JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY

ARTICLE

pubs.acs.org/JAFC

Predicting the Age and Type of Tuocha Tea by Fourier Transform Infrared Spectroscopy and Chemometric Data Analysis

Lu Xu,*[†] De-Hua Deng,*[†] and Chen-Bo Cai[‡]

[†]College of Chemistry and Chemical Engineering, Anyang Normal University, Anyang 455002, People's Republic of China

[‡]Department of Chemistry and Life Science, Chuxiong Normal University, Chuxiong 675000, People's Republic of China

ABSTRACT: Fourier transform infrared (FTIR) spectroscopy combined with chemometric multivariate methods was proposed to discriminate the type (unfermented and fermented) and predict the age of tuocha tea. Transmittance FTIR spectra ranging from 400 to 4000 cm^{-1} of 80 fermented and 98 unfermented tea samples from Yunnan province of China were measured. Sample preparation involved finely grinding tea samples and formation of thin KBr disks (under 120 kg/cm^2 for 5 min). For data analysis, partial least-squares (PLS) discriminant analysis (PLSDA) was applied to discriminate unfermented and fermented teas. The sensitivity and specificity of PLSDA with first-derivative spectra were 93 and 96%, respectively. Multivariate calibration models were developed to predict the age of fermented and unfermented teas. Different options of data preprocessing and calibration models were investigated. Whereas linear PLS based on standard normal variate (SNV) spectra was adequate for modeling the age of unfermented tea samples (RMSEP = 1.47 months), a nonlinear back-propagation-artificial neutral network was required for calibrating the age of fermented tea (RMSEP = 1.67 months with second-derivative spectra). For type discrimination and calibration of tea age, SNV and derivative preprocessing played an important role in reducing the spectral variations caused by scattering effects and baseline shifts.

KEYWORDS: brick tea, age, FTIR, multivariate calibration, pattern recognition

INTRODUCTION

As one of the most popular beverages in the world, tea (*Camellia sinensis* L.) has also been regarded as a natural medicine for over 4000 years (since arguably as early as 2700 BC).[1] Tea gains its popularity because of its pleasurable aroma, taste, and putative healthy effects.[2,3] Tea plants (*C. sinensis*) are widely distributed in over 30 countries and play a significant role in their economies. According to the degree of fermentation, teas can be generally classified into three major types: unfermented green teas, partially fermented or semifermented oolong and paochong teas, and fully fermented black and pu-erh (red) teas.[4] According to morphological and chemical diversities, some authors also suggest teas should be classified into five principal varieties.[5]

The quality and chemical composition of teas depend on various factors, including species, season, age of the leaves (plucking position), climate, and horticultural conditions (soil, water, minerals, fertilizers, etc.).[6] Numerous papers in the literature have been devoted to investigations of the chemical compositions of teas influenced by the above-mentioned factors.[7−14] Such studies are crucial for understanding the biological and pharmaceutical properties of various teas but usually lack a comprehensive view of chemical compositions. In traditional sensory analysis, the quality of teas is evaluated by professional tea tasters.[15] Because the process of training a skilled tea taster may take years and is very expensive, it would be attractive to evaluate tea quality by some nonhuman techniques.

As a promising alternative approach to the traditional methods of chemical and sensory analysis, the combination of spectrometry and chemometric methods[16,17] has been widely used in food analysis. The rationale behind such techniques is that chemical compositions of samples are characterized by measured multivariate spectra; useful information concerning tea quality can be extracted by multivariate calibration and/or pattern recognition methods. Some advantages of spectrometry analysis include the following: (1) it requires no or less sample preparation; (2) the analysis time and cost are largely reduced compared with chemical analysis, so it is very suitable to analyze batch samples; (3) it is a nondestructive or noninvasive analysis method and can be used potentially for online analysis. Among various spectroscopic methods, near-infrared (NIR) spectroscopy is the most popular for noninvasive analysis of food products, but recently reported applications of mid-infrared (MIR) in food analysis have significantly increased.[17]

Tuocha is a compressed brick tea produced in the Yunnan province of China and consumed in large quantities in Central Asia, southwestern China, and other areas. It has been used as a food (mainly in parts of Central Asia and Tibet) and beverage as well as a folk medicine.[18] Tuocha can be blocks of whole or finely ground black tea, green tea, or postfermented tea leaves that have been packed in molds and pressed into block form. Whereas fermented and unfermented tuocha teas have different flavors and health effects, it is also recognized that the quality increases with age,[19] in contrast to green tea, which is unfermented and consumed as fresh as possible. Moreover, during the long storage of brick teas, both fermented and unfermented teas undergo some complicated chemical changes, so age is an important feature of the effects and flavors of brick tea.[20−26]

Received: July 4, 2011
Revised: September 5, 2011
Accepted: September 7, 2011
Published: September 07, 2011

This research is aimed to develop a precise and reliable model to predict the age and type of tuocha by Fourier transform infrared (FTIR) spectroscopy and chemometrics. Both fermented and unfermented tuocha samples of known age were collected as samples for training the prediction model. For data mining, partial least-squares discriminant analysis (PLSDA)[27] was applied to discriminate different types of tuocha. Linear partial least-squares regression (PLSR)[28] and nonlinear back-propagation-artificial neural network (BP-ANN)[29] were performed to relate the age of tea samples to the measured FTIR spectra. To remove undesirable factors in the raw data, different strategies of data preprocessing including smoothing,[30] first- and second-order derivatives,[30] standard normal variate (SNV),[31] and detrend[31] were also investigated. For different analytical objectives, given the performances of different models are similar or have no significant differences, the models with least complexity and least preprocessing were sought to ensure the generalization of models.

MATERIALS AND METHODS

Teas. Eighty fermented and 98 unfermented tuocha samples were analyzed. The tea samples were obtained from the market branch of Xiaguan Tuocha Group Co., Ltd. (Dali, Yunnan, China). All of the tea samples retained integral packaging and the original labels indicating detailed sample information. By the time of analysis, the age of samples ranged from 51 to 6 months and from 42 to 3 months for fermented and unfermented teas, respectively. All of the samples are made of green tea leaves. The detailed information concerning samples is shown in Table 1. All of the samples were stored in a cool, dark, and dry area with integral packaging before spectrometry analysis.

FTIR Spectroscopy. Sample preparation involved finely grinding tea samples followed by preparation of KBr pellets. Samples were manually ground into fine particles using an agate pestle and mortar. Then, 10 mg (1:30 w/w) of each powder sample was mixed with 290 mg (29:30 w/w) of KBr (Sigma Chemical Co., St. Louis, MO). KBr pellets were prepared by exerting a pressure of 120 kg/cm^2 for approximately 5 min in a pellet press (Tuopu Instrument., Tianjin, China). To examine whether the variation in pellet thickness cause significant interference in the measured spectra, different pellets were prepared from the same sample and their FTIR spectra were compared.[32] The measured FTIR spectra were nearly identical to their average spectrum used for analysis.

FTIR spectra were collected using a Nicolet 380 FTIR spectrometer (Thermo Scientific, Waltham, MA) in the wavelength range of 400—4000 cm^{-1}. For each pellet, 64 scans were performed with a resolution of 4 cm^{-1} at room temperature using OMNIC software. An increase in scanning time did not significantly improve the signal. The average of the 64 scans was used as a raw spectrum for further data analysis. The scanning interval was 1.929 cm^{-1}; therefore, each spectrum contained 1868 individual points for chemometric analysis.

Preprocessing and Outlier Detection. The performance and reliability of chemometric analysis depend largely on proper implementation of data preprocessing when the measured spectra are subject to significant noises, baselines, and other undesirable factors. Although various preprocessing methods have been developed, it is well-known that preprocessing not only can improve certain qualities of the spectra but also is likely to degrade the data in certain other aspects.[33] Considering the lack of sufficient prior information concerning the measured spectra, different options were investigated to optimize data pretreatment.

Smoothing was frequently used to remove part of the random noise present in the signal and enhance the signal-to-noise ratio (SNR). The algorithm of polynomial fitting by the Savitzky and Golay (S-G) method[30] was applied for this purpose because of its popularity and simplicity. Taking derivatives can enhance spectral differences and remove baseline and background, so first and second derivatives were also adopted. Because derivatives tend to decrease the SNR by enhancing noise, the derivative spectra were computed by S-G algorithms.[30] SNV[31] was originally designed to reduce scattering effects in the spectra but was also proved to be effective in correcting the interference caused by variations in pellet thickness or optical path. Although the influence of the thickness of pellets was found to be insignificant in this work, SNV was performed to reduce the possible variations caused by scattering effects or uneven mixing of KBr and tea powders. To further reduce the spectral variations in baseline shifts and curvilinearity caused by powdered or densely packed samples, detrending with a second-degree polynomial[31] was used after SNV transformation.

To avoid the masking effects in outlier detection, robust principal component analysis (rPCA)[34] was performed to detect outliers in the

Table 1. Analyzed Tea Samples

code	brand	age (months)	class[a]
1F (4[b])	Xiaguan tuocha	51	F
2F (5)	Canger tuocha	49	F
3F (5)	Pu-erh tuocha	46	F
4F (3)	Xiaguan Jincha[c]	44	F
5F (5)	Xiaguan tuocha	44	F
6F (4)	Yunnan Qizi Bing	40	F
7F (4)	Pu-erh tuocha	35	F
8F (3)	Xiaguan Jincha[c]	33	F
9F (5)	Xiaguan tuocha	33	F
10F (6)	Yunnan Qizi Bing	29	F
11F (4)	Pu-erh tuocha	27	F
12F (5)	Xiaguan tuocha	23	F
13F (5)	Canger tuocha	18	F
14F (4)	Yunnan Qizi Bing	14	F
15F (5)	Xiaguan Jincha[c]	14	F
16F (5)	Pu-erh tuocha	12	F
17F (4)	Yunnan Qizi Bing	8	F
18F (4)	Xiaguan tuocha	6	F
1N (6)	Canger tuocha	42	N
2N (3)	Xiaguan Brick[c]	42	N
3N (6)	Xiaguan tuocha	38	N
4N (7)	Xiaguan tuocha	33	N
5N (4)	Canger tuocha	32	N
6N (6)	Xiaguan tuocha	28	N
7N (4)	Xiaguan Jincha[c]	27	N
8N (4)	Xiaguan tuocha	25	N
9N (4)	Xiaguan Brick[c]	24	N
10N (4)	Canger tuocha	24	N
11N (6)	Xiaguan tuocha	20	N
12N (5)	Xiaguan Jincha[c]	18	N
13N (7)	Canger tuocha	18	N
14N (5)	Xiaguan tuocha	16	N
15N (5)	Xiaguan Jincha[c]	16	N
16N (5)	Xiaguan tuocha	13	N
17N (6)	Canger tuocha	8	N
18N (3)	Xiaguan Brick[c]	8	N
19N (4)	Xiaguan tuocha	5	N
20N (4)	Xiaguan tuocha	3	N

[a] F, fermented teas; N, unfermented teas. [b] Sample size of teas of different batches of the same age. [c] Teas excluded from model training used solely for prediction.

original data set. The algorithm involved the centering by the L^1 median, the stepwise search for orthogonal directions, the use of the Q_n estimator, and the search in the direction of the data points. Because the FTIR spectral data were high-dimensional (for the raw spectra, $p = 1868$), the improved version by Hubert et al.[35] was adopted, which was more numerically stable for high-dimensional data and had a much lower computational cost. In terms of the computed score distance (SD) and orthogonal distance (OD), an rPCA diagnosis plot yields a classification of the samples into four groups: regular data (with small SD and small OD), good PCA-leverage points (with large SD and small OD), orthogonal outliers (with small SD and large OD), and bad PCA-leverage points (with large SD and large OD).

Multivariate Analysis. The Kennard and Stone (K-S) algorithm[36] was used to split the measured spectra sets into a training set and test set. The aim of this algorithm was to select a representative training set in such a way that the objects are scattered uniformly in the range of training samples. Because the distributions of fermented and unfermented teas were not the same, the K-S method was performed separately for the fermented and nonfermented teas. For pattern recognition, the training sets for fermented and nonfermented teas were combined to form a new training set. PLSDA was applied to the discrimination of nonfermented and fermented teas. For discriminant analysis, sensitivity and specificity were used to evaluate the performance of classification models,[37] which are defined as

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

and

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

where TP, FN, TN, and FP denote the numbers of true positives, false negatives, true negatives, and false positives, respectively.

To model the relationship between tea age and FTIR spectra, linear PLS and nonlinear BP-ANN were performed. For a PLS model, a crucial problem is the determination of the number of PLS components or latent variables. It is well-known that selecting too few latent variables is insufficient to explain the response variable, whereas models with too much complexity will include the y-uncorrelated data variances and have a bad prediction performance. Therefore, the F test method proposed by Haaland and Thomas[38,39] combined with Monte Carlo cross-validation (MCCV)[40] was applied to estimating model complexity. The F test based on MCCV was employed as follows. First, random sampling of the training set was performed with a given percent of left-out samples, and then a PLS model with a given model complexity was built on the selected samples to predict the left-out samples. Second, step 1 was repeated for B ($B = 100$ in this paper) times, and the pooled predicted residual sum of squares (PRESS) value was computed. Third, steps 1 and 2 were repeated to obtain the PRESS values for PLS models with different numbers of PLS components. Finally, the F test of each PRESS value was performed, and the fewest PLS components with a PRESS value not significantly larger than the minimum PRESS value were selected.

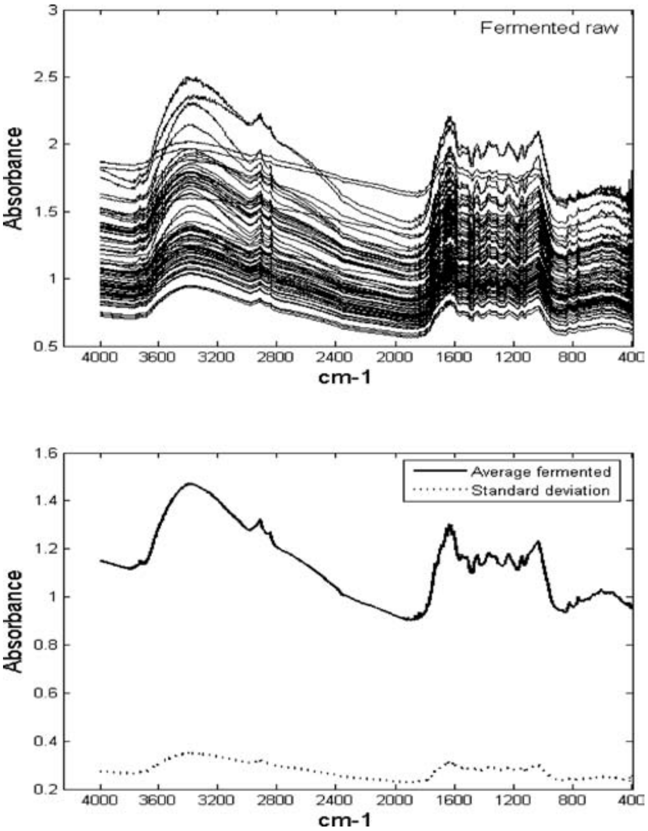

Figure 1. Average and standard deviation raw spectra of 80 fermented tuocha teas.

A major concern with ANN is that it tends to be overfitted; therefore, the structure and parameters of BP-ANN should be carefully optimized. Because a three-layer ANN with one hidden layer is sufficient to simulate most nonlinear relationships and because the instability of the network increases with the growth of hidden layers, a three-layer BP-ANN model was developed. The optimization of BP-ANN parameters is as follows. First, the input variables were rescaled. The original variables were transformed such that the input variables ranged from 0 to 1:

$$x_i = 0.8\,\frac{v_i - v_i^{\min}}{\text{range }(v_i)} + 0.1$$

v_i is the original value and $v_i^{\min}$ the minimum value of the ith predictor variable. Second, the number of nodes in the hidden layer was determined. Too few hidden nodes might be inadequate to reflect the complex relationship between predictor variables and response variables, whereas too many hidden nodes require more computation time and training samples and tend to incorporate a great deal of noise and overlapping information into the model. In this paper, the following experiential rule was adopted to determine the number of hidden nodes:

$$h \geq \sqrt{p + N_o}$$

h is the number of hidden nodes, p is the number of input nodes, and N_o is the number of the output nodes (which equals the number of response variables). With this regulation, the candidate values of h could be tried one-by-one until the best one was found in terms of the squared error loss function. Third, a log-sigmoid function was chosen as the transfer function for the hidden layer because of its desirable nonlinearity and PURELIN linear transfer function for the output layer due to its stability. Fourth, the learning rate was adjusted. A high learning rate can speed the training of ANN and reduce the risk of obtaining a local minimum, but it is also likely to make the network oscillatory and nonconvergent. Therefore, the learning rate was in the range of 0.05−0.8 and adjusted according to the error function. Finally, to reduce the complexity of BP-ANN, PLS components rather than original variables were used as input variables. This has some advantages in the case of the "large p, small n" problem: (1) reducing the size of the network; (2) increasing the calculation speed; (3) reducing the risk of overfitting; (4) denoising by omitting some trivial PLS latent variables and maintaining the stability of the neural network.

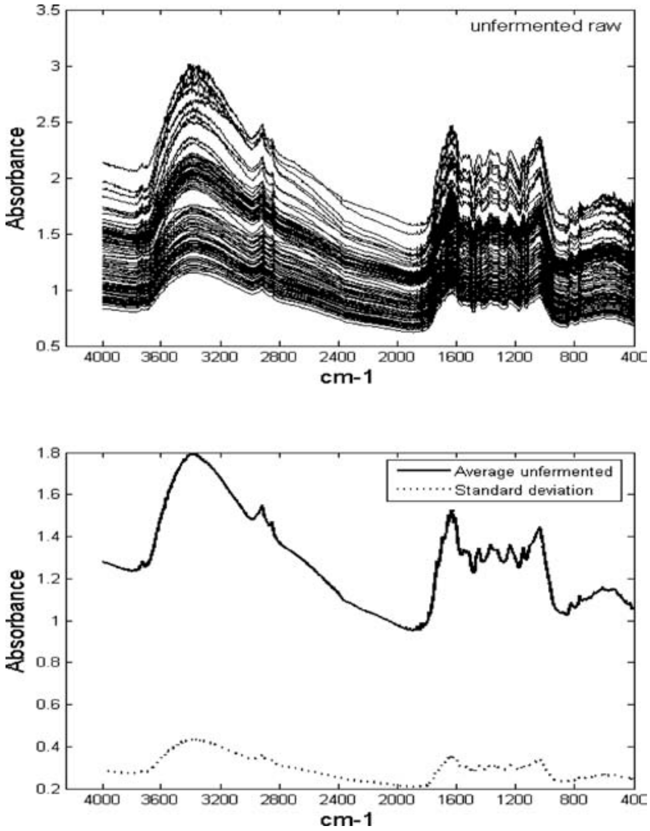

Figure 2. Average and standard deviation raw spectra of 98 unfermented tuocha teas.

RESULTS AND DISCUSSION

All of the data analyses were performed on Matlab 7.0.1 (Mathworks, Sherborn, MA). The toolbox LIBRA for robust analysis[41] was used to perform rPCA diagnosis.

Spectral Data. The average absorption spectra and standard deviation spectra of the 80 fermented and 98 unfermented teas are shown in Figures 1 and 2, respectively. For both types of teas, the highest absorbance in the average spectrum is around 1.5 absorbance units, well within the linear photometric range of the instrument detector.[42] The spectra had broad absorbance bands and were contaminated with significant baselines, so explanation and attribution of bands were very difficult. Standard deviation spectra reflect the variance contributions of different wavelengths. By comparison of the standard deviation spectra with the mean spectra, the spectra of unfermented teas demonstrate a stronger linear relationship than those of fermented teas. This could be attributed to the more complex composition changes of fermented teas.[20−23] Although the spectra of unfermented and fermented teas have similar absorbance bands, the relative intensities and positions of absorbance are slightly different. In Table 1, a brand of fermented (Xiaguan Jincha, 11 samples) and two brands of unfermented teas (Xiaguan Jincha, 14 samples; Xiaguan Brick, 10 samples) were purposely left out to form independent samples for testing calibration models. Therefore, the above 3 brands (35 samples in all) were completely excluded from model training.

Outlier detection was performed on the raw spectra by rPCA. The diagnosis plots are demonstrated in Figure 3. The significance level was 0.05, and the number of principal components (PCs) was evaluated by robust PRESS values. Orthogonal outliers (with small SD and large OD) and bad PCA-leverage points (with large SD and large OD) were excluded from discrimination and calibration models. Because there might be considerable spectral difference between teas with a large age gap, good PCA-leverage points (with large SD and small OD) were reserved to maintain the representativeness of training objects and a wide linear range of calibration models. Five (four orthogonal outliers and one bad PCA-leverage point) and six outliers (four orthogonal outliers and two bad PCA-leverage points) were detected for fermented and unfermented teas, respectively. The K-S algorithm was then applied to splitting the remaining 64 fermented and 68 unfermented teas into training and prediction samples. The training/prediction set contains 48/16 samples for fermented teas and 51/17 samples for unfermented teas. Because the left-out samples were considered solely for prediction, the final test sets contained 27 samples for fermented teas and 41 for unfermented teas. The raw spectra were pretreated by S-G smoothing, first and second S-G derivatives, SNV, and

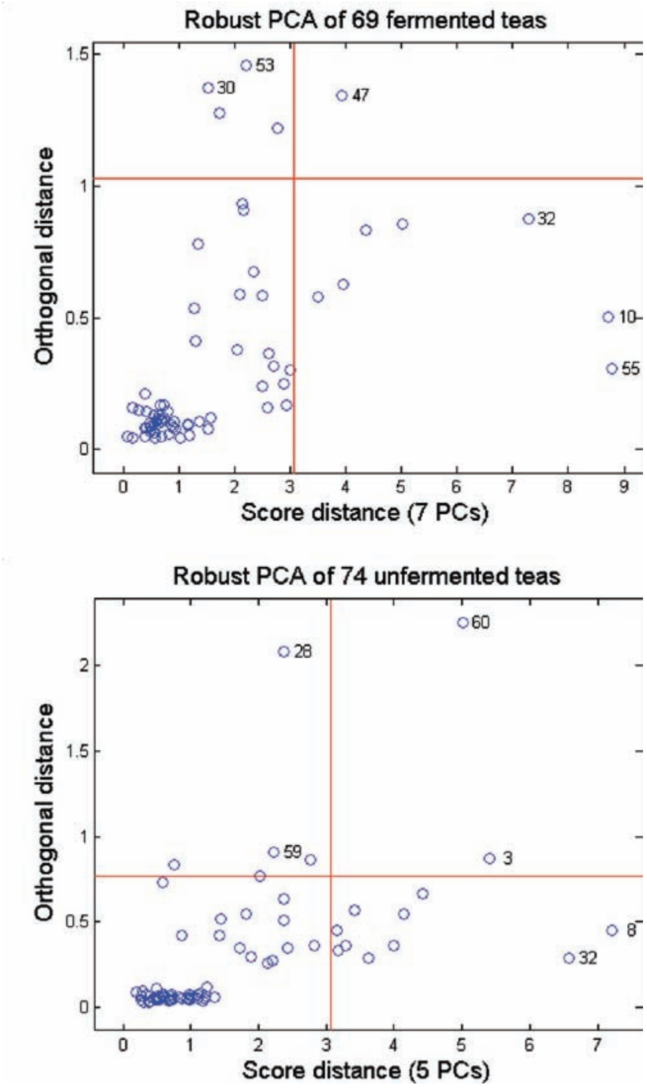

Figure 3. Robust PCA outlier diagnosis plots for fermented and unfermented teas. The red lines segment the samples into four classes: regular data (small SD, small OD), good PCA-leverage points (large SD, small OD), orthogonal outliers (small SD, large OD), and bad PCA-leverage points (large SD, large OD).

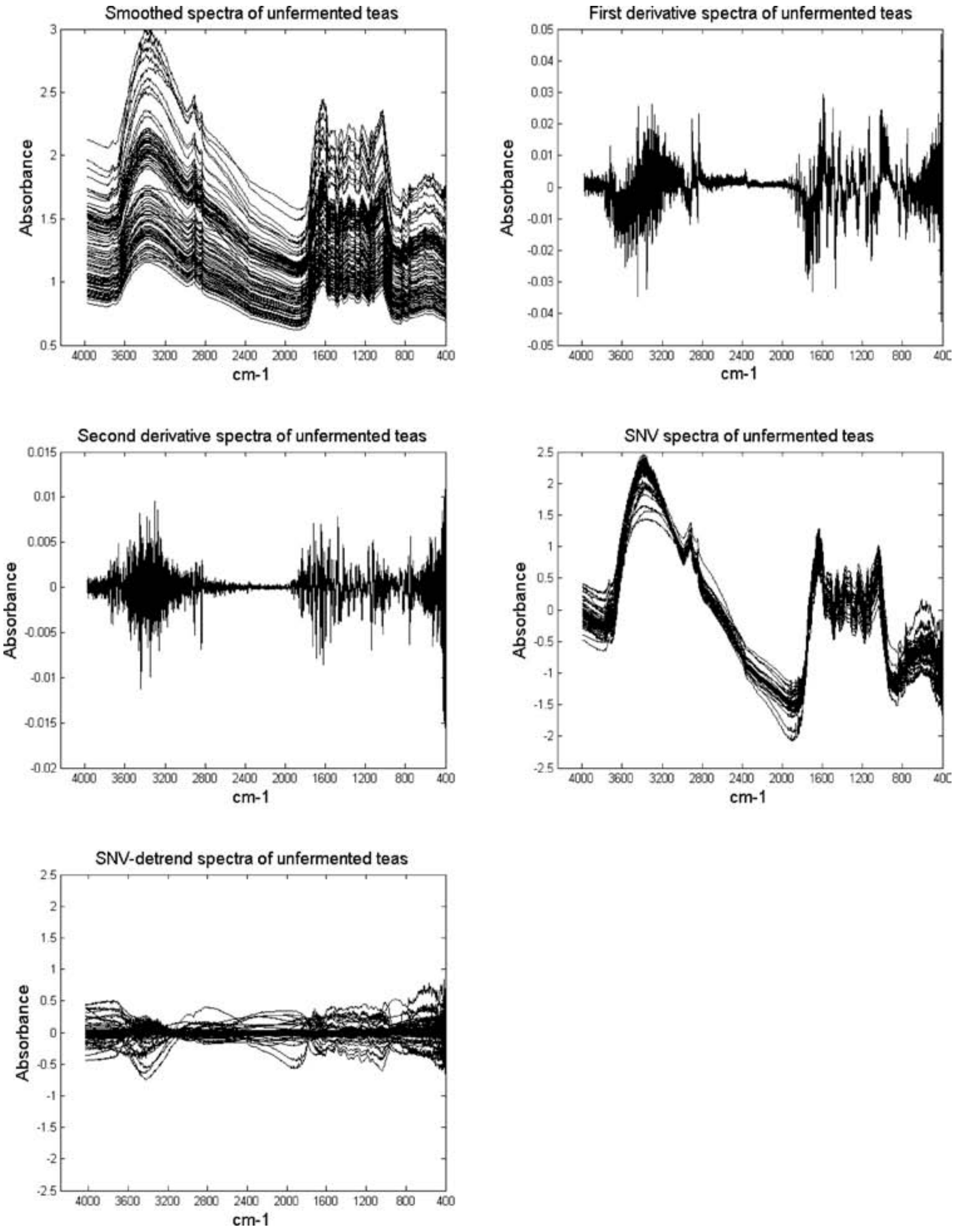

Figure 4. Smoothed, first-derivative, second-derivative, SNV, and SNV-detrend spectra of 98 unfermented teas.

SNV-detrend, and the preprocessed data are demonstrated in Figures 4 and 5.

Modeling of Age. Although nonlinear models are more flexible and accurate to model complex nonlinear relationships, they usually have less stability and generalization ability and tend to be overfitted. Although preprocessing can improve some aspects of data and models, because it treats the data with a presupposed model, it can also bring uncertainty to the predictions of new data. Therefore, the objective of tea age calibration is to develop models with less model complexity and preprocessing given the difference in model performances is not significant. For fermented and unfermented teas, linear PLS models were developed to relate the FTIR spectra to the age of teas. The number of PLS components was determined by *F* test of MCCV. Root mean squared error of prediction (RMSEP) of the test samples was used to evaluate the accuracy of calibration models.

The results of different models and preprocessing techniques are demonstrated in Table 2. The most effective and economic models for modeling the age of teas in this study are highlighted in bold. For unfermented teas, PLS models based on SNV (RMSEP = 1.47) and SNV-detrend spectra (RMSEP = 1.54) obtained the best predictions for test samples. As seen from Figure 4 and Table 2, preprocessing by taking derivatives and SNV can in general reduce the model complexity and enhance model accuracy by removing the baseline variations and backgrounds. By examining the differences between the RMSEC and RMSEP values, SNV and SNV-detrend spectra were more stable for predicting new samples compared with second-derivative spectra.

For fermented teas, the best linear PLS models based on SNV-detrend and second-derivative spectra obtained RMSEP values of 2.20 and 2.38, respectively. Raw, smoothed, and first-derivative

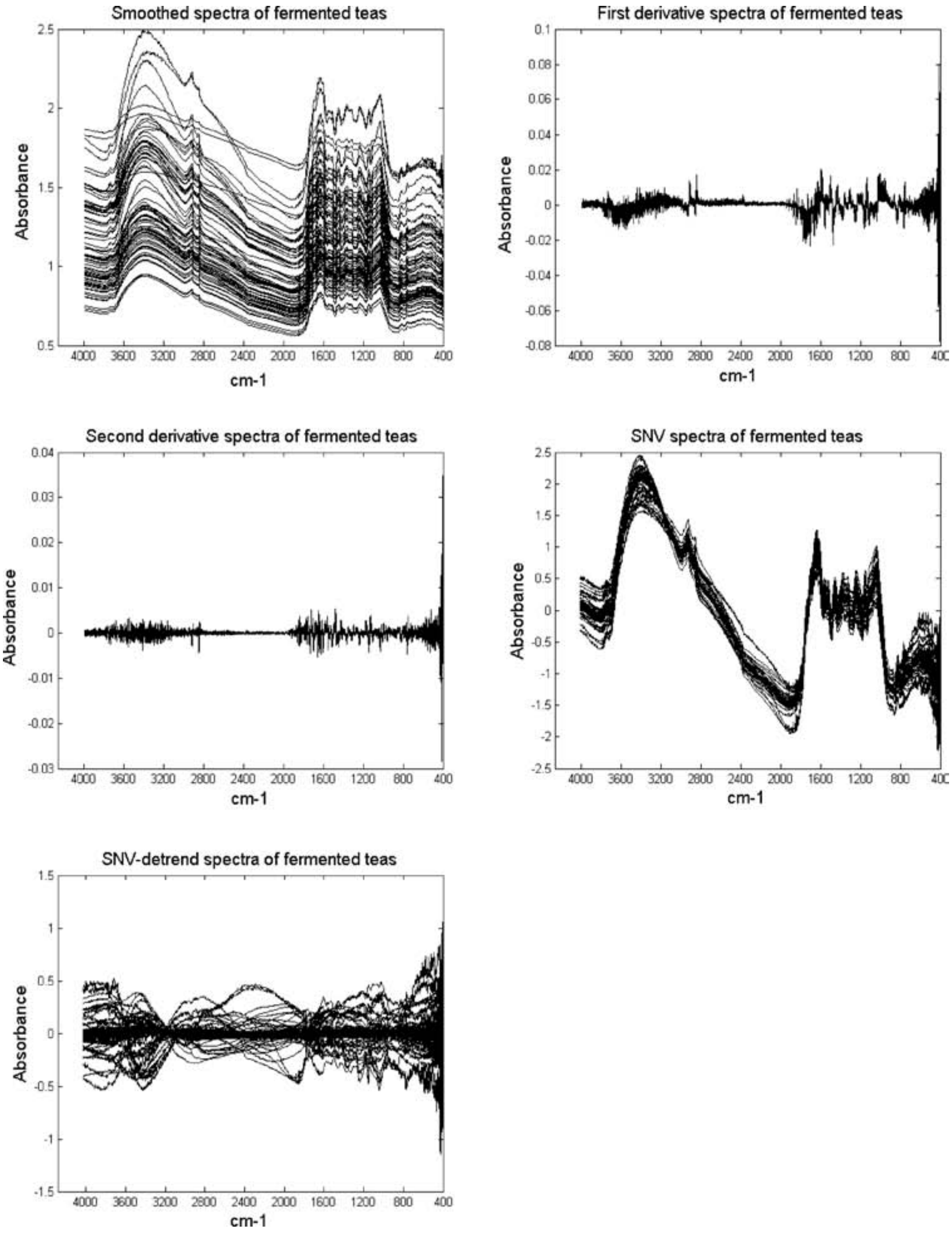

Figure 5. Smoothed, first-derivative, second-derivative, SNV, and SNV-detrend spectra of 80 fermented teas.

spectra achieved inferior calibration accuracy, which is very similar to the unfermented models. This might be partially attributed to the baseline variations. It can also be seen from Figures 4 and 5 that first-derivative spectra still have some baseline variations compared with second derivative. Therefore, nonlinear BP-ANN models were built on second-derivative, SNV, and SNV-detrend spectra. Compared with linear PLS models, BP-ANN had better accuracy. BP-ANN based on second derivative and SNV obtained an RMSEP of 1.67 and 1.95, respectively. To validate the BP-ANN models, cross-validation and multiple initializations of input weights were performed. For BP-ANN, PLS components rather than the original wavelengths were used as input variables, and the number of input nodes was optimized by cross-validation. By cross-validation, BP-ANN models were selected to have a root mean squared error of cross validation (RMSECV) not significantly larger than the root mean squared error of calibration (RMSEC). By multiple initializations, the network was retrained with 200 different initial weights yielding different final weight settings and, thus, different predictions of the test set. The box and whisker plot of RMSEP represents the range of the prediction error and stability of prediction. Figure 6 demonstrates the box and whisker plots of the BP-ANN models based on different preprocessings, indicating the prediction of BP-ANN models was stable. The advantage of BP-ANN models over linear PLS models can be explained by fitting the complexity of chemical composition, as well as the changes during fermentation.[23−26] As seen from the number of PLS latent variables, the PLS models for fermented teas had more model complexity in general than those of unfermented teas, also indicating the greater complexity of spectra and chemical compositions of fermented teas. Correlation plots between the actual and predicted tea ages are shown in Figure 7.

Table 2. Summary of Results for Age Models of Fermented and Unfermented Teas

type	model	pretreatment	RMSEC (months)	RMSEP (months)	LVs[a]
unfermented	PLS	raw data	1.75	2.05	7
		smoothing	1.64	1.78	8
		first derivative	1.69	1.85	7
		second derivative	1.60	1.96	5
		SNV[b]	1.29	1.47	6
		SNV-detrend	1.38	1.54	5
fermented	PLS	raw data	2.27	2.48	9
		smoothing	2.32	2.47	10
		first derivative	2.38	2.55	8
		second derivative	2.03	2.38	7
		SNV	2.19	2.48	7
		SNV-detrend	2.05	2.20	8
	BP-ANN	**second derivative**	1.53	1.67	12[c]
		SNV	1.78	1.95	11
		SNV-detrend	1.90	2.23	11

[a] LVs, number of PLS latent variables. [b] The most effective and economic models for modeling the age of teas in this study are highlighted in bold. [c] Number of PLS components as BP-ANN inputs determined by cross-validation.

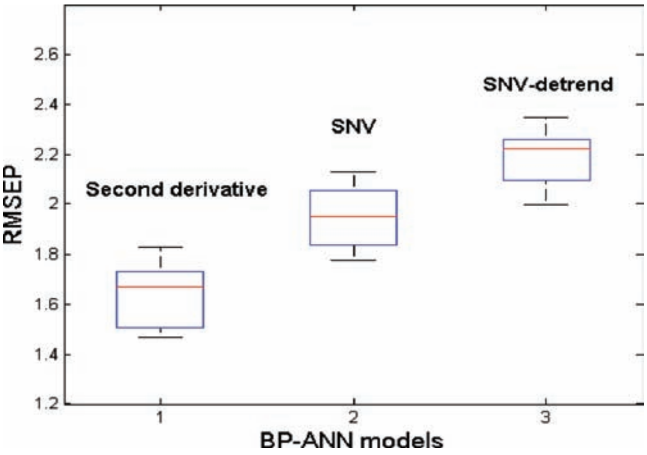

Figure 6. Box and whisker plots of RMSEP obtained for 200 different initial input weights for BP-ANN models with three preprocessing methods. Each plot indicates the minimum, lower quartile, median, upper quartile, and maximum of RMSEP.

Discriminating Fermented and Unfermented. PLSDA was performed to distinguish fermented teas from unfermented. The prediction set contained 41 unfermented plus 27 fermented teas. Sensitivity and specificity of prediction were used to evaluate the classification performance. The unfermented teas were denoted "positives" and the fermented teas as "negatives". With different data preprocessing, the results of PLSDA models in prediction are shown in Table 3. In terms of sensitivity and specificity, preprocessing procedures except smoothing could improve the classification performance and reduce model complexity. Among various data preprocessing techniques, first-derivative (sensitivity = 0.93, specificity = 0.96) and SNV-detrend spectra (sensitivity = 0.95, specificity = 0.93) obtained the most significant improvement on classification power. The results obtained by different preprocessings demonstrate that when the objective is to classify fermented and unfermented teas, the spectral variations caused by scattering and baseline shifts play a more important role than a lower SNR.

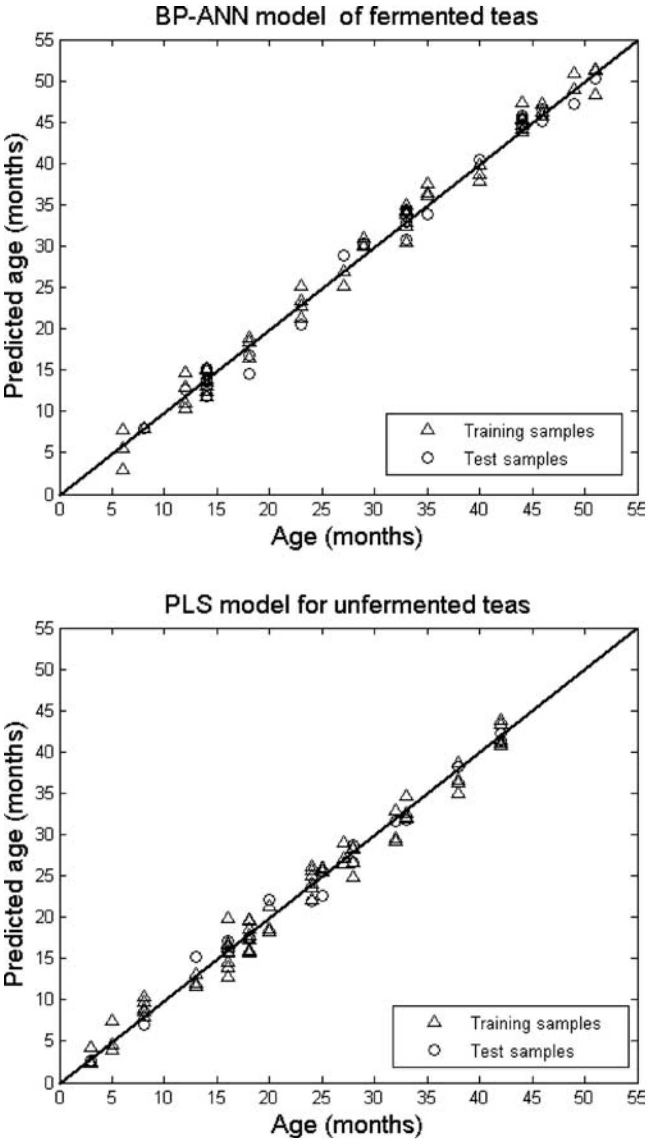

Figure 7. Correlation plots between the actual and predicted ages of teas by the most effective models. BP-ANN model for fermented teas was based on second-derivative spectra, and PLS model for unfermented teas was obtained by SNV spectra.

This research showed that FTIR coupled with chemometrics provided an accurate and practical method to predict the type and age of tuocha teas. Although we can hardly perform an exhaustive sampling of all types of tuocha teas, this study built a good model for type and age authentication of some representative teas. The results are useful for the quality control and routine analysis of tuocha in its market branch. In addition, derivative and SNV were successfully applied to reducing the influence of spectral variations by removing baseline shifts and scattering effects. The combination of FTIR and chemometric analysis would provide an alternative method to the expensive sensory analysis. Reliable predictions were obtained for a fermented tea and two unfermented

Table 3. Results by PLSDA Models with Different Data Preprocessings[a]

data preprocessing	TP	FN	TN	FP	sensitivity	specificity
raw (8[b])	36	5	24	3	0.88	0.89
smoothing (9)	33	8	22	5	0.80	0.81
first derivative (7)[c]	38	3	26	1	0.93	0.96
second derivative (6)	38	3	24	3	0.93	0.89
SNV (6)	37	4	26	1	0.90	0.96
SNV-detrend (7)	39	2	25	2	0.95	0.93

[a] TP, number of true positives; FN, number of false negatives; TN, number of true negatives; FP, number of false positives. Here "unfermented" is positive and "fermented" is negative. [b] Numbers in parentheses indicate numbers of PLSDA latent variables. [c] The most effective and economic models are highlighted in bold.

teas that were similar to but designedly excluded from the calibration samples, indicating good generalization performance of the method. However, caution should be taken when predictions are made for unknown samples of different origins. In such cases, calibration transfer techniques might be required to reduce or correct the bias caused by uncalibrated spectral variations.

AUTHOR INFORMATION

Corresponding Author

*E-mail: (L.X.) lxchemo@163.com; (D.-H.D.) ddh@aynu.edu.cn.

Funding Sources

This work is financially supported by the Basic Research Plans on Natural Science of the Education Department of Henan Province (No. 2008A150001 and 2011A430001).

REFERENCES

(1) McKenzie, J. S.; Jurado, J. M.; de Pablos, F. Characterisation of tea leaves according to their total mineral content by means of probabilistic neural networks. *Food Chem.* **2010**, *123*, 859–864.

(2) Kuroda, Y.; Hara, Y. Antimutagenic and anticarcinogenic activity of tea polyphenols. *Mutat. Res.* **1999**, *436*, 69–97.

(3) Grahan, H. N. The polyphenols of tea — biochemistry and significance — a review. In *Xve Journées Internationales Group Polyphenols*; DTA: Lisbon, Portugal, 1992; Vol. 2, pp 32—43.

(4) Lin, J. K.; Lin, C. L.; Liang, Y. C.; Lin-Shiau, S. Y.; Juan, I. M. Survey of catechins, gallic acid, and methylxanthines in green, oolong, pu-erh, and black teas. *J. Agric. Food Chem.* **1998**, *46*, 3635–3642.

(5) Chen, Q.; Zhao, J.; Fang, C. H.; Wang, D. Feasibility study on identification of green, black and Oolong teas using near-infrared reflectance spectroscopy based on support vector machine (SVM). *Spectrochim. Acta A* **2007**, *66*, 568–574.

(6) Lin, Y. L.; Juan, I. M.; Chen, Y. L.; Liang, Y. C.; Lin, J. K. Composition of polyphenols in fresh tea leaves and associations of their oxygen-radical-absorbing capacity with antiproliferative actions in fibroblast cells. *J. Agric. Food Chem.* **1996**, *44*, 1387–1394.

(7) Shao, W.; Powell, C.; Clifford, M. N. The analysis by HPLC of green, black and pu'er teas produced in Yunnan. *J. Sci. Food Agric.* **1995**, *69*, 535–540.

(8) Goto, T.; Yoshida, Y.; Kiso, M.; Nagashima, H. Simultaneous analysis of individual catechins and caffeine in green tea. *J. Chromatogr., A* **1996**, *749*, 295–299.

(9) Horie, H.; Mukai, T.; Kohata, K. Simultaneous determination of qualitative important components in green tea infusions using capillary electrophoresis. *J. Chromatogr., A* **1997**, *758*, 332–335.

(10) Horie, H.; Kohata, K. Application of capillary electrophoresis to tea quality estimation. *J. Chromatogr., A* **1998**, *802*, 219–223.

(11) Arce, L.; Ríos, A.; Valcírcel, M. Determination of anticarcinogenic polyphenols present in green tea using capillary electrophoresis coupled to a flow injection system. *J. Chromatogr., A* **1998**, *827*, 113–120.

(12) Bronner, W. E.; Beecher, G. R. Method for determining the content of catechins in tea infusions by high-performance liquid chromatography. *J. Chromatogr., A* **1998**, *805*, 137–142.

(13) Wang, H.; Helliwell, K.; You, X. Isocratic elution system for the determination of catechins, caffeine and gallic acid in green tea using HPLC. *Food Chem.* **2000**, *68*, 115–121.

(14) Fernández, P. L.; Martín, M. J.; González, A. G.; Pablos, F. HPLC determination of catechins and caffeine in tea. Differentiation of green, black and instant teas. *Analyst* **2000**, *125*, 421–425.

(15) Gall, G. L.; Colquhoun, I. J.; Defernez, M. Metabolite profiling using ^{1}H NMR spectroscopy for quality assessment of green tea, *Camellia sinensis* (L.). *J. Agric. Food Chem.* **2004**, *52*, 692–700.

(16) Downey, G. Food and food ingredient authentication by mid-infrared spectroscopy and chemometrics. *Trends Anal. Chem.* **1998**, *17*, 418–424.

(17) 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.

(18) Wu, S. C.; Yen, G. C.; Wang, B. S.; Chiu, C. K.; Yen, W. J.; Chang, L. W.; Duh, P. D. Antimutagenic and antimicrobial activities of pu-erh tea. *LWT—Food Sci. Technol.* **2007**, *40*, 506–512.

(19) Sano, M.; Takenaka, Y.; Kojima, R. Effects of pu-erh tea on lipid metabolism in rats. *Chem. Pharm. Bull.* **1986**, *34*, 221–228.

(20) Kuo, K. L.; Weng, M. S.; Chang, C. T.; Tsai, Y. J.; Lin-Shiau, S. Y.; Lin, J. K. Comparative studies on the hypolipidemic and growth suppressive effects of oolong, black, pu-erh, and green tea leaves in rats. *J. Agric. Food Chem.* **2005**, *53*, 480–489.

(21) Xu, X. Q.; Mo, H. Z.; Yan, M. C.; Zhu, Y. Analysis of characteristic aroma of fungal fermented Fuzhuan brick-tea by gas chromatography/mass spectrophotometry. *J. Sci. Food Agric.* **2007**, *87*, 1502–1504.

(22) Qian, Z. M.; Guan, J.; Yang, F. Q.; Li, S. P. Identification and quantification of free radical scavengers in pu-erh tea by HPLCDAD-MS coupled online with 2,2′-azinobis(3-ethylbenzthiazolinesulfonic acid) diammonium salt assay. *J. Agric. Food Chem.* **2008**, *56*, 11187–11191.

(23) Wang, B. S.; Yu, H. M.; Chang, L. W.; Yen, W. J.; Duh, P. D. Protective effects of pu-erh tea on LDL oxidation and nitric oxide generation in macrophage cells. *LWT—Food Sci. Technol.* **2008**, *41*, 1122–1132.

(24) Mo, H. Z.; Zhu, Y.; Chen, Z. M. Microbial fermented tea — a potential source of natural food preservatives. *Trends Food Sci.Technol.* **2008**, *19*, 124–130.

(25) Duh, P. D.; Yen, G. C.; Yen, W. J.; Wang, B. S.; Chang, L. W. Effects of pu-erh tea on oxidative damage and nitric oxide scavenging. *J. Agric. Food Chem.* **2004**, *52*, 8169–8176.

(26) Zhou, Z. H.; Zhang, Y. J.; Xu, M.; Yang, C. R. Puerins A and B, two new 8-C substituted flavan-3-ols from pu-er tea. *J. Agric. Food Chem.* **2005**, *53*, 8614–8617.

(27) Barker, M.; Rayens, M. Partial least squares for discrimination. *J. Chemom.* **2003**, *17*, 166–173.

(28) Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109–130.

(29) Pérez-Marín, D.; Garrido-Varo, A.; Guerrero, J. E.; Gutiérez-Estrada, J. C. Use of artificial neural networks in near-infrared reflectance spectroscopy calibrations for predicting the inclusion percentages of wheat and sunflower meal in compound feedingstuffs. *Appl. Spectrosc.* **2006**, *60*, 1062–1069.

(30) Savitzky, A.; Golay, M. J. E. Smoothing and differentiation of data by simplified least-squares procedures. *Anal. Chem.* **1964**, *36*, 1627–1639.

(31) Barnes, R. J.; Dhanoa, M. S.; Lister, S. J. Standard normal variate transformation and detrending of near infrared diffuse reflectance. *Appl. Spectrosc.* **1989**, *43*, 772–777.

(32) Garip, S.; Gozen, A. C.; Severcan, F. Use of Fourier transform infrared spectroscopy for rapid comparative analysis of *Bacillus* and *Micrococcus* isolates. *Food Chem.* **2009**, *113*, 1301–1307.

(33) Helland, I. S.; Naes, T.; Isaksson, T. Related versions of the multiplicative scatter correction methods for pre-processing spectroscopic data. *Chemometr. Intell. Lab.* **1995**, *29*, 233–241.

(34) Stanimirova, I.; Walczak, B.; Massart, D. L.; Simeonov, V. A comparison between two robust PCA algorithms. *Chemom. Intell. Lab.* **2004**, *71*, 83–95.

(35) Hubert, M.; Rousseeuw, P. J.; Verboven, S. A fast method for robust principal components with applications to chemometrics. *Chemom. Intell. Lab.* **2002**, *60*, 101–111.

(36) Kennard, R. W.; Stone, L. Computer aided design of experiments. *Technometrics* **1969**, *11*, 137–148.

(37) Forina, M.; Armanino, C.; Leardi, R.; Drava, G. A class-modelling technique based on potential functions. *J. Chemom.* **1991**, *5*, 435–453.

(38) Haaland, D. M.; Thomas, E. V. 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.

(39) Haaland, D. M.; Thomas, E. V. Partial least-squares methods for spectral analyses. 2. Application to simulated and glass spectral data. *Anal. Chem.* **1988**, *60*, 1202–1208.

(40) Xu, Q.-S.; Liang, Y.-Z. Monte Carlo cross validation. *Chemom. Intell. Lab.* **2001**, *56*, 1–11.

(41) Verboven, S.; Hubert, M. LIBRA: a MATLAB library for robust analysis. *Chemom. Intell. Lab.* **2005**, *75*, 127–136.

(42) Foss NIR Systems. *Installation Manual for NIR Systems Scanning Spectrophotometers*; NIR Systems, Silver Spring, MD, 1991.