Near-Infrared Spectroscopy Combined with Partial Least Squares Discriminant Analysis Applied to Identification of Liquor Brands

Bin Yang[1], Lijun Yao[1,2], Tao Pan[1*]

[1]Department of Optoelectronic Engineering, Jinan University, Guangzhou, China
[2]Guangzhou SonDon Network & Technology Co., Ltd., Guangzhou, China
Email: *tpan@jnu.edu.cn

Abstract

The identification of liquor brands is very important for food safety. Most of the fake liquors are usually made into the products with the same flavor and alcohol content as regular brand, so the identification for the liquor brands with the same flavor and the same alcohol content is essential. However, it is also difficult because the components of such liquor samples are very similar. Near-infrared (NIR) spectroscopy combined with partial least squares discriminant analysis (PLS-DA) was applied to identification of liquor brands with the same flavor and alcohol content. A total of 160 samples of Luzhou Laojiao liquor and 200 samples of non-Luzhou Laojiao liquor with the same flavor and alcohol content were used for identification. Samples of each type were randomly divided into the modeling and validation sets. The modeling samples were further divided into calibration and prediction sets using the Kennard-Stone algorithm to achieve uniformity and representativeness. In the modeling and validation processes based on PLS-DA method, the recognition rates of samples achieved 99.1% and 98.7%, respectively. The results show high prediction performance for the identification of liquor brands, and were obviously better than those obtained from the principal component linear discriminant analysis method. NIR spectroscopy combined with the PLS-DA method provides a quick and effective means of the discriminant analysis of liquor brands, and is also a promising tool for large-scale inspection of liquor food safety.

Keywords

Identification of Liquor Brands, Near-Infrared Spectroscopy, Partial Least Squares Discriminant Analysis, Principal Component Linear Discriminant Analysis

1. Introduction

Chinese liquor is a distilled spirit mainly made from grain and obtained using distiller's yeast. This type of liquor also contains abundant micronutrients and active ingredients. Moderate drinking results in positive effects to various aspects. China is the leading country in liquor production and consumption. Unfortunately, many fake products are being sold in the market because liquor occupies sufficient market share and is highly profitable. These fake liquors are generally composed of the low-cost inferior liquors that counterfeit famous liquor brands and are prepared by simple dilution of industrial ethanol. They not only cause economic losses to producers of famous liquor brands but also pose serious threat to the health of consumers. As an important part of liquor quality inspection, the identification of liquor brands is increasingly attracting considerable attention.

Chinese liquor is a complex mixtures and composed mainly of water and ethanol; the remaining components contain hundreds of trace elements with various contents. Identification of liquor brands usually requires the determination of various feature components and their content recipes using traditional instrumental analysis methods (e.g., high-performance liquid chromatography). Such a detection method is complicated, costly, and cannot meet the needs of large-scale applications. Currently, identification of liquor products relies mainly on the sensory judgment of tasters. This method results in problems (e.g., great subjectivity and low precision) and is difficult to conduct in large-scale promotions. Thus, developing a simple and effective identification method of liquor brands is valuable.

Chemometric developments have demonstrated the significant potential of the near-infrared (NIR) spectroscopy analysis method in rapid and reagent-less measurement. This method is a powerful tool for quantitative analysis in various fields, such as agriculture [1] [2] [3] [4] [5], food [6] [7] [8] [9], environment [10], and medicine [11] [12] [13] [14] and so on. The NIR quantitative analysis has been used for determining the main components of liquor, including ethanol [7], ethylacetate [8], and aldehydes [9]. However, the components and contents are different in liquors of various brands that have diverse raw materials and production processes. Therefore, the identification of liquor brands is still difficult by the quantitative analysis of the above-mentioned conventional components.

