Discriminant Analysis of Liquor Brands Based on Moving-Window Waveband Screening Using Near-Infrared Spectroscopy

Jie Zhong1, Jiemei Chen2, Lijun Yao1,3, Tao Pan1*

1Department of Optoelectronic Engineering, Jinan University, Guangzhou, China
2Department of Biological Engineering, Jinan University, Guangzhou, China
3Guangzhou SonDon Network & Technology Co., Ltd, Guangzhou, China

Email: *466945939@qq.com

Abstract

Partial least squares discriminant analysis (PLS-DA) with integrated moving-window (MW) waveband screening was applied to the discriminant analysis of liquor brands with near-infrared (NIR) spectroscopy. Luzhou Laojiao, a popular liquor with strong fragrant flavor, was used as the identified liquor brand (160 samples, negative, 52 vol alcoholicity). Liquors of 10 other brands with strong fragrant flavor were used as the interferential brands (200 samples, positive, 52 vol alcoholicity). The Kennard-Stone algorithm was used for the division of modeling samples to achieve uniformity and representativeness. Based on the MW-PLS-DA, a simplified optimal model set with 157 wavebands was further proposed. This set contained five types of wavebands corresponding to the NIR absorption bands of water, ethanol, and other micronutrients (i.e., acids, aldehydes, phenols, and aromatic compounds) in liquor for practical choice. Using five selected simple models with 4775 - 4239, 7804 - 6569, 6264 - 5844, 9435 - 7896, and 12066 - 10373 cm\(^{-1}\), the validation recognition rates were obtained as 99.3% or higher. Results show good prediction performance and low model complexity, and also provided a valuable reference for designing small dedicated instruments. The proposed method is a promising tool for large-scale inspection of liquor food safety.

Keywords

Liquor Brands, Near-Infrared Spectroscopy, Partial Least Squares Discriminant Analysis, Moving-Window Waveband Screening, Simplified Optimal Model Set
1. Introduction

Chinese liquor is a distilled spirit mainly made from grain and obtained using distiller’s yeast, which is a complex mixture and composed mainly of water and ethanol as well as micronutrients and active ingredients, including acids, aldehydes, phenols, and aromatic compounds. Unfortunately, because of the huge market share and high profitability, many fake liquors are being sold in the market, which not only causes economic losses to the producers of liquor brands but also poses a threat to consumers’ health. At present, identification of liquor brands usually requires the determination of various components and their content recipes using traditional analysis methods (e.g., high-performance liquid chromatography), which are complex and costly. Another method is the sensory judgment of tasters, which has great subjectivity and relies on the experience. The above methods are difficult to conduct in large-scale promotions.

With the developments of chemometric and sensor technology, near-infrared (NIR) spectroscopy has been proven to be a significant potential tool in the rapid and reagent-less measurement of various fields. It is reported that NIR quantitative analysis has been applied to determine the main components of liquor, such as ethanol [1], ethyl acetate [2], and aldehydes [3]. However, the components and contents are different in liquors of various brands. Therefore, it is still difficult to identify the liquor brands by quantitative analysis of the above conventional components.

Spectral discriminant analysis uses the spectral overall features to identify and to classify samples; its bases are that the spectral similarities of the samples of the same types and the spectral differences among samples of different types. Principal component analysis-linear discriminant analysis (PCA-LDA) is the commonly well-performed method for spectral discriminant analysis, which has been applied in the identification of liquor brands [4] [5]. Partial least squares-discriminant analysis (PLS-DA) is more effective than the PCA-LDA method in theory and practice [6] [7] [8], which has been applied in the identification of liquor brands [9]. However, neither of the literatures [4] and [5] strictly used the liquor brands of the same flavor and ethanol content for discriminant analysis. Literature [9] shown the experimental result of identifying liquor brands with the same flavor and ethanol content, but model only used the entire spectral region without any waveband selection and the prediction recognition rates required further improved.

Appropriate wavelength selection is essential for mitigating disturbance, improving prediction accuracy and simplifying the model, especially for the complex samples with multiple components. However, the above works [1] [2] [3] [4] [5] [9] on liquor brands identification are still based on the whole spectral region because of algorithm complexity. In the quantitative analysis of the NIR spectrum, moving-window waveband screening [10]-[15] combined with the PLS method can extract information effectively, eliminate noise disturbances, and improve predictive capability.

