RAPID DETERMINATION OF GREEN TEA ORIGINS BY NEAR INFRARED SPECTROSCOPY AND MULTI-WAVELENGTH STATISTICAL DISCRIMINANT ANALYSIS

X. G. Zhuang [1,2*], X. S. Shi [1], H. F. Wang [1], L. L. Wang [3], J. X. Fang [3]

[1] The 41st Research Institute of CETC, Qingdao, China; e-mail: xingangzhuang@163.com
[2] Science and Technology on Electronic Test and Measurement Laboratory, Qingdao, China
[3] Advanced Research Center for Optics, Shandong University, Jinan, China

A new simple classification modeling procedure, multi-wavelength statistical discriminant analysis (MW-SDA), is proposed for the identification of Shandong green tea origins coupled with near–infrared (NIR) spectroscopy. After smoothing and first derivative preprocessing, seven characteristic wavelengths (CW) were selected by enlarging the detailed information of preprocessed spectra. Then, for each characteristic wavelength, a classification threshold is calculated according to the differences in absorbance value, which can best separate the spectra for different origins. Based on the seven CWs and corresponding thresholds, seven classifiers were obtained, which form the classification model. The performance of the calibration model was evaluated according to sensitivity, specificity, and classification accuracy. Analysis results indicated that MW-SDA can be used well to build classification models. The predicted precision of the last model in prediction set was: sensitivity = 1, specificity = 0.967, and accuracy = 98.3%.

Keywords: *multi-wavelength statistical discriminant analysis, NIR spectroscopy, green tea, origin.*

БЫСТРОЕ ОПРЕДЕЛЕНИЕ ПРОИСХОЖДЕНИЯ ЗЕЛЕНОГО ЧАЯ С ПОМОЩЬЮ СПЕКТРОСКОПИИ БЛИЖНЕЙ ИК ОБЛАСТИ И МНОГОВОЛНОВОГО СТАТИСТИЧЕСКОГО ДИСКРИМИНАНТНОГО АНАЛИЗА

X. G. Zhuang [1,2*], X. S. Shi [1], H. F. Wang [1], L. L. Wang [3], J. X. Fang [3]

УДК 543.42:663.952.76

[1] 41-й Научно-исследовательский институт CETC, Циндао, Китай; e-mail: xingangzhuang@163.com
[2] Научно-техническая лаборатория электронных испытаний и измерений, Циндао, Китай
[3] Исследовательский центр оптики, Университет Шаньдуна, Цзинань, Китай

(Поступила 8 ноября 2017)

Для идентификации происхождения зеленого чая Шаньдуна предложена простая процедура классификации, основанная на многоволновом статистическом дискриминантном анализе в сочетании со спектроскопией ближнего ИК диапазона. После сглаживания спектра и предварительной обработки первой производной выбраны семь характерных длин волн, для каждой рассчитан порог классификации в соответствии с различиями в поглощении, которые могут наилучшим образом выделить спектры чая из разных источников происхождения. На основе характерных длин волн и соответствующих пороговых значений получены семь классификаторов, формирующих классификационную модель. Эффективность калибровочной модели оценивалась по чувствительности, избирательности и точности классификации. Результаты показали, что многоволновой статистический дискриминантный анализ может успешно использоваться для построения классификационных моделей. Прогнозируемые характеристики модели: чувствительность 1, избирательность 0.977 и точность 98.3%.

Ключевые слова: *многоволновой статистический дискриминантный анализ, спектроскопия ближнего инфракрасного диапазона, зеленый чай, происхождение.*

Introduction. As a traditional beverage, green tea is popular in East and Southeast Asian countries. It is not only looked upon as a tasty beverage in daily life, but also regarded as a kind of medicinal drink [1, 2]. Researches indicate that green tea plays a major role in the treatment of periodontal diseases [3] and some cancers [4, 5]. For instance, the catechins in green tea have antiviral activity against feline calicivirus [6]. In addition, studies suggest that polyphenols have potent antioxidant and anti-aging benefits [7, 8]. Compared with the green tea in southern China, Shandong green tea (Laoshan green tea and Rizhao green tea) has unique advantages, such as penetrating fragrance, thick mellow taste, and long brewing time, due to the special geographical location and large temperature difference between day and night. Nowadays, these two kinds of green tea enjoy immense popularity, especially in northern China. However, adulteration is becoming quite severe with the development of the green tea market due to lack of convenient and effective detection methods [9]. Therefore, an efficient and easy-to-use green tea detection method needs to be developed urgently.