Spectral discriminant analysis uses computer pattern recognition to identify and classify samples on the basis of collected spectral data. Instead of the quantitative analysis for some components of samples, its bases are the spectral overall features including the spectral similarities of the same type samples and the spectral differences among the different type samples. Spectral discriminant analysis mainly includes two main algorithms. The first one is the extraction of feature information; the known effective methods for extraction are principal component analysis (PCA) [1] [2] and partial least squares (PLS) [1] [2] [3] [4] [5] [10] [11] [12] [13] [14]. The second one is the classifier algorithm; the com-

mon and valid methods for classification are linear discriminant analysis (LDA) [1] [2].

Chinese liquor brands use three main flavors, namely, strong fragrant, mild fragrant, and sauce fragrant flavors. Compared with the liquor samples with same flavor, the difference of the liquor components and the corresponding spectra with different flavors are more obvious [15]. Water and ethanol are the main components in liquor. The spectra of liquors with different ethanol contents are remarkably diverse.

The identification of liquor brands is very important for food safety. Most of the fake liquors are usually made into the products with the same flavor and alcohol content as regular brand, so the identification for the liquor brands with the same flavor and the same alcohol content is essential. However, it is also difficult because the components of such liquor samples are very similar. As far as we know, there is currently barely no effective discriminant method for liquor samples with the same flavor and alcohol content.

The present study focused on the identification method for liquor brands with the same flavor and ethanol content. Although difficult, such a method is important and effective. The partial least squares discrimimant analysis (PLS-DA) method combined with the NIR spectroscopy were employed for the spectral discriminant analysis of liquor brands, and the principal component linear discriminate analysis (PCA-LDA) method is also performed for comparison.

2. Materials and Methods

2.1. Experimental Materials, Instruments, and Measurement Methods

As a kind of popular liquor brand with strong fragrant flavor in China, Luzhou Laojiao was used as the identified liquor brand. Liquors of the 10 other kinds of brands with strong fragrant flavor and the same ethanol content were used as the interferential samples. The collected liquor samples totalled 360. The identified liquor samples (160 bottles, negative) were composed of 160 bottles of Luzhou Laojiao Danya Erqu liquor (Luzhou Laojiao Group Co., Ltd., 52 vol, 125 mL). The interfering liquor samples (200 bottles, positive) were composed of 20 bottles from 10 other liquor brands. The 10 liquor brands were (1) Bainian Hutu (Value Pack) (Hutu Alcohol Co., Ltd., 52 vol, 125 mL), (2) Dukang Taibai (Luoyang Dukang Holdings Co., Ltd., 52 vol, 125 mL), (3) Shixiantaibai Xiaojiu-xiaojiuxian (Chongqing Shixiantaibai Alcohol Co., Ltd., 52 vol, 125 mL), (4) Luzhou Laojiufang Xiaoluzhou (Luzhou Laojiufang Sales Co., Ltd., 52 vol, 125 mL), (5) Tangchao Laojiao Gujiuwang (Tangchao Laojiao Co., Ltd., 52 vol, 125 mL), (6) Wudang Xiaojiufan (Xianzunniangjiu Co., Ltd., 52 vol, 125 mL), (7) Jingjuyuan Pingjian (Luzhou Jingjuyuan Alcohol Co., Ltd., 52 vol, 125 mL), (8) Tianxiafu (Anhui Tianxiafu Alcohol Co., Ltd., 52 vol, 125 mL), (9) Guifeizuijiu Xiaozui (Sichuan Guifeizuijiu Alcohol Co., Ltd., 52 vol, 125 mL), and (10) Kongfujia Shengshidatao (Kongfujia Alcohol Co., Ltd., 52 vol, 125 mL). A sample was extracted from each bottle of liquor and used for spectral measurement.