In the current study, moving-window (MW) waveband screening was inte-
grated to PLS-DA (MW-PLS-DA) and employed for the NIR spectral discriminant analysis of liquor brands. Furthermore, the optimal model set and its simplified method were further proposed, and the simple models with high accuracy were obtained.

The spectra of liquor samples of different flavors (or different ethanol contents) are remarkably diverse [1] [2] [3] [4]. This work is focused on the identification for liquor brands with the same flavor and ethanol content. Although difficult, such a method is important and essential.

2. Materials and Methods

2.1. Experiment

Luzhou Laojiao, a popular liquor with strong fragrant flavor in China, was used as the identified brand (negative). A total of 160 bottles of Luzhou Laojiao Danya Erqu liquor (52 vol alcoholicity) were collected. Liquors of 10 other brands with strong fragrant flavor and the same ethanol content (52 vol) were used as the interferential brands (200 bottles, positive), which were composed of 20 bottles from each liquor brand, including 1) Bainian Hutu, 2) Dukang Taibai, 3) Shixian Taibai, 4) Luzhou Laojufang, 5) Tangchao Laojiao, 6) Wudang Xiaojiufan, 7) Jingiuyuan Pingjian, 8) Tianxiafu, 9) Guifeizuijiu, and 10) Kongfujia Shengshidatao. A total of 360 liquor samples were extracted and used for spectral measurement.

The instrument was a VERTEX 70 FT-NIR Spectrometer (Bruker, Germany) equipped with a transmission accessory and a 1 mm cuvette. An InGaAs detector was used. Twelve scans were added to each spectrum. The entire scanning region was 14,994 - 3996 cm⁻¹ at a wavenumber interval of 3.857 cm⁻¹, with 2852 wavenumbers. Each sample was measured twice, and the mean value was used for modeling and validation. The spectra were obtained at 25°C ± 1°C and 45% ± 1% RH.

2.2. Sample Division

Initially, 60 negative and 80 positive samples were randomly selected into the independent validation set (140 samples), while the rest of 100 negative and 120 positive samples were used for modeling set (220 samples). Then, using the Kennard-Stone algorithm [16], the negative and positive modeling samples were further equally divided into calibration and prediction samples, to fulfill uniformity and representativeness.

2.3. Integrated MW-PLS-DA Method

All sub-waveband were traversed for modeling, using the following two parameters [11] [12] [13] [14] [15]: (1) initial wavenumber (I) and (2) number of wavenumbers (N). The search range was 14,994 - 3996 cm⁻¹ with 2852 wavenumbers, and I was set to $I \in \{14994,14990,\ldots,3996\}$. To cut down the volume of work and guarantee representativeness, $N$ was set to...
The obtained wavebands were used to establish the calibration and prediction models of PLS-DA. The process can refer to [6] [7] [8] [9]. Here, the positive and negative samples were assigned to the category value 1 and 0 respectively, then the quantitative calculation was carried out, the samples are classified by the 0.5 as the threshold. Where, the number of PLS factors (F) was set to $F \in \{1,2,\ldots,20\}$.

On the basis of the predicted category of the samples and their genuine brand type, it is easy to calculate the prediction recognition rate denoted as $P_{REC}$. According to the maximum $P_{REC}$, the optimal parameters (i.e. I, N and F) were selected and then the optimal MW-PLS-DA models were obtained.

### 2.4. Optimal Model Set and Its Simplification

Given that the optimal MW-PLS-DA models corresponded to the maximum $P_{REC}$ (denoted as $P_{REC}^*$) were not usually unique, the optimal model set and its simplification method were further proposed for the appropriate selection of wavebands. The optimal waveband set can be expressed as follows:

$$\Lambda^* = \{(I^{(e)}, E^{(e)}) \mid q = 1,2,3,\cdots,Q\} ,$$

where $I$ and $E$ are the initial and ending wavenumbers, respectively, and $Q$ is the number of optimal wavebands. If a containing relationship exists between two optimal wavebands, then

$$(I^{(i)}, E^{(i)}) \subseteq (I^{(j)}, E^{(j)}), \ 1 \leq i \neq j \leq Q .$$

The latter contained redundant wavenumbers, which must be removed from the optimal model set. The same processing was performed for each optimal waveband. Accordingly, the simplified optimal model set (denoted as $\Omega^*$) can be obtained. In the set of $\Omega^*$, no containing relationship existed between any two wavebands.

### 2.5. Model Validation

The validation group that containing 60 negative and 80 positive samples (total 140 samples) as well as out of the modeling optimization procedure were used for verifying the selected models screened using MW-PLS-DA method. According to the predicted category of validation samples and their genuine brand type, it is easy to calculate the validation recognition rate denoted as $V_{REC}$. Furthermore, the validation recognition rates of negative and positive samples can be calculated and were denoted as $V_{REC}^-$ and $V_{REC}^+$, respectively.

The computer platform was developed using Matlab 2012a.

### 3. Results and Discussion

#### 3.1. Full PLS-DA Model

The NIR spectra ranging from 14,994 to 3996 cm$^{-1}$ of liquor samples for 160
Luzhou Laojiao (negative, upper) and 200 Non-Luzhou Laojiao (positive, lower) are plotted, as shown in Figure 1. There were no apparent differences of spectra for direct discriminant analysis, since the given spectra of negative and positive samples were overlapping.

The PLS-DA model based on the entire scanning region (14,994 - 3996 cm\(^{-1}\)), called full PLS-DA, was first established. The optimal F was 7, and the corresponding \( P_{\text{REC}} \) was 99.1%. However, the adopted waveband contained a large number of wavenumbers (\( N = 2852 \)), which may include redundant wavenumbers. Therefore, the model complexity must be further reduced.

3.2. Simplified Optimal Model Set with MW-PLS-DA

The waveband selection was performed using the MW-PLS-DA method. The maximum \( P_{\text{REC}} \) achieved 100% (\( P_{\text{REC}}^* \)), and the optimal waveband set \( \Lambda^* \) contained 37,870 wavebands. The corresponding 2D diagram for initial and ending wavenumbers is shown in Figure 2(a). In the set of \( \Lambda^* \), a large amount of containing relationship was easily observed. Therefore, \( \Lambda^* \) must be further simplified. Using the simplification method mentioned above, the simplified optimal model set \( \Omega^* \) contained only 157 models. The 2D diagram is shown in Figure 2(b). No containing relationship was observed in \( \Omega^* \). In the average spectra, 157 wavebands were marked to clearly observe their position, as in Figure 3.

The wavebands of the simplified optimal model set \( \Omega^* \) could be divided into two parts, as follows.

The first part was associated mainly with the NIR characteristic absorption bands of water and ethanol. At 4347 cm\(^{-1}\), the absorption band related to the characteristic absorption of ethanol [3] could be applied to a quantitative analysis of ethanol in liquor. Two wavebands existed in \( \Omega^* \) containing the absorption band. These wavebands were 4775 - 4239 and 4772 - 4235 cm\(^{-1}\), the number

![Figure 1. NIR spectra of liquor samples for Luzhou Laojiao (160 negative, upper) and Non-Luzhou Laojiao (200 positive, lower).](image-url)
of wavenumbers $N$ were both 140, and the corresponding optimal $F$ were 7 and 8.

At 5128 and 6896 cm$^{-1}$, the absorption bands were related to the O-H stretch first overtone and second overtone of water [17]. A total of 55 wavebands in $\Omega^2$ was around the band at 6896 cm$^{-1}$. Among them, the waveband (7804 - 6569 cm$^{-1}$) was of low model complexity ($N = 320$) with the corresponding $F$ of 8.

The second part was associated mainly with the NIR characteristic absorption bands of other micronutrients (i.e., acids, aldehydes, phenols, and aromatic compounds) in liquor.

Both varieties showed absorption bands at 5586 cm$^{-1}$ related with the C-H stretch first overtone of acids, aldehydes, phenols; bands at 5917 cm$^{-1}$ were
related with either the C-H\(^3\) stretch first overtone or the C-H first overtone of aromatic groups \cite{17}. These bands were contained in 42 wavebands of \(\Omega'\). Among them, the waveband (6264 - 5844 cm\(^{-1}\)) was of low model complexity (\(N = 110\)) with the corresponding \(F\) of 7.