Near-infrared (NIR) spectroscopy is one of the most rapidly developed analytical tools in recent decades [10]. It has been considered as a replacement of the traditional chemical analytical methods, such as high-performance liquid chromatography (HPLC) [11], capillary electrophoresis [12, 13] and colorimetric measurements. As a novel analytical tool, NIR spectroscopy has been extensively applied in the detection of food [14], petrochemical industry [15], and agriculture [16]. For example, combined with multivariate calibration methods, NIR spectroscopy has been widely applied in component analysis of olive, milk, fresh pork, and other products [17–21].

As applied to green tea, NIR spectroscopy has been used for category and origin identification [22]. Chen et al. investigated the feasibility for discrimination of roast green tea according to geographical origin by Fourier transform near infrared reflectance (FT-NIR) spectroscopy and supervised pattern recognition [23]. Combined with FT-NIR and partial least squares (PLS), NIR spectroscopy was successfully used for geographical origins identification of oolong tea (Anxi-Tieguanyin) [24]. Moreover, Zhao et al. used NIR spectroscopy and support vector machine (SVM) to discriminate green tea, black tea, and oolong tea [25]. Xu et al. reported a geographical indication identification method for a kind of Chinese green tea (Anji-white) by class modeling techniques and NIR spectroscopy [26]. Obviously, NIR spectroscopy has been successfully used to discriminate the origin and categories of tea. However, most of the classification models were built by complex algorithms, such as pattern recognition, PLS, and SVM, which was not conducive to embedded development and online application.

In this paper, we proposed a new classification method – multi-wavelength statistical discriminant analysis (MW-SDA). In contrast to conventional chemometric algorithms, the theory of MW-SDA is simple but effective. Taking green tea as the sample, this study aims to develop an efficient and nondestructive method for green tea origin identification by NIR spectroscopy and MW-SDA.

Experimental. *Spectrum pretreatment*. In this study, 200 representative green tea samples (100 Laoshan green tea samples and 100 Rizhao green tea samples) were collected from Laoshan and Rizhao, Shandong province, China. All 200 samples were randomly divided into a calibration set (70 Laoshan and 70 Rizhao green tea samples) and a prediction set at the ratio of 7:3. Therefore, the prediction set had 30 positive (Laoshan green tea) objects and 30 negative (Rizhao green tea) objects. In this study, Laoshan green tea and Rizhao green tea were respectively regarded as positive objects and negative objects.

Spectrum collection. The NIR spectra in the range of 1050-2500 nm were collected in the reflectance mode using an AvaSpec-NIR256/2.5TEC spectrometer (Avantes, The Netherlands). For each sample, 30±0.1 g tea leaf was filled into a 200 ml breaker to collect spectra by a standard diffuse fiber optic probe. The distance between probe and tea leaf was kept at 10 mm. Ten spectra were collected from different parts of every sample, and each spectrum was the average of 40 scans. The raw spectral data were measured in 6.4 nm intervals. The mean of the 10 spectra was applied in the subsequent analysis. After that, the region of 1300–2300 nm was selected for further analysis since both ends of the spectra exhibited a high level of noise. The room temperature was kept at 25°C, and the humidity was kept at an ambient level.

Spectrum preprocessing methods. All spectra were collected from tea leaf without grinding. Thus, the particle sizes of green tea samples were significantly different, which resulted in poor spectral reproducibility. To sharpen the poor peak resolution, the raw spectra were preprocessed before building classification models. Standard normal variate transformation (SNV) and the derivative are used for baseline correction [27]. SNV is a mathematical transformation method of the $\log(1/R)$ spectra used to remove slope variation. Compared with SNV, the derivative can eliminate overlapped spectral lines and enhance small spectral differences [25, 28].

Basic principle of multi-wavelength statistical discriminant analysis (MW-SDA). MW-SDA consists of four basic steps. Steps 1 and 2 are repeated until the optimal characteristic wavelengths and classification thresholds are determined (see Fig. 1).

1) Spectrum preprocessing. In contrast to mid-infrared (MIR) spectroscopy, the most intensive bands in NIR spectral region belong to the fundamental frequency and overtone vibration of hydrogen-containing functional groups. As a result, the NIR spectral curve is mostly comparatively smooth without sharp absorption peak, since the absorption bands are closely spaced. In addition, the inhomogeneity of the sample surface leads to baseline drift in the diffuse reflectance absorbance spectrum. Spectral pretreatment aims at removing the baseline drift and magnifying the details information of raw spectra.

2) After spectrum pretreatment, the spectra of train set samples were drawn by MATLAB. To facilitate samples identification, the spectra of different origins were plotted by different colors. After that, we need manually magnify the spectral image and look for the spectral variables that have remarkable difference in some absorbance values. The selected n ($n \geq 3$) variables are regarded as characteristic wavelength (CW). In generally, n is an odd. Then a classification threshold is calculated by a simple cycle program for each characteristic wavelength, which can well separate the spectra for different origins. As a result, n classifiers are obtained. If the classification performance is unacceptable, return to the first step and once again process the raw spectra to select characteristic wavelengths and thresholds.

3) By using the statistics analysis method, all the n classifiers make up a classification model. Then the new classification model is further used to identify the origins of green tea samples in the prediction set. As a result, each sample will get n prediction results according to the n classifiers.

4) For each sample in the prediction set, the final prediction origins are calculated by the n prediction results in step 3, according to majority role. For example, a sample has k Laoshan green tea prediction results and $n - k$ Rizhao green tea prediction results. If $k > n - k$, the sample is identified as Laoshan green tea, otherwise, Rizhao green tea. So, all prediction results of samples in prediction set can be obtained accordingly.

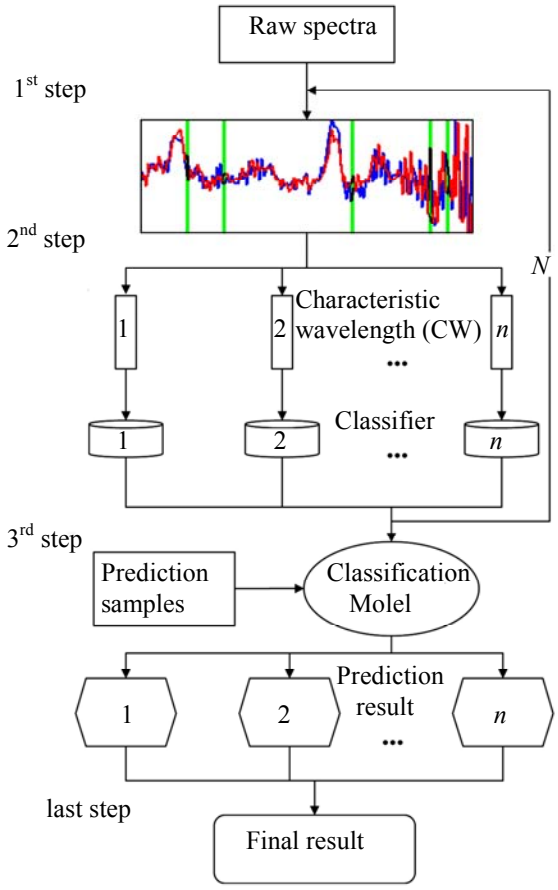

Fig. 1. Schematic picture of the MW-SDA classification method.

Software. For the spectra collection, AvaSoft (AvaSpecNIR256/2.5TEC systems) was used. All data analysis was done using a self-developed NIR analysis software ARCO-NIR, which was developed in MATLAB programming language by MATLAB 2010a (The Mathworks Inc., Natick, MA).

Results and discussion. *Spectrum preprocessing*. The spectra processed by SNV and the first derivative are presented in Fig. 2a. Both SNV and first derivative can well correct the baseline. Compared with SNV, the first derivative not only eliminates the influence caused by the heterogeneity of tea samples, but also enhances the repeatability and reproducibility of green tea spectra. However, it also enhances the noise while sharpening the poor peak resolution. As a consequence, the raw spectra were first smoothed by the Savitzky–Golay algorithm [29] before the first derivative.

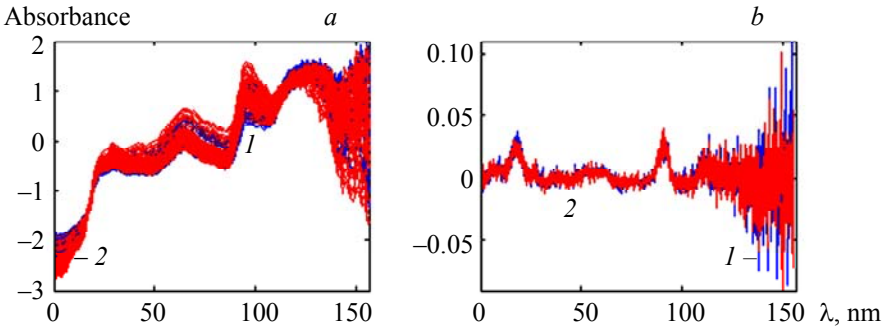

Fig. 2. Spectra for Laoshan (1) and Rizhao (2) green tea obtained from SVN data (a) and first derivative data (b).

Selecting CWs and building classifiers. After smoothing and first derivative preprocessing, the detailed information of raw spectra was enlarged and the processed spectra of two kinds of green tea presented significant differences in some spectral variables. Magnifying the local details information of the spectra, seven variables were selected as characteristic wavelengths, which showed obvious differences in absorbance value. As presented in Fig. 3, the selected seven characteristic wavelengths were 1700.16, 1752.68, 1939.54, 1990.02, 2040.14, 2071, and 2077.17 nm, respectively. By examining Fig. 3, the direction of peak curves of Laoshan and Rizhao green tea were basically opposite for all the seven characteristic wavelengths. Besides, the obvious clustering feature of absorbance value indicated that the seven characteristic wavelengths can be used to build classifiers for green tea origins identification.

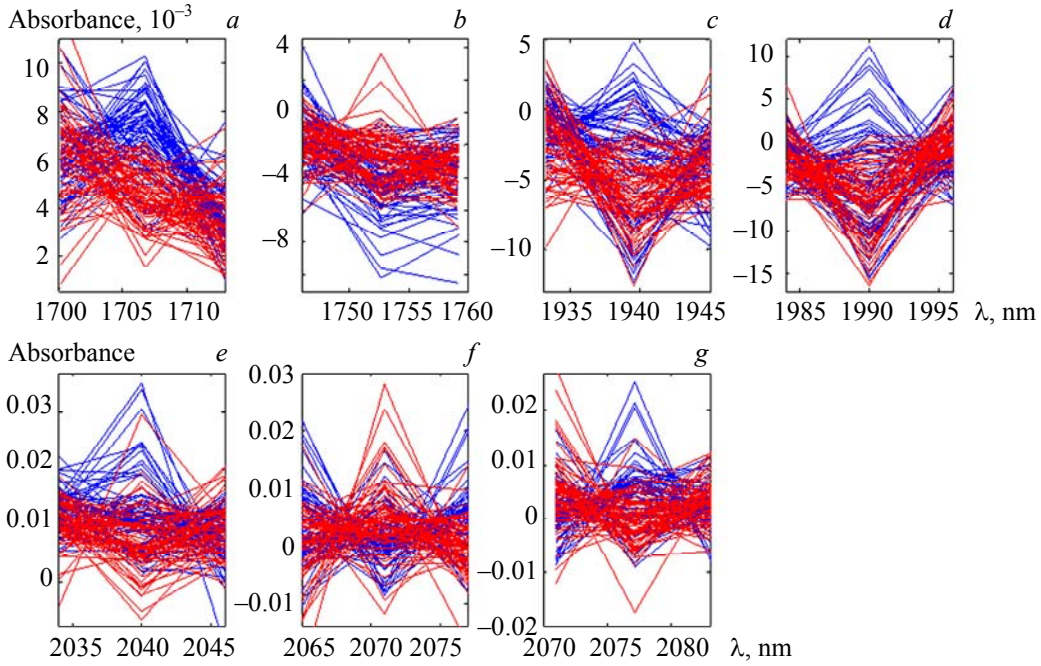

Fig. 3. The selected seven characteristic wavelengths.

For each characteristic wavelength, the classification threshold was calculated by a cyclical function with the following starting values: the minimum absorbance; stop value: the maximum absorbance; step size: 0.0001. With each iteration, the classification accuracy of green tea origins was calculated. The numeric value with optimal accuracy was regarded as the classification threshold, which formed a classifier. If there is more than one numeric value with the best classification effect, the middle one was regarded as the threshold. Following the methods, all seven classification thresholds were calculated and collected in Table 1. By comparing the absorbance values with the classification thresholds, the green tea origins could be identified. For example, the absorbance value below the threshold was for Laoshan green tea, and Rizhao green tea was indicated by the classifiers 2 and 6.

TABLE 1. The Seven Characteristic Wavelengths and Corresponding Classification Thresholds

Classifier	CW, nm	Threshold	High absorbance	Low absorbance
1	1700.16	0.0067	L	R
2	1752.68	–0.003	R	L
3	1939.54	–0.0034	L	R
4	1990.02	–0.001	L	R
5	2040.14	0.01	L	R
6	2071	0.0018	R	L
7	2077.17	0.0044	L	R

N o t e. CW is characteristic wavelength, L and R are Laoshan and Rizhao green tea.

As a result, the seven CWs and corresponding thresholds form seven classifiers (classification model). Then the 140 samples in calibration set were used to evaluate the identification effect of the classification model, according to sensitivity, specificity, and accuracy. The evaluation results are collected in Table 2. Classifier 7 gave the best result while classifier 4 gave the worst. Each sample in the calibration set will be given seven prediction results, and the final prediction result was determined by the maximum value. As presented in Table 2, both Laoshan and Rizhao green tea can be totally correctly identified, which indicated that the classifiers were acceptable.

TABLE. 2. Results from the Classification Model in Calibration Set

Classifier	TP	FN	TN	FP	Sensitivity	Specificity	Accuracy, %
1	54	16	69	1	0.771	0.986	87.857
2	55	15	52	18	0.786	0.743	76.429
3	60	10	65	5	0.857	0.929	89.286
4	34	26	67	3	0.486	0.957	72.143
5	62	8	62	8	0.886	0.886	88.571
6	62	8	62	8	0.886	0.886	88.571
7	64	6	63	7	0.914	0.9	90.714
Total	70	0	70	0	1	1	100

N o t e. TP, true positive; FP, false positive; TN, true negative; FN, false negative; "positive" and "negative" represent Laoshan and Rizhao green tea.

Classification results of the model. After the classification model was successfully built, it will be used to predict the green tea origins in the prediction set. The prediction results of seven classifiers are collected in Table 3, and all recognition rates were more than 75%. Similarly, each sample will get seven prediction origins, and all prediction results of 60 samples are collected in Fig. 4. For each sample, it can be recognized as Laoshan green tea and Rizhao green tea with different proportions, which was presented by a pie chart. The final prediction origin was judged according to the larger proportion. As seen in Fig. 4, all 30 Laoshan green tea samples were correctly identified, and one Rizhao green tea cannot be recognized. The accuracy of identification was 98.333%, as presented in Table 3.

TABLE 3. Results from the Classification Model in Prediction Set

Classifier	TP[a]	FN[b]	TN[c]	FP[d]	Sensitivity	Specificity	Accuracy (%)
1	24	6	27	3	0.8	0.9	85
2	25	5	21	9	0.833	0.7	76.667
3	26	4	27	3	0.867	0.9	83.333
4	22	8	30	0	0.733	1	86.667
5	30	0	25	5	1	0.833	91.667
6	23	7	26	4	0.767	0.867	81.667
7	28	2	26	4	0.933	0.867	90
Total	30	0	29	1	1	0.967	98.333

N o t e. As in Table 2.

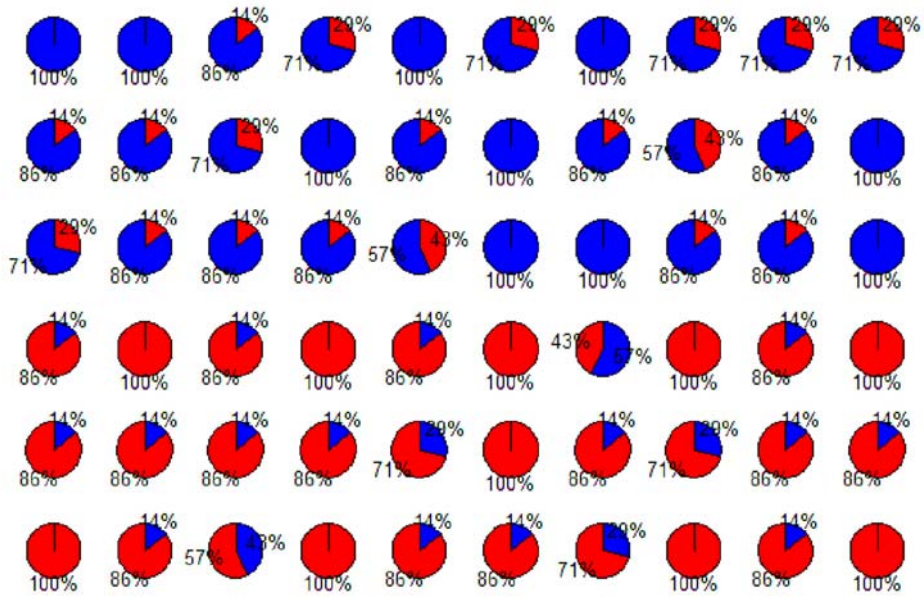

Fig. 4. The prediction result for Laoshan (•) and Rizhao (•) green tea origins.

As a natural organic beverage, the quality characteristics of green tea are closely related to the growing environment. Both Laoshan and Rizhao are from coastal areas of Shandong province, China. However, there are still slight differences in geomorphological environment and temperature, etc. Compared with Rizhao, Laoshan is mostly mountain terrain and the growth temperature is lower. Coupled with a small difference in their stir-frying-technology, these two kinds of green tea show some differences in the phenol ammonia ratio, which is greatly linked to the great quantity of functional groups (i.e., C–H, O–H, C=O, N–H, etc.). This could well explain that the first derivative spectra of Laoshan and Rizhao green tea samples have seven wavelengths that have great differences in absorbance value.

Conclusion. The present study has demonstrated that MW-SDA can be used well to build classification model in NIR analysis. MW-SDA has several other significant advantages when compared with traditional chemometrics methods, such as model transfer, simple structure, and easy operation. Starting with the spectral curve, the MW-SDA classification model can be easily built without complex professional knowledge. The simple structures make it easy for nonprofessionals. Secondly, the succinct modeling procedure is easily to embed into the instrument for on-line analysis. Finally, the structural features of MW-SDA make it conducive to model transfer between different spectrometers. The overall results have revealed that MW-SDA is powerful in green tea origins identification and is expected to be applied in other classification problems.

Acknowledgment. This work was financially supported by the State Key Laboratory of Science and Technology on Electronic Test and Measurement (No. 6142001180307), the National Basic Research Program of China (No. JSJL2016210A001, JSJL2018210C003), and State Key Laboratory of Sensor Technology Fund (No. SKT1507).

REFERENCES

1. T. Ikeda, S. Kanaya, T. Yonetani, A. Kobayashi, E. Fukusaki, *J. Agric. Food. Chem.*, **55**, 9908–9912 (2007).
2. Q. Chen, J. Zhao, S. Chaitep, Z. Guo, *Food Chem*, **113**, 1272–1277 (2009).
3. H. Arab, A. Maroofian, S. Golestani, H. Shafaee, K. Sohrabi, A. Forouzanfar, *J. Med. Plants. Res.*, **5**, 5465–5469 (2011).
4. I. C. Hou, S. Amarnani, M. T. Chong, A. Bishayee, *World J. Gastroenterol.*, **19**, 3713–3722 (2013).
5. L. X. Sang, B. Chang, X. H. Li, M. Jiang, *Nutr. Cancer*, **65**, 802–812 (2013).
6. E. G. Oh, K. L. Kim, S. B. Shin, K. T. Son, H. J. Lee, T. H. Kim, Y. M. Kim, E. J. Cho, D. K. Kim, E. W. Lee, M. S. Lee, I. S. Shin, J. H. Kim, *Food. Sci. Biotechnol*., **22**, 593–598 (2013).
7. T. Mostafa, D. Sabry, A. M. Abdelaal, I. Mostafa, M. Taymour, *Andrologia*, **45**, 272–277 (2013).
8. E. C. Yiannakopoulou, *Free Radic. Res.*, **47**, 667–671 (2013).
9. X. G. Zhuang, L. L. Wang, Q. Chen, X. Y. Wu, J. X. Fang, *Sci. China Technol. Sci*., **60**, 84–90 (2017).
10. M. Blanco, I. Villarroya, *Trends Anal. Chem*., **21**, 240–250 (2002).
11. K. Wei, L.Y. Wang, J. Zhou, W. He, J. M. Zeng, Y. W. Jiang, H. Cheng, *Food. Chem.*, **130**, 720–724 (2012).
12. D. A. El-Hady, N. A. El-Maali, *Talanta*, **76**, 138–145 (2008).
13. P. Li, S.Q. Dong, Q. J. Wang, Y. Z. Fang, *Chin. J. Org. Chem*., **26**, 485–488 (2008).
14. A. Alishahi, H. Farahmand, N. Prieto, D. Cozzolino, *Spectrochim. Acta A*, **75**, 1–7 (2010).
15. Y. Roggo, P. Chalus, L. Maurer, C. Lema-Martinez, A. Edmond, N. Jent, *J. Pharm. Biomed. Anal*., **44**, 683–700 (2007).
16. A. M. Mouazen, J. De Baerdemaeker, H. Ramon, *Soil, Till. Res.*, **80**, 171–183 (2005).
17. Y. T. Liao, Y. X. Fan, F. Cheng, *Meat Sci.*, **86**, 901–907 (2010).
18. Y. J. Xie, Z. A. Wang, W. P. Hu, S. Xu, *Anal. Bioanal. Chem.*, **404**, 3189–3194 (2012).
19. C. H. Latorre, R. M. P. Crecente, S. G. Martin, J. B. Garcia, *Food. Chem.*, **141**, 3559–3565 (2013).
20. O. Galtier, N. Dupuy, Y. Le Dreau, D. Ollivier, C. Pinatec, J. Kister, J. Artaud, *Anal. Chim. Acta*, **595**, 136–144 (2007).
21. M. J. Martelo-Vidal, F. Dominguez-Agis, M. Vazquez, *Aust. J. Grape Wine Res.*, **19**, 62–67 (2013).
22. N. S. Ye, *Crit. Rev. Food. Sci. Nutr*., **52**, 775–780 (2012).
23. Q. Chen, J. Zhao, H. Lin, *Spectrochim. Acta, A*, **72**, 845–850 (2009).
24. S. M. Yan, J. P. Liu, L. Xu, X. S. Fu, H. F. Cui, Z. Y. Yun, X. P. Yu, Z. H. Ye, *J. Anal. Methods Chem.*, 704971 (2014).
25. J. Zhao, Q. Chen, X. Huang, C. H. Fang, *J. Pharm. Biomed. Anal.*, **41**, 1198–1204 (2006).
26. L. Xu, P. T. Shi, X. S. Fu, H. F. Cui, Z. H. Ye, C. B. Cai, X. P. Yu, *J. Spectrosc*., **2013**, 1–8 (2013).
27. Y. Sun, Z. H. Du, X. Yin, K. X. Xu, *Spectrosc. Spectr. Anal.*, **28**, 2282–2284 (2008).
28. X. G. Zhuang, L. L. Wang, X. Y. Wu, J. X. Fang, *J. Infrared Millim. Wave*, **35**, 200–205 (2016).
29. A. Savitzky, M. J. E. Golay, *Anal. Chem.*, **36**, 1627–1639 (1964).