Spectral measurement was performed using a VERTEX 70 FT-NIR Spectrometer (Bruker Co., Germany) equipped with a transmission accessory and a 1 mm cuvette. Twelve scans of symmetrical interferograms at an 8 cm^{-1} resolution were added to each spectrum. The scanning scope of the spectrum ranged from 14994 cm^{-1} to 3996 cm^{-1} at a 3.857 cm^{-1} wavenumber interval, with 2852 wavenumbers. An InGaAs detector was used for the entire scanning region. Each liquor sample was measured twice, and the mean value of the measurements was used for modeling and validation. The spectra were obtained at 25˚C ± 1˚C and 45% ± 1% RH. The time of acquisition of an NIR spectrum was about 0.5 min.

2.2. Calibration, Prediction, and Validation Process

The Kennard-Stone (K-S) algorithm [16] [17] is an effective method for sample division in experimental planning. The objective is to select a maximally diverse subset from a large set of candidate samples. Thus, the subset can uniformly and sufficiently represent the entire sample space. The algorithm assumes that a "distance" between two samples can be defined, and the value is low when the two samples are similar and high when the samples are dissimilar.

A framework of calibration, prediction, and validation was performed to produce objective models. To ensure modeling representativeness and integrity, the calibration, prediction, and validation sets must all contain negative and positive samples. First, 60 negative and 80 positive samples were randomly selected for validation. The remaining samples (100 negative and 120 positive) were used for modeling. Using the K-S algorithm, the modeling samples were further divided into calibration (50 negative and 60 positive) and prediction (50 negative and 60 positive) sets to achieve uniformity and representativeness. Then, all models were established for the calibration and prediction sets, and the modeling parameters were optimized on the basis of the prediction recognition rate. Finally, the selected model was revalidated against the validation samples excluded from the modeling process.

2.3. PCA-LDA Method

PCA-LDA is the commonly well-performed method for spectral discriminant analysis [1] [2]. According to the principal component of the cumulative variance contribution rate to select the number of principal components, the first three principal components usually represent most of the information provided by the original variables. The two-dimensional PCA models with the combinations of any two in the first three principal components were usually adopted to next LDA procedure. The optimal principal component combination was selected according to the maximum *P_REC*. The detailed procedure can be found in the previous study [1] [2].

Based on principal component analysis, the PCA-LDA method uses the principal component vector of the spectral matrix to achieve qualitative discrimination of samples.

2.4. PLS-DA Method

Unlike the PCA-LDA method, the PLS-DA method classified the results of PLS quantitative analysis based on the assignment method, and then achieved qualitative discrimination of samples [18] [19]. In the PLS-DA method, the process for calibration and prediction is as follows. (1) The category variables of calibration samples were defined, and the value was assigned to 1 (or 0) for each positive (or negative) sample. (2) The number of PLS factors (F) was set from 1 to 20, and the PLS regression coefficients for each F were calculated on the basis of the spectra and categorical variables of all calibration samples. (3) On the basis of the obtained PLS coefficients and the spectrum of each prediction sample, the corresponding predictive values ($\tilde{V}_P$) of a category variable were further calculated for each F; when $\tilde{V}_P > 0.5$, the category variables (V_P) of prediction samples were assigned to 1 and the samples were determined as positive; otherwise, V_P values were assigned to 0, and the samples were determined as negative. (4) Referring to the genuine brand type of each prediction sample and the number of correctly recognized prediction samples, the prediction recognition rate can be calculated easily and was denoted as *P_REC*. The optimal number of PLS factors (F) was selected according to the maximum *P_REC*.

2.5. Model Validation

The validation samples excluded from the modeling optimization process were used to validate the optimal model of PLS-DA method. According to the genuine brand type of each validation sample and the number of correctly recognized validation samples, the validation recognition rate can be calculated easily and was denoted as *V_REC*. Moreover, the corresponding validation recognition rates of negative and positive samples can be calculated and were denoted as *V_REC$^-$* and *V_REC$^+$*, respectively.