The small peak around 8333 cm\(^{-1}\) arose from the second overtones of C-H with stretching vibrations of acids, aldehydes, and phenols \cite{4}. A total of 49 wavebands in \(\Omega'\) were around the band at 8333 cm\(^{-1}\). Among them, the waveband (9435 - 7896 cm\(^{-1}\)) was of low model complexity (\(N = 400\)) with the corresponding \(F\) of 9.

The band at 11,235 cm\(^{-1}\) related with the C-H third overtone of aromatic compounds was contained by the remaining 9 wavebands in \(\Omega'\). Among them, the waveband (12,066 - 10,373 cm\(^{-1}\)) was of low model complexity (\(N = 440\)) with the corresponding \(F\) of 7.

Liquor samples with different ethanol contents have significantly different water contents. Hence, the wavebands in the first part of \(\Omega'\) were suitable to identify those samples. In this study, ethanol contents of samples were the same. Therefore, the wavebands in the second part of \(\Omega'\) were appropriate for the discriminant analysis.

### 3.3. Independent Validation

The randomly selected validation samples excluded from the modeling optimization process were used to validate the five selected simple wavebands (4775 - 4239, 7804 - 6569, 6264 - 5844, 9435 - 7896, and 12,066 - 10,373 cm\(^{-1}\)). The corresponding parameters and validation effects are summarized in Table 1. The validation recognition rates (\(V_{REC}\)) were 99.3% or higher, and the number of wavenumbers (\(N\)) were 440 or less. Furthermore, the full PLS-DA model (14,994 - 3996 cm\(^{-1}\)) was used to validate for comparison. The \(V_{REC}\) and \(N\) were 98.6%
Table 1. Parameters and validation effects of the five selected models screened using the MW-PLS-DA method.

<table>
<thead>
<tr>
<th>Waveband (cm⁻¹)</th>
<th>N</th>
<th>F</th>
<th>V_REC</th>
<th>V_REC⁻</th>
<th>V_REC⁺</th>
</tr>
</thead>
<tbody>
<tr>
<td>4775 - 4239</td>
<td>140</td>
<td>7</td>
<td>99.3%</td>
<td>100%</td>
<td>98.8%</td>
</tr>
<tr>
<td>7804 - 6569</td>
<td>320</td>
<td>8</td>
<td>99.3%</td>
<td>100%</td>
<td>98.8%</td>
</tr>
<tr>
<td>6264 - 5844</td>
<td>110</td>
<td>7</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>9435 - 7896</td>
<td>400</td>
<td>9</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>12,066 - 10,373</td>
<td>440</td>
<td>7</td>
<td>99.3%</td>
<td>100%</td>
<td>98.8%</td>
</tr>
</tbody>
</table>

and 2852, respectively (see also in [9]). The results indicate that, the validation effects of the five selected models were superior to the full PLS-DA model in two aspects of prediction performance and model complexity.

4. Conclusions

Most of the fake liquors are usually made into the products with the same flavor and ethanol content as regular brand, so the identification for such liquor samples is essential. However, it is also difficult because their components are very similar.

In the present study, the MW-PLS-DA was integrated and successfully applied to the NIR spectral discriminant analysis of liquor brands with the same flavor and ethanol content. A simplified optimal model set with 157 wavebands was further proposed based on the MW-PLS-DA. The five types of wavebands in the simplified optimal model set corresponded to the NIR absorption bands of water, ethanol, and other micronutrients (i.e., acids, aldehydes, phenols, and aromatic compounds) in liquor. According to the differences in components and NIR absorption features of objects, an appropriate model can be selected from them.

The experimental results indicate that the selected models achieved high prediction recognition rates with low model complexity, and provide a valuable reference for designing small dedicated instrument. The proposed method is a promising tool for large-scale inspection of liquor food safety.

Acknowledgements

This work was supported by the Science and Technology Project of Guangdong Province of China (No. 2014A020213016, No. 2014A020212445).

References


