HYPERSPECTRAL IMAGING AND SPA-LDA QUANTITATIVE ANALYSIS FOR DETECTION OF COLON CANCER TISSUE

X. Yuan, D. Zhang*, Ch. Wang, B. Dai, M. Zhao, B. Li

*Ministry of Education, Optical Instrument and Systems Engineering Center,
Shanghai Key Laboratory of Modern Optical System, University of Shanghai for Science and Technology,
Shanghai, 200093, China; e-mail: dwzhang@usst.edu.cn*

Hyperspectral imaging (HSI) has been demonstrated to provide a rapid, precise, and noninvasive method for cancer detection. However, because HSI contains many data, quantitative analysis is often necessary to distill information useful for distinguishing cancerous from normal tissue. To demonstrate that HSI with our proposed algorithm can make this distinction, we built a Vis-NIR HSI setup and made many spectral images of colon tissues, and then used a successive projection algorithm (SPA) to analyze the hyperspectral image data of the tissues. This was used to build an identification model based on linear discrimination analysis (LDA) using the relative reflectance values of the effective wavelengths. Other tissues were used as a prediction set to verify the reliability of the identification model. The results suggest that Vis-NIR hyperspectral images, together with the spectroscopic classification method, provide a new approach for reliable and safe diagnosis of colon cancer and could lead to advances in cancer diagnosis generally.

Keywords: *colon cancer detection, image classification, successive projection algorithm – linear discrimination analysis, Vis-NIR hyperspectral imaging.*

ПОЛУЧЕНИЕ ГИПЕРСПЕКТРАЛЬНЫХ ИЗОБРАЖЕНИЙ И ИХ КОЛИЧЕСТВЕННЫЙ АНАЛИЗ МЕТОДАМИ ПОСЛЕДОВАТЕЛЬНЫХ ПРОЕКЦИЙ И ЛИНЕЙНОГО ДИСКРИМИНАНТНОГО АНАЛИЗА ДЛЯ ОБНАРУЖЕНИЯ РАКА ТОЛСТОЙ КИШКИ

X. Yuan, D. Zhang*, Ch. Wang, B. Dai, M. Zhao, B. Li

УДК 543.42:616-006.04

*Центр оптических приборов и систем Министерства образования КНР,
Шанхайский научно-технический университет, Шанхай, 200093, Китай;
e-mail: dwzhang@usst.edu.cn*

(Поступила 3 августа 2016)

Показано, что гиперспектральные изображения (ГСИ) позволяют реализовать быстрый, точный и неинвазивный метод выявления рака. Поскольку ГСИ содержат большой объем информации, необходимо проведение количественного анализа, обеспечивающего отбор данных для выявления злокачественной опухоли. Для демонстрации возможности выявления рака с помощью ГСИ в сочетании с предложенным нами алгоритмом создана установка для регистрации ГСИ в видимой и ближней ИК области спектра и получено большое количество ГСИ ткани толстой кишки. Анализ ГСИ ткани проведен с помощью метода последовательных проекций. Полученные данные использованы для построения идентификационной модели, основанной на линейном дискриминантном анализе и значениях относительной отражающей способности на эффективных длинах волн. В качестве прогнозного набора для проверки надежности созданной идентификационной модели взяты другие ткани. Показано, что совместное применение ГСИ и метода спектроскопической классификации обеспечивает новый подход к надежной и безопасной диагностике рака толстой кишки и может привести к успехам в диагностике рака в целом.

Ключевые слова: *обнаружение рака толстой кишки, классификация изображений, метод последовательных проекций и линейный дискриминантный анализ, гиперспектральное изображение в видимой и ближней ИК области.*

Introduction. Cancer of the colon is a malignant disease affecting millions of persons. The colon is a major part of the large intestine, and its cancer is a major cause of death in the Western and industrialized world [1]. According to [2], 34,000 new cases of colon cancer are diagnosed every year. During the year 2000, there were 16,250 deaths from colon cancer in the UK alone. Traditionally, colon cancer is diagnosed using microscopic analysis of histopathological colon samples by professional pathologists [3]. Until now, histopathological analysis was the most widely respected method for cancer diagnosis. However, that method is invasive, expensive, greatly depends on the judgments of pathologists, and much time is needed for obtaining the results [4]. Therefore, new techniques that are simple and enable rapid, accurate detection of colon cancer would be very helpful [5].

In recent years, many methods have been proposed for detecting colon cancer. For instance, Esgiar et al. proposed a promising method that uses textural features, because there are obvious differences between the textures of normal and malignant colon tissue [6]. Tosun et al. put forward the first object-oriented texture analysis based on a segmentation method comprising objective definition, texture definition, and segmentation into three parts [7]. The segmentation and classification techniques have been further divided using the Tosun et al. technology. Researchers [8–10] analyzed gene expression files of patients and employed algorithms on their data set of gene expressions from normal and malignant colon tissues. However, the above methods need precise and costly equipment, along with substantial time to detect and analyze the samples. The most problematic issues are the very complicated and manual steps.

Hyperspectral imaging (HSI), also known as chemical or spectroscopic imaging, is an emerging technology that integrates conventional imaging and spectroscopy to attain both spatial and spectral information from a scene under observation [11]. In the medical field, HSI is a novel method of generating spectral characteristics of regions of interest based on the chemical compositions of the substances [12]. For example, Hamed et al. [13] used infrared hyperspectral imaging to detect gastric cancer in 10 human subjects after resection, and they successfully distinguished differences between the cancerous and noncancerous tissues. Zacher et al. [14] analyzed the properties of human skin by using the HIS system according to different concentrations and localizations of absorbing and scattering in human tissues, which result in varying spectral reflectance. Compared to computed tomography (CT) and magnetic resonance imaging (MRI), HSI has several advantages. The HSI imaging system can produce three-dimensional information, including both two-dimensional spatial images and the third spectral dimension. That characteristic makes it one of the most powerful approaches for obtaining accurate information to distinguish between benign and malignant tissue. However, because of the enormous data content of the HSI images, it is a burdensome task for computers. Therefore, there is a need for an effective method to extract the most useful information without losing the details. In the colon cancer detection field, Kashif et al. [15–17] used a support vector machine algorithm to classify colon tissue into benign and malignant parts through cell detection, and they achieved relatively high classification accuracy. Yang et al. [18] used Fourier transform infrared hyperspectral microscopic images to histologically analyze stained tissue sections. Maggioni et al. [19] developed and applied a digital mirror device to conduct HSI data analysis of normal and malignant colon tissues. However, the huge time cost of the above methods limits them to use at the cytological level, and their sensitivity and accuracy also need to be improved.

This study aimed to provide a new method to safely distinguish cancerous colon regions from normal regions during surgery based on a hyperspectral imaging system in real-time and in vivo. We developed a novel way to combine HSI data to distinguish benign colon tissue from malignant colon tissue. A successive projection algorithm was explored as a strategy to select the effective characteristic wavelengths that include the mostly useful information, and linear discrimination analysis (LDA) was developed to determine whether a colon region is cancerous or noncancerous.

Experimental. In this study, four patients who were suffering from colon cancer underwent a total colectomy at the Tenth People's Hospital, Shanghai, China. After the surgery, a pathology inspection was made of the colon tissues to ensure that the tissues included both cancerous and noncancerous parts. After that, we took four samples to a laboratory and prepared them for hyperspectral imaging as soon as possible to maintain their freshness. The samples were divided into two groups. Two of them were selected randomly and served as the first group, the calibration set; the other two samples served as the second group, the prediction set. Finally, we captured the hyperspectral images and data from the colon samples.

Hyperspectral imaging instrumentation. To capture the experiment data, we used a Vis-NIR hyperspectral imaging system. We used an imaging spectrograph (VNIRImSpectorV10E, ZOLIX, China) with an interest wavelength range of 400 to 1000 nm, meaning that the spectral cube data contained the spectra of each

pixel from 400 to 1000 nm. The light-sensitive surface area of the CCD was 8.978×6.708 mm (length×width), and 1392×1040 pixels. The spectrometer had a high spectral resolution of 2.73 nm, with a slit width of 30 μm.

To obtain the spectral data, we positioned the HSI camera (400 to 1,000 nm) directly above the colon tissue, which was placed at the focal plane of the camera. The relations between the two light sources and the different points on the colon tissue were determined by the spectral response of the HSI camera, which was proportional to the amount of light received by it [20]. To achieve the best imaging quality, the CCD spectral response was kept between 3200 and 3500 (80% of the maximum spectral response). The scan time was from 5 to 10 s. Figure 1 shows a pixel selected randomly from the colon tissue. The hyperspectral image had a sequence of reflectance in various spectral wavelengths, as well as spatial (*x*-axis and *y*-axis) information.

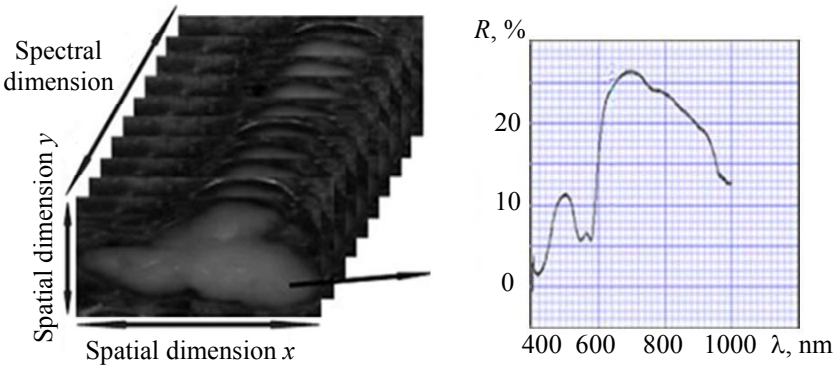

Fig. 1. Schematic view of the hyperspectral image of colon tissue and reflectance for each wavelength in the corresponding pixel.

Data normalization. The CCD camera in the Vis-NIR HSI system collecting the signal intensity conveyed information relating to not only the quality attributes of the tested sample but also to the sensitivity of the detector used and the sample illumination source [21]. The problem of spectral non-uniformity of the illumination device and the influence of the dark current were eliminated through normalization of the image data to reveal the normalized reflectance of the samples. A standard reference white board was placed in the field of view, and its data were utilized as a white reference. The dark currents were captured by keeping the camera shutter off. Before we processed the data, the raw data were normalized to find the relative reflectance using the following equation:

$$R(\lambda) = [I_{raw}(\lambda) - I_{dark}(\lambda)]/[I_{white}(\lambda) - I_{dark}(\lambda)], \quad (1)$$

where $R(\lambda)$ is the calculated relative reflectance value for each wavelength, $I_{raw}(\lambda)$ is the raw data radiance value of a given pixel, and $I_{dark}(\lambda)$ and $I_{white}(\lambda)$ are the dark current and the white board radiance acquired for each line and spectral band of the sensor, respectively.

Successive projection algorithm. A successive projection algorithm (SPA) employs simple operations in a vector space to obtain subsets of variables with small collinearity [22]. The SPA comprises 2 phases. The first phase consists of projections carried out on the X matrix, which generates K chains of M variables each. Each element in a chain is selected in order to display the least collinearity with the previous ones. The construction of each chain starts from one of the variables x_k, $k = 1, \ldots, K$, and follows the operations described below:

Step 1 (initialization): Let $z^1 = x_k$ (the vector that defines the initial projection operations) $x_j^1 = x_j$, $j = 1, \ldots, K$, SEL(1, k) = k, $i = 1$ (iteration counter).

Step 2: Calculating the matrix P^i of the projection onto the subspace orthogonal to z^i as

$$P^i = I - \frac{z^i (z^i)^T}{(z^i)^T z^i}, \quad (2)$$

where I is an (N×M) identity matrix.

Step 3: Calculating the projected vectors x_j^{i+1} as

$$x_j^{i+1} = P^i x_j^i, \tag{3}$$

for all $j = 1, \ldots, K$.

Step 4: Determining the index j^* of the largest projected vector and storing this index in element $(i+1, k)$ of the SEL matrix:

$$j^* = \arg_{j=1,\ldots,K} \max \|x_j^{i+1}\|, \tag{4}$$

$$\mathrm{SEL}(i+1, k) = j^*. \tag{5}$$

Step 5: Let $z^{i+1} = x_{j^*}^{i+1}$ (the vector that defines the projection operations for the next iteration).

Step 6: Let $i = i + 1$. If $i < M$, return to Step 2.

The second phase of the SPA consisted of evaluating candidate subsets of variables extracted from the chains generated in the first phase. The candidate subset of m variables starting from x_k was defined by the index set: {SEL(1, k), SEL(2, k), …, SEL(m, k)}. Because m ranged from 1 to M, and k ranged from 1 to K, a total of $M \times K$ subsets of variables was tested [23]. Different prediction performance metrics could be used to choose the best variable subset [24]. For this purpose, the present work adopted the root mean-square error of the prediction value obtained by applying the resulting partial least-squares regression (PLSR) model to an independent validation set.

Finally, selections of the effective wavelengths were conducted based on the calibration set.

Results and discussion. *Spectral diagrams.* We aimed to establish a relatively perfect calibration set; however, taking into consideration the number of samples, we randomly selected 55 cancerous pixels and 55 noncancerous pixels in all from the two samples of the first group. After capturing the HSI image, to lend credibility to the performance we soaked the colon tissues in formaldehyde for 2 months, and a professional staff made pathology slides. High- resolution microscopy was used to inspect the samples; we used hematoxylin and eosin (H&E)-stained preparations. The aim was to ensure that we could correctly distinguish cancerous pixels from noncancerous ones. The result is shown in Fig. 2: the values are the average reflectance of the pixels from the cancerous and noncancerous tissue regions. It was interesting to find that the cancerous tissue had higher reflectance intensity than the normal tissue. Both the cancerous and normal tissues had an intensity absorption range of 520 to 580 nm. Therefore, these spectral signatures were employed to automatically find the best method for classifying hyperspectral images.

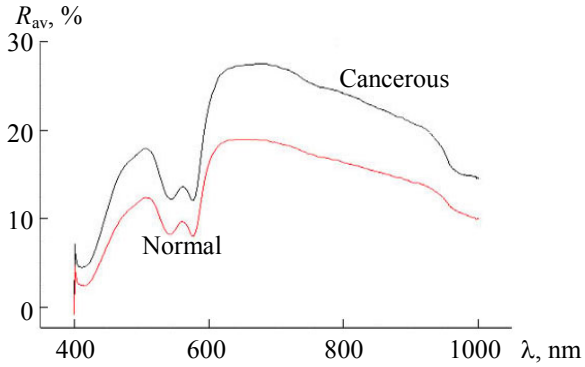

Fig. 2. Average reflectance spectra of cancerous tissue and normal tissue.

Results of the effective wavelength selection. We used SPA to select effective wavelengths for reducing redundant and correlated information without losing valuable details. We obtained the best selection results by the smallest root mean-square error of the prediction value of PLSR. Figure 3 shows the selection results: the SPA was executed using the GUI interface of MATLAB 2011b, and 11 effective wavelengths were selected: 405.235, 406.42, 408.794, 414.736, 422.48, 468.76, 559.489, 577.506, 594.342, 957.948, and 1000.31 nm.

These 11 effective wavelengths selected by SPA were used to analyze the spectral features of the colon tissue samples. To build the identification models, reaction center (RC) values corresponding to these 11 wavelengths were used as independent variables to replace the RC values corresponding to the whole spectral wavelengths.

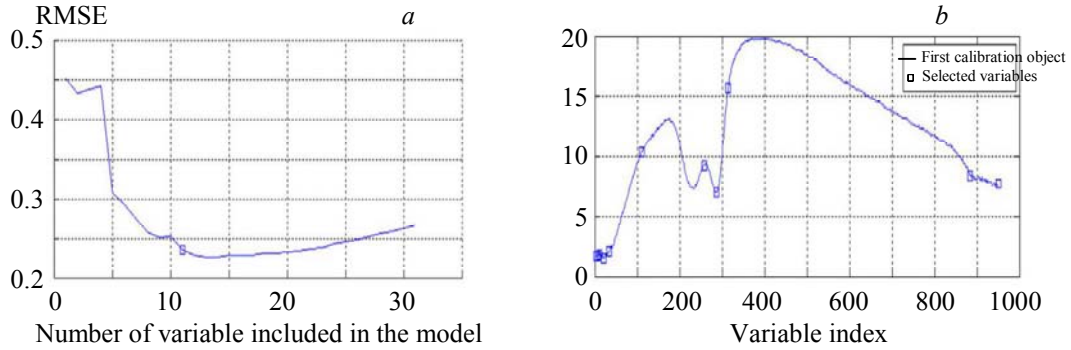

Fig. 3. Selected wavelengths: (a) number of wavelengths: root mean-squared error of prediction of PLSR curve, and (b) spatial distribution of wavelengths.

LDA based on effective wavelengths. In this analysis, the reflective values of the effective wavelengths were used as the variables of the LDA models. The type of colon tissue, as well as the normal part and the cancerous part, was set as the classification standard. All 110 groups of the data were processed on the LDA of IBM SPSS Statistics 19.0. Finally, we used a typical discriminant function: the corresponding 11 effective wavelengths were regarded as independent variables; F is the value of the typical discriminant function:

$$F = 2.109 - 0.064X_{405.235\ \mathrm{nm}} + 2.085X_{406.42\ \mathrm{nm}} + 2.607X_{408.794\ \mathrm{nm}} - 2.893X_{414.736\ \mathrm{nm}} - 1.780X_{422.48\ \mathrm{nm}} -$$
$$-1.022X_{468.76\ \mathrm{nm}} + 4.455X_{559.489\ \mathrm{nm}} - 2.112X_{577.506\ \mathrm{nm}} - 1.482X_{594.342\ \mathrm{nm}} + 1.710X_{957.948\ \mathrm{nm}} - 1.422X_{1{,}000.31\ \mathrm{nm}}. \quad (6)$$

From the typical discriminant function that we know, when the data F value is greater than zero, it represents a normal pixel, otherwise it represents a cancerous pixel. Finally, no sample data were miscalculated in the calibration set. Therefore, we achieved 100% accuracy for the LDA models in the calibration set. After the calibration, the LDA models were established to quickly identify the type of colon tissue. First, we needed to acquire the reflectance values corresponding to the 11 selected effective wavelengths from the normal and cancerous pixels of the colon tissues. We randomly selected 16 cancerous pixels and 16 noncancerous pixels from the two samples of the second group. Then 32 pixels of the data from the prediction set were used to determine the reliability of the LDA models. Thus, we could get a new series of F values through the sample data. Eventually, only three sets of the sample data were miscalculated in the prediction set, including two cancerous sets and one normal set. This resulted in an accuracy of 90.625% in the prediction set of the LDA models. The results of the LDA analysis were obtained on MATLAB 2011b.

To ensure the validity of our experiment, the positions of the pixels on the colon tissues must match those of the pixels on the pathological slides. Figure 4a shows an instance of the performance of one tissue with the cancerous region outlined in blue. This result was verified by the result of the corresponding H&E-stained preparation of normal and abnormal colon tissues, which was magnified using high-resolution microscopy. The result was confirmed by a professional and is shown in Fig. 4b. This was done mainly to check the validity of our experiment results.

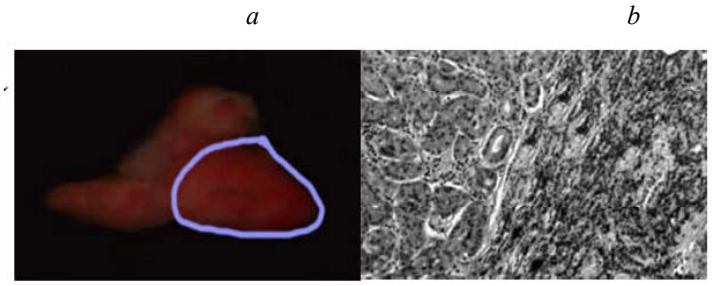

Fig. 4. Cancerous and noncancerous regions of colon tissue: (a) cancerous region outlined in blue by a pathologist, and (b) enlarged cancerous cells.

Performance criteria for the cancer detection are the false negative rate (FNR) and the false positive rate (FPR), which were calculated for each hyperspectral image. For a pixel not detected as tumorous, the detection was considered a false negative if the pixel was an actual tumor pixel according to a pathological examination. The FNR was defined as the number of false negative pixels divided by the total number of tumor pixels. For a pixel detected as tumorous, the detection was considered a false positive if the pixel was not a tumor. The FPR was defined as the number of false positive pixels divided by the total number of normal tissue pixels. The FNR in the SPA-LDA method was 12.5%, which shows 87.5% specificity; the FPR was 6.25%, which shows 93.75% sensitivity for a hyperspectral image.

Conclusion. We developed a new classification method based on SPA-LDA for hyperspectral images to be used for cancer detection. The SPA-LDA classifier was used to distinguish cancerous and noncancerous regions in human tissues. To enable validation, each sample was made on a pathology slide. The positions of the pixels that we extracted from the tissues were compared with the ones from the pathology slides. Through 11 selected efficient wavelengths, we detected a difference between the cancerous and normal tissues. We achieved high classification accuracy, including the 100% recognition rate and 90.625% prediction rate. An accurate examination with relatively safe margin resections will significantly reduce the mortality risk. Although the final decision as to whether tissue is cancerous or normal still belongs to physicians, our method could dramatically decrease the burden and provide referenced advice for them. Consequently, the numerical results of the FNR and FPR also revealed that the Vis-NIR HSI system together with the SPA-LDA method is an efficient tool for detecting colon cancer.

REFERENCES

1. R. S. Houlston, *J. Clin. Pathol.*, **54**, 206–214 (2001).
2. Cancer Research UK, Bowel Cancer Fact Sheet, May (2002).
3. S. Rathore, M. Hussain, A. Ali, A. Khan, *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **10**, 545–563 (2013).
4. Z. Liu, H. J.Wang, Q. L. Li, *J. Sensors*, **12**, 162–174 (2012).
5. N. Fujioka, Y. Morimoto, T. Arai, M. Kikuchi, *Cancer Detect. Prevent.*, **28**, 32–36 (2004).
6. A. N. Esgiar, R. N. Naguib, B. S. Sharif, M. K. Bennett, A. Murray, *IEEE Trans. Inform. Technol. Biomed.*, **2**, 197–203 (1998).
7. A. B. Tosun, M. Kandemir, C. Sokemnsuer, C. Gunduz-Demir, *Pattern Recogn.*, **42**, 1104–1112 (2009).
8. U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, *Proc. Natl. Acad. Sci. USA*, **96**, 6745–6750 (1999).
9. G. Marian, H. Patrick, B. Sandra, B. H. Amanda, W. Danny, V. Sudhir, S. Richard, L. Torsten, B. Heinz, J. D. Michael, B. M. Ghadimi, R. Thomas, *J. Cancer Res*., **67**, 41–56 (2007).
10. S. Yajima, M. H. Ishii, K. Aoyagi. K. Yoshimatsu, H. Kaneko, N. Yamamoto, T. Teramoto, T. Yoshida, Y. Matsumura, H. Sasaki, *Int. J. Oncol*., **31**, 1029–1037 (2007).
11. R. S. Gad, N. Vetrekar, I. Nazareth, J. S. Parab, *IEEE Xplore*, 741–744 (2012).
12. J. H. Lee, C. H. Won, *IEEE Xplore*, 1416–1419 (2009).
13. H. Akbari, K. Uto, Y. Kosugi, K. Kojima, N. Tanaka, *Cancer Sci*., **102**, 852–857 (2011).
14. A. Zacher*, Computer Vision and Graphics*, Springer-Verlag, 376–384 (2010).
15. K. Rajpoot, N. Rajpoot, In *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2004*, Springer, Berlin Heidelberg, 829–837 (2004).
16. K. M. Rajpoot, N. M. Rajpoot, M. J. Turner, *Hyperspectral Colon Tissue Cell Classification*, 829–837 (2004).
17. K. Masood, N. Rajpoot, In *Medical Image Understanding & Analysis* (2007).
18. Y. Chen, D. Niedieker, F. Großerüschkamp, M. Horn, A. Tannapfel, A. K. Thieltges, K. Gerwert, A. Mosig, *J. BMC Bioinform.*, **16**, 144–147 (2015).
19. M. Maggioni, G. L. Davis, F. J. Warner, F. B. Geshwind, A. C. Coppi, R. A. DeVerse, R. R. Coifman. *Algorithms from Signal and Data Processing Applied to Hyperspectral Analysis: Discriminating Normal and Malignant Microarray Colon Tissue Sections Using a Novel Digital Mirror Device System* (2006).
20. B. Li, Y. Zhou, M. Zhao, B. Hou, D. Zhang, Q. Wang, *J. Food Process Preserv*., **41**, No. 1, 12749 (2016).
21. W. Di, D. W. Sun, H. Yong, *Innov. Food Sci. Emerg. Technol*., **16**, 361–372 (2012).
22. M. C. U. Araújo, T. C. B. Saldanha, R. K. H. Galvão, T. Yoneyama, H. C. Chame, V. Visani, *Chemom. Intell. Lab. Syst.*, **57**, 65–73 (2001).
23. R. K. H. Galvão, M. C. U. Araújo, W. D. Fragoso, E. C. Silva, G. E. José, H. Paiva, *J. Chemom. Intell. Lab. Syst.*, **92**, 83–91 (2008).
24. E. R. Ziegel, *Technometrics*, **17**, 108–110 (2002).