3. Results and Discussion

The NIR spectra in the entire scanning region (14994 - 3996 cm^{-1}) of 200 Luzhou Laojiao (negative) and 160 non-Luzhou Laojiao (positive) liquor samples are shown in Figure 1. At 5128 and 6896 cm^{-1}, the absorption bands related with the OH stretch first overtone and second overtone of water, respectively [20]. The spectral features of Figure 1(a) and Figure 1(b) were compared. Given that the spectra of negative and positive samples were overlapping, no obvious spectral differences were obtained for direct discriminant analysis.

3.1. PCA-LDA Model

Using the method in Section 2.3, the PCA-LDA model was first established. The optimal principal component combination was PC$_1$ - PC$_2$, and the corresponding *P_REC* was 92.4%.The corresponding modeling parameters and effects were summarized in Table 1.
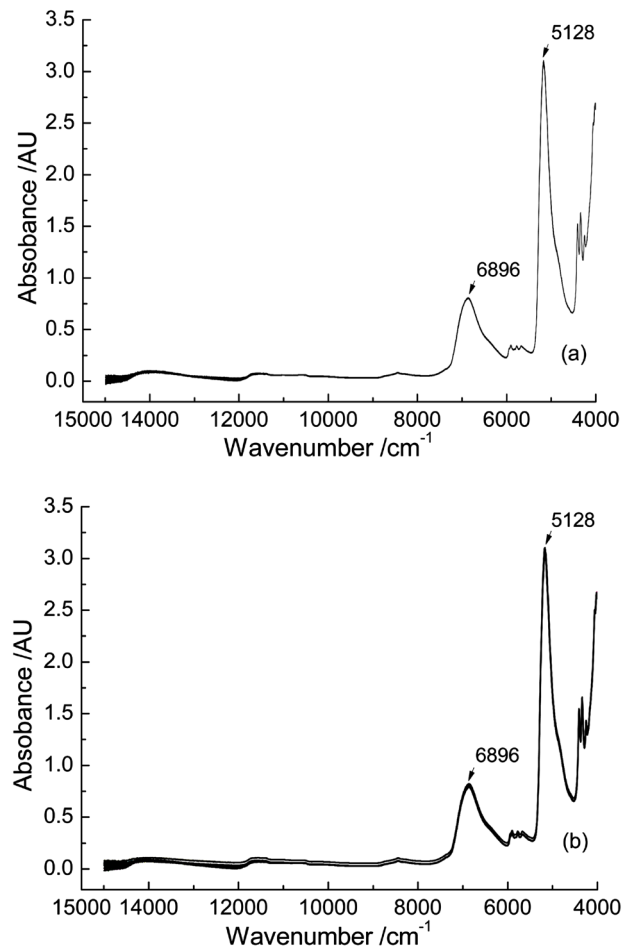
Figure 1. Near-infrared spectra of liquor for (a) Luzhou Laojiao (200 samples) and (b) non-Luzhou Laojiao (160 samples).

Table 1. Modeling parameters and effects of PCA-LDA and PLS-DA models.

Methods	F	PCC	P_REC
PCA-LDA	-	1, 2	92.4%
PLS-DA	7	-	99.1%

Note: PCC: principal component combination; F: number of PLS factors; P_REC: prediction recognition rate.

3.2. PLS-DA Model

The PLS-DA model was established according to the method in Section 2.4. The optimal number of PLS factors (F) was 7, and the corresponding P_REC was 99.1%. The corresponding modeling parameters and effects were also summarized in Table 1. The result indicates that the NIR spectroscopy combined with the PLS-DA method achieved good performance for the discriminant analysis of liquor brands, which was obviously better than that obtained from PCA–LDA model.

3.3. Validation

The randomly selected validation samples excluded from the modeling
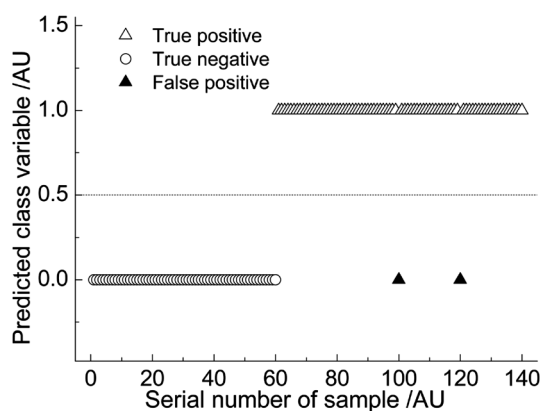
Figure 2. Validation recognition of the optimal PLS-DA model.

optimization process were used to validate the optimal model of PLS-DA method ($F = 7$). The corresponding validation recognition rates V_REC^-, V_REC^+, and V_REC achieved 96.7%, 100% and 98.6%, respectively. As shown in Figure 2, the validation samples were clearly divided into two parts using the different predicted class variables. Among them, only two samples were wrongly discriminated.

4. Conclusions

Chinese liquor is a popular alcoholic beverage, and occupies huge market share in China. The identification of liquor brands is of great significance for liquor food safety. There is currently barely no effective discriminant method for liquor samples with the same flavor and alcohol content because their chemical components are very similar.

In the present study, the PLS-DA method was successfully applied to the NIR spectral discriminant analysis of liquor brands with the same flavor and ethanol content. The experimental results indicate that the optimal PLS-DA model achieved high prediction recognition rate for the identification of liquor brands, and were obviously better than the results obtained from the PCA-LDA method. NIR spectroscopy combined with the PLS-DA method provides a quick and effective means of the identification of liquor brands, and is also a promising tool for large-scale inspection of liquor food safety.

Further wavelength selection can usually improve the spectral prediction effect, and reduce the scope of waveband, which will be the direction of the future researches.

Acknowledgements

References

[1] Williams, P. and Norris, K. (2001) Near-Infrared Technology in the Agricultural and Food Industries. American Association of Cereal Chemists, USA.

[2] Eriksson, L., Johansson, E., Kettaneh-Wold, N., Trygg, J., Wikström, C. and Wold, S. (2006) Multi- and Megavariate Data Analysis Part I: Basic Principles and Applications. Umetrics Academy, Umea, Sweden.

[3] Chen, H.Z., Pan, T., Chen, J.M. and Lu, Q.P. (2011) Waveband Selection for NIR Spectroscopy Analysis of Soil Organic Matter Based on SG Smoothing and MWPLS Methods. *Chemometrics and Intelligent Laboratory Systems*, **107**, 139-146. https://doi.org/10.1016/j.chemolab.2011.02.008

[4] Pan, T., Li, M.M. and Chen, J.M. (2014) Selection Method of Quasi-Continuous Wavelength Combination with Applications to the Near-Infrared Spectroscopic Analysis of Soil Organic Matter. *Applied Spectroscopy*, **68**, 263-271. https://doi.org/10.1366/13-07088

[5] Lyu, N., Chen, J.M., Pan, T., Yao, L.J., Han, Y. and Yu, J. (2016) Near-Infrared Spectroscopy Combined with Equidistant Combination Partial Least Squares Applied to Multi-Index Analysis of Corn. *Infrared Physics & Technology*, **76**, 648-654. https://doi.org/10.1016/j.infrared.2016.01.022

[6] Liu, Z.Y., Liu, B., Pan, T. and Yang, J.D. (2013) Determination of Amino Acid Nitrogen in Tuber Mustard Using Near-Infrared Spectroscopy with Waveband Selection Stability. *Spectrochimica Acta. Part A*: *Molecular and Biomolecular Spectroscopy*, **102**, 269-274. https://doi.org/10.1016/j.saa.2012.10.006

[7] Qu, F.F., Ren, D., Wang, J.H., Zhang, Z., Lu, N. and Meng, L. (2016) An Ensemble Successive Project Algorithm for Liquor Detection Using Near Infrared Sensor. *Sensors*, **16**, 89-102. https://doi.org/10.3390/s16010089

[8] Liu, J.X., Zhang, W.W., Han, S.H., Li, X., Li, P.Y., Yang, G.D., Yang, Y., Xu, B.C. and Luo, D.L. (2016) Rapid Detection of Caproic Acid and Acetic Acid in Liquor Base Based on Fourier Transform Near-Infrared Spectroscopy. *Food Science*, **37**, 181-185.

