Bin *et al.*
Journal of Analytical Science and Technology (2022) 13:18
https://doi.org/10.1186/s40543-022-00328-3

Journal of Analytical Science and Technology

RESEARCH ARTICLE Open Access

Identification of different parts of *Panax notoginseng* based on terahertz spectroscopy

Li Bin*, Han Zhao-yang, Cai Hui-zhou, Yang A-kun and Ou Yang Ai-guo

Abstract

In this paper, the combined terahertz time-domain spectroscopy (THz-TDS) and chemometrics method is proposed to identify four different parts of *Panax notoginseng* rapidly and nondestructively. The research objects of the taproot, scissor, rib, and hairy root of *P. notoginseng* are taken. The refractive index, absorption coefficient, time-domain, and frequency-domain spectra of the samples are analyzed. It is found that the terahertz spectra of different parts of *P. notoginseng* are significantly different, so the absorption coefficient of samples is selected to establish models. Firstly, the baseline correction, multiple scattering correction, and normalization algorithms are used to preprocess the absorption coefficient in 0.5–2.0 THz to remove noise. Then, the Kennard–Stone (KS) algorithm is used to divide the model set and the prediction set at the ratio of 3:1, and the successive projection algorithm (SPA) is used to select the characteristic frequency points of the samples. Finally, the chosen characteristic variables are input into the support vector machine (SVM) and linear discriminant analysis (LDA) algorithm to establish the qualitative analysis models, respectively. In the SPA-SVM models, the performance of the SPA-SVM model under the linear kernel function by baseline is best, the accuracy of the training set of it is 99.50%, and the accuracy of the test set of it is 99.25%. In the SPA-LDA models, the performance of the SPA-LDA model by baseline is best, and the accuracy of the training set of it is 100%, and the accuracy of the test set of it is 100%. And the value of cumulative variance contribution is proposed to assess whether the variable is good or bad to model. The results show that the combined THz-TDS and chemometrics method can be used to realize rapid, accurate, and nondestructive identification of different parts of *P. notoginseng*.

Keywords: THz-TDS, *Panax notoginseng*, Qualitative analysis, Chemometrics

Introduction

Panax notoginseng is an essential herbal medicine in China, and it has an excellent effect on the treatment of cardiovascular disease. The main practical component of *P. notoginseng* is saponin. In contrast, the different parts of *P. notoginseng* have various contents of saponins. The rib and hairy root of *P. notoginseng* have a lower range of saponins than others. The saponins in the scissors of *P. notoginseng* are not absorbed by the body of people easily (Yun et al. 2019). Although the different parts of *P. notoginseng* are different in morphological, it is a great challenge to identify the powder form of the different parts of P. notoginseng, which are similar in color and smell. Some undesirable businessmen use the scissor, ribs, and hairy root powders as taproot powder to sell to get higher interest. Those affect the benign development of the *P. notoginseng* industry seriously and impact the curative effects of the patients with the disease of cardiovascular. Therefore, it has great practical significance to identify the powder form of the different parts of *P. notoginseng*.

In recent years, many scholars have studied the quality of *P. notoginseng*. Li and Chen (2019) used the liquid chromatography-tandem mass spectrometry to identify

*Correspondence: libingioe@126.com

National and Local Joint Engineering Research Center of Fruit Intelligent Photoelectric Detection Technology and Equipment, Institute of Optical-Electro-Mechatronics Technology and Application, East China Jiao Tong University, Nanchang 330013, China

the pesticide residues in *P. notoginseng*. Yun et al. (2018) detected the content of polysaccharides in *P. notoginseng* using near-infrared spectroscopy. Shen et al. (2019) used laser-induced breakdown spectroscopy (LIBS) to see the contents of six metal nutrients in *P. notoginseng* samples from eight production areas. Meng et al. (2018) took pictures of *P. notoginseng* and its counterfeits under an electronic mirror to identify their micro-traits. Huang et al. (2014) explored the identification of traditional Chinese medicine molecules by the molecular identification method. Hong and Wang (2018) evaluated the medication and efficacy of *P. notoginseng* by analyzing its characteristics, physicochemical, