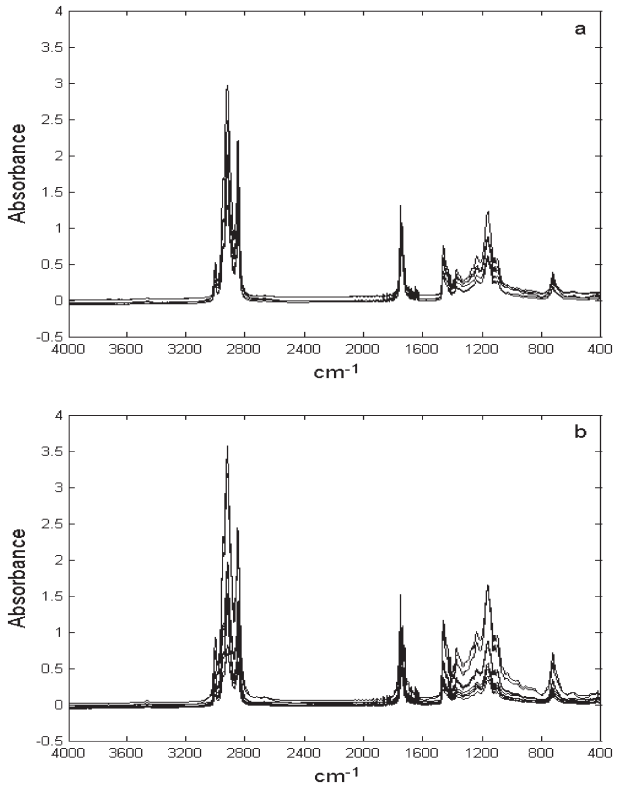

Figure 2. Some of the raw mid-infrared spectra of (a) pure and (b) adulterated peanut oils with doping concentrations ranging from 3% to 90%.

adulterated peanut oils becomes very subtle with low doping concentrations. Therefore, chemometric class models are necessary to extract the useful information from spectral data for characterizing pure peanut oils. The significance level for class models is set to be 0.05.

With the 90 pure peanut oil samples, robust PCA [39] is performed and no outliers are detected. The algorithm by Kennard and Stone [40] is then used to form a representative training set of 70 samples and test set (test set 1) of 20 samples. The aim of this algorithm is to select a training set in such way that the objects are scattered uniformly around the training samples. For both robust PCA and Kennard–Stone algorithms, the codes included in the widely distributed toolbox TOMCAT [41] are used. The 110 doped peanut oil samples are used as a test set (test set 2).

For the 70 training samples, the first two PCs account for 85.3% of the total data variances. For a SIMCA model, the decision region proposed in Ref. [29] is used to reduce the risk of having large number of objects wrongly rejected. By using scores predicted by leave-one-out cross-validation rather than the original scores obtained after PCA on the class objects, this procedure inflates the within-class component variances and is

Table I. Adulterated peanut oils with different levels of doping concentrations

Doping levels	3%	5%	10%	20%	30%	40%	50%	60%	70%	80%	90%
Number of samples	10	10	10	10	10	10	10	10	10	10	10

shown to lead to a reduction of the number of false outliers. To determine the number of significant PCs, we used different methods based on factor indicator function (IND), residual standard deviation (RSD), and cumulative percentage variance [42], and these give somewhat different results. Because the first six PCs explain 96.8% of the data variances, for a fair comparison, the results of SIMCA models with four to six PCs are reported and listed in Table II. SIMCA model with six PCs seems to provide the best training and predicting results as shown in Figure 3. The wrongly predicted samples for test set 1 (containing 20 pure peanut oils) and test set 2 (containing 110 doped peanut oils) are 1 and 4, respectively. In Figure 3(c), the samples in test set 2 are arranged according to an ascending doping concentration. Seen from Figure 3(c), the higher the doping concentrations, the higher the predicted *s* values above the critical value. Moreover, the doping levels of four wrongly accepted doped oils are 3%, which seems to be the lowest doping concentration that can be detected by SIMCA.

For PLSCM, MCCV with 20% left-out samples is used to determine the number of PLS components and the sampling time is 100. The PRESS values by MCCV are subject to the *F*-test proposed by Refs [25] and [26]. As suggested, a significance of 0.25 is used. The lowest PRESS value is obtained by nine PLS components, and PLSCM with six latent variables obtains a PRESS value not significantly larger than the minimum value according to the *F*-test. The results of PLSCM are also listed in Table II. Figure 4 demonstrates the results obtained by PLSCM with six latent variables. The results are similar to those of SIMCA, as the predicted response values become farther from 1 with the increasing of doping concentrations. The concentrations of three wrongly accepted samples are also 3%. The six PLS latent variables account for 84.9% of the data variances. For this data set, the estimated mean of **e** is 0.023, so the center of PLSCM confidence interval is just slightly deviated from 1. Although the correction of confidence interval in equation (9) makes little difference in rejecting or accepting a sample compared with equation (8) for the current problem, we believe this correction is necessary especially when **e** has a larger mean value.

The results indicate that the performance of PLSCM is comparable with that of SIMCA. Because PLSCM can be performed as in the routines of multivariate calibration, it seems to be easier to use in terms of determining model complexity and decision region.

Table II. The numbers of wrongly predicted samples by soft independent modeling of class analogy and partial least squares class model

Models	Training set (70[a])	Test set 1 (20)	Test set 2 (110)
SIMCA(4[b])	6	2	8
SIMCA(5)	5	1	5
SIMCA(6)	2	1	4
PLSCM(6)	2	0	3

SIMCA, soft independent modeling of class analogy; PLSCM, partial least squares class model.
[a]Total number of samples in a data set.
[b]The number of components or latent variables in the model.

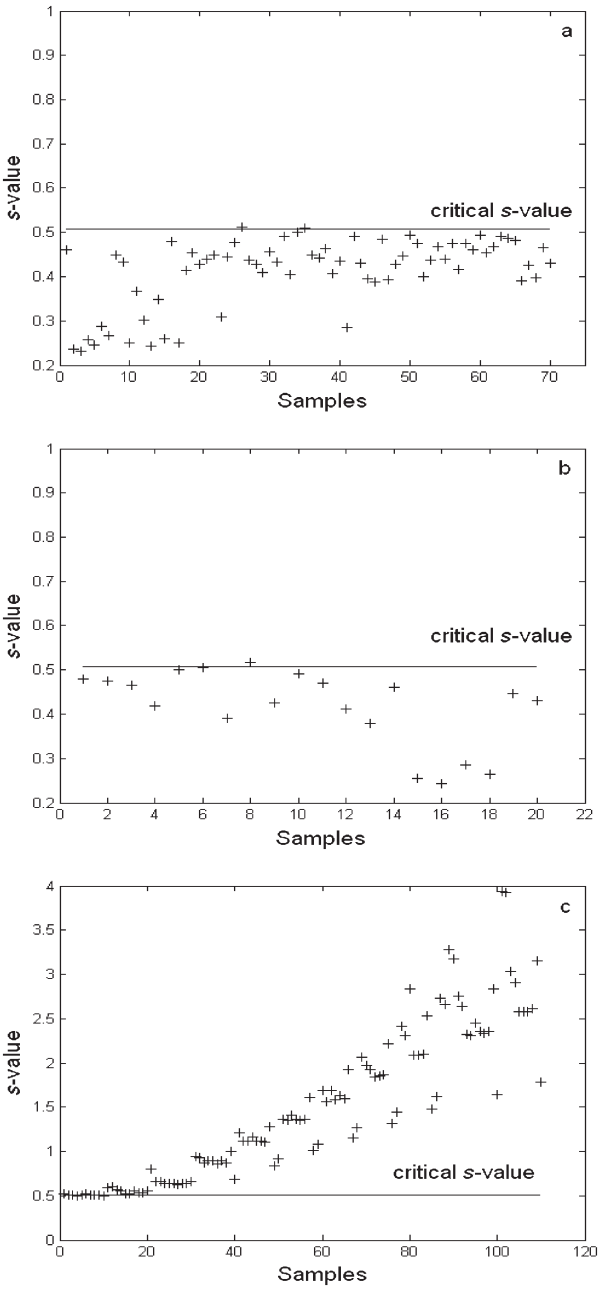

Figure 3. The results by soft independent modeling of class analogy with six principal components for (a) training, (b) predictions of test set 1, and (c) predictions of test set 2.

5. CONCLUSIONS

The identification of doped peanut oils by MIR spectrometry is tackled by PLSCM and SIMCA. The results demonstrate PLSCM has comparable performance with SIMCA. For outlier detection, when the doping concentration is as low as 3%, both methods wrongly accept some new samples. Because PLSCM can be performed in the framework of multivariate calibration, determining model complexity and decision region for PLSCM seems more straightforward than for SIMCA. Moreover, for a more accurate estimation of PLSCM confidence interval, a correction to the original PLSCM is made to estimate the mean of

Journal of CHEMOMETRICS

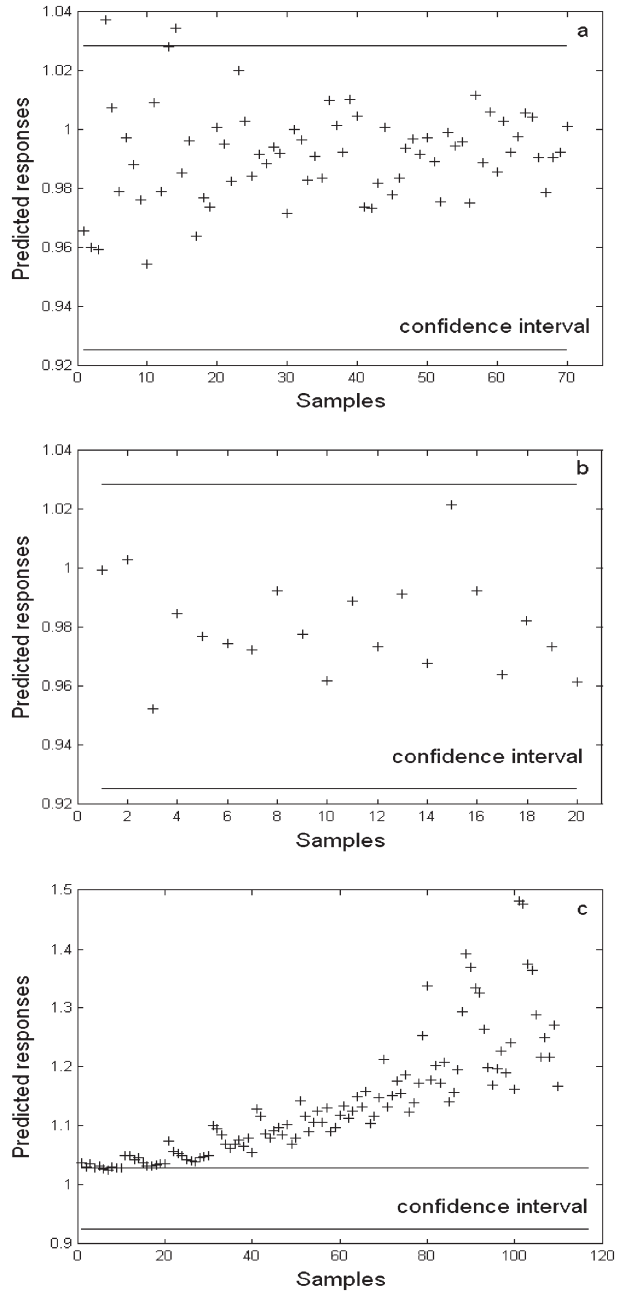

Figure 4. The results by partial least squares class model with six components for (a) training, (b) predictions of test set 1, and (c) predictions of test set 2.

model errors by MCCV resampling. If PLSCM has some advantages, they should be attributed to the well-established foundations of PLS.

PLSCM estimates the class center and class dispersion in the framework of a latent variable projection model, where the estimation of regression is shrunk to stabilize model variance. The intuitive geometric meanings of PLSCM are briefly discussed. By projecting the sample spectra onto the mean spectrum, PLSCM latent variables consider both the explained variances and within-class closeness. This feature seems to be of interest when the objective is to characterize a single class.

Acknowledgements

This work is financially supported by the Basic Research Plans on Natural Science of the Education Department of Henan Province (Nos. 2008A150001 and 2011A430001).

REFERENCES

1. Jung S, Swift D, Sengoku E, Patel M, Teulé F, Powell G, Moore K, Abbott A. The high oleate trait in the cultivated peanut (*Arachis hypogaea* L.): I. Isolation and characterization of two genes encoding microsomal oleoyl-PC desaturases. *Mol. Gen. Genet.* 2000; **263**: 796–805.
2. Jiang L, Hua D, Wang Z, Xu S. Aqueous enzymatic extraction of peanut oil and protein hydrolysates. *Food Bioprod. Process.* 2010; **88**: 233–238.
3. Oyinlola A, Ojo A, Adekoya LO. Development of a laboratory model screw press for peanut oil expression. *J. Food Eng.* 2004; **64**: 221–227.
4. Alishahi A, Farahmand H, Prieto N, Cozzolino D. Identification of transgenic foods using NIR spectroscopy: a review. *Spectrochim. Acta, Part A* 2010; **75**: 1–7.
5. Huang H, Yu H, Xu H, Ying Y. Near infrared spectroscopy for on/in-line monitoring of quality in foods and beverages: a review. *J. Food Eng.* 2008; **87**: 303–313.
6. Toher D, Downey G, Murphy TB. A comparison of model-based and regression classification techniques applied to near infrared spectroscopic data in food authentication studies. *Chemom. Intell. Lab. Syst.* 2007; **89**: 102–115.
7. Downey G. Food and food ingredient authentication by mid-infrared spectroscopy and chemometrics. *Trends Anal. Chem.* 1998; **17**: 418–424.
8. Karoui R, Downey G, Blecker C. Mid-infrared spectroscopy coupled with chemometrics: a tool for the analysis of intact food systems and the exploration of their molecular structure-quality relationships—a review. *Chem. Rev.* 2010; **110**: 6144–6168.
9. Stuart BH. *Infrared spectroscopy: fundamentals and applications*. John Wiley & Sons: Chichester, UK, 2004; 137–165.
10. Munck L, Nørgaard L, Engelsen SB, Bro R, Andersson CA. Chemometrics in food science—a demonstration of the feasibility of a highly exploratory, inductive evaluation strategy of fundamental scientific significance. *Chemom. Intell. Lab. Syst.* 1998; **44**: 31–60.
11. Marini F, Magrì AL, Bucci R, Balestrieri F, Marini D. Class-modeling techniques in the authentication of Italian oils from Sicily with a protected denomination of origin (PDO). *Chemom. Intell. Lab. Syst.* 2006; **80**:140–149.
12. Forina M, Oliveri P, Jäger H, Römisch U, Smeyers-Verbeke J. Class modeling techniques in the control of the geographical origin of wines. *Chemom. Intell. Lab. Syst.* 2009; **99**: 127–137.
13. Derde MP, Massart DL. Comparison of the performance of the class modelling techniques UNEQ, SIMCA, and PRIMA. *Chemom. Intell. Lab. Syst.* 1988; **4**: 65–93.
14. Frank IE, Friedman JH. Classification: oldtimers and newcomers. *J. Chemometrics* 1989; **3**: 463–475.
15. Forina M, Oliveri P, Lanteri S, Casale M. Class-modeling techniques, classic and new, for old and new problems. *Chemom. Intell. Lab. Syst.* 2008; **93**: 132–148.
16. Li D, Lloyd GR, Duncan JC, Brereton RG. Disjoint hard models for classification. *J. Chemometrics* 2010; **24**: 273–287.
17. Wold S, Sjöström M. In *Chemometrics: Theory and Applications*, ACS Symposium Series, vol. **52**, Kowalski BR (ed.). American Chemical Society: Washington, USA, 1977; 243–282.
18. Hotelling H. In *Techniques of Statistical Analysis*, Eisenhart C, Hastay MW, Wallis WA (eds.). McGraw-Hill: N.Y., 1947; 111–184.
19. Rosenblatt M. Remarks on some nonparametric estimates of a density function. *Ann. Math. Stat.* 1956; **27**: 832–837.
20. Coomans D, Broeckaert I. *Potential Pattern Recognition in Chemical and Medical Decision Making*, Research Studies Press: Letchworth, UK, 1986.
21. Forina M, Armanino C, Leardi R, Drava G. A class-modelling technique based on potential functions. *J. Chemometrics* 1991; **5**: 435–453.
22. Xu L, Fu HY, Jiang N, Yu XP. A new class model based on partial least square regression and its applications for identifying authenticity of bezoar samples. *Chin. J. Anal. Chem.* 2010; **38**: 175–180.