[9] Zhang, W.W., Liu, J.X., Han, S.H., Pan, Y.O., Li, X., Li, P.Y., Xu, B.C. and Luo, D.L. (2016) Determination of Aldehydes in Liquor Base Based on Fourier Transform Near-Infrared Spectroscopy. *Food Science*, **37**, 111-115.

[10] Pan, T., Chen, Z.H., Chen, J.M. and Liu, Z.Y. (2012) Near-Infrared Spectroscopy with Waveband Selection Stability for the Determination of COD in Sugar Refinery Wastewater. *Analytical Methods*, **4**, 1046-1052. https://doi.org/10.1039/c2ay05856a

[11] Xie, J., Pan, T., Chen, J.M., Chen, H.Z. and Ren, X.H. (2010) Joint Optimization of Savitzky-Golay Smoothing Models and Partial Least Squares Factors for Near-Infrared Spectroscopic Analysis of Serum Glucose. *Chinese Journal of Analytical Chemistry*, **38**, 342-346. https://doi.org/10.3724/sp.j.1096.2010.00342

[12] Pan, T., Liu, J.M., Chen, J.M., Zhang, G.P. and Zhao, Y. (2013) Rapid Determination of Preliminary Thalassaemia Screening Indicators Based on Near-Infrared Spectroscopy with Wavelength Selection Stability. *Analytical Methods*, **5**, 4355-4362. https://doi.org/10.1039/c3ay40732b

[13] Han, Y., Chen, J.M., Pan, T. and Liu, G.S. (2015) Determination of Glycated Hemoglobin Using Near-Infrared Spectroscopy. *Chemometrics And Intelligent Laboratory Systems*, **145**, 84-92. https://doi.org/10.1016/j.chemolab.2015.04.015

[14] Yao, L.J., Lyu, N., Chen, J.M., Pan, T. and Yu, J. (2016) Joint Analyses Model for Total Cholesterol and Triglyceride in Human Serum with Near-Infrared Spectroscopy. *Spectrochimica Acta Part A*, **159**, 53-59. https://doi.org/10.1016/j.saa.2016.01.022

[15] Zhang, A.N. and Zhang, J.H. (2010) Tutorial of Liquor Production and Blending. Science Press, Beijing, 140-146.

[16] Kennard, R.W. and Stone, L.A. (1969) Computer-aided Design of Experiments.

Technometrics, **11**, 137-148. https://doi.org/10.1080/00401706.1969.10490666

[17] Claeys, D.D., Verstraelen, T., Pauwels, E., Stevens, C.V., Waroquier, M. and Speybroeck, V.V. (2010) Conformational Sampling of Macrocyclic Alkenes Using a Kennard-Stone-Based Algorithm. *Journal of Physical Chemistry A*, **114**, 6879-6887. https://doi.org/10.1021/jp1022778

[18] Barker, M. and William, R. (2003) Partial Least Squares for Discrimination. *Journal of Chemometrics*, **17**, 166-173. https://doi.org/10.1002/cem.785

[19] Miguel, P.E. and Michel, T. (2003) Prediction of Clinical Outcome with Microarray Data: A Partial Least Squares Discriminant Analysis (PLS-DA) Approach. *Human Genetics*, **112**, 581-592.

[20] Lu, W.Z., Yuan, H.F. and Xu, G.T. (2000) Modern near Infrared Spectroscopy Analytical Technology. China Petrochemical Press, Beijing, 29-31.