23. Wold H. Soft modeling: the basic design and some extensions. In *Systems Under Indirect Observation* (Vols. **I and II**), Jöreskog K-G, Wold H (eds.). North-Holland: Amsterdam, 1982.
24. Wold S, Johansson E, Cocchi M. PLS—partial least squares projections to latent structures. In *3D QSAR in Drug Design, Theory, Methods, and Applications*, Kubinyi H (ed.). ESCOM Science Publishers: Leiden, 1993; 523 – 550.
25. Haaland DM, Thomas EV. Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information. *Anal. Chem.* 1988; **60**: 1193 – 1202.
26. Haaland DM, Thomas EV. Partial least-squares methods for spectral analyses. 2. Application to simulated and glass spectral data. *Anal. Chem.* **60** (1988) 1202 – 1208.
27. Xu QS, Liang YZ. Monte Carlo cross validation. *Chemom. Intell. Lab. Syst.* 2001;**56**: 1 – 11.
28. Xu QS, Liang YZ, Du YP. Monte Carlo cross-validation for selecting a model and estimating the prediction error in multivariate calibration. *J. Chemometrics* 2004; **18**: 112 – 120.
29. De Maesschalck R, Candolfi A, Massart DL, Heuerding S. Decision criteria for soft independent modelling of class analogy applied to near infrared data. *Chemom. Intell. Lab. Syst.* 1999; **47**: 65 – 77.
30. Wold S. Pattern recognition by disjoint principal components models. *Pattern Recogn.* 1976; **8**: 127 – 139.
31. Bro R, Kjeldahl K, Smilde AK, Kiers HAL. Cross-validation of component models: a critical look at current methods. *Anal. Bioanal. Chem.* 2008; **390**: 1241 – 1251.
32. Wold S, Sjöström M. Comments on a recent evaluation of the SIMCA method. *J. Chemometrics* 1987; **1**: 243 – 245.
33. Albano C, Blomqvist G, Coomans D, Dunn III WJ, Edlund U, Eliasson B, Hellberg S, Johansson E, Nordén B, Joknels D, Sjöström M, Söderström B, Wold H, Wold S. Pattern recognition by means of disjoint principal components models SIMCA. Philosophy and Methods, Proceedings of the Symposium on Applied Statistics, Copenhagen, Jan. 22, 1981.
34. Kvalheim OM, Oygard K, Grahl-Nielsen O. SIMCA multivariate data analysis of blue mussel components in environmental pollution studies. *Anal. Chim. Acta* 1983; **150**: 145 – 152.
35. Gemperline PJ, Webber LD, Cox FO. Raw materials testing using soft independent modeling of class analogy analysis of near-infrared reflectance spectra. *Anal. Chem.* 1989; **61**: 138 – 144.
36. Li HD, Liang YZ, Xu QS. Support vector machines and its applications in chemistry. *Chemom. Intell. Lab. Syst.* 2009; **95**:188 – 198.
37. Czekaj T, Wu W, Walczak B. About kernel latent variable approaches and SVM. *J. Chemometrics* 2005; **19**: 341 – 354.
38. Tax DMJ, Duin RPW. Support vector data description. *Mach. Learn.* 2004; **54**: 45 – 66.
39. Rousseeuw PJ, Croux C. Alternatives to the median absolute deviation. *J. Am. Stat. Assoc.* 1993; **88**: 1273 – 1283.
40. Kennard RW, Stone LA. Computer aided design of experiments. *Technometrics* 1969; **11**: 137 – 148.
41. Daszykowski M, Serneels S, Kaczmarek K, Van Espen P, Croux C, Walczak B. TOMCAT: a MATLAB toolbox for multivariate calibration techniques. *Chemom. Intell. Lab. Syst.* 2007; **85**:269 – 277.
42. Malinowski ER. *Factor Analysis in Chemistry* (3rd edn). Wiley: New York, USA, 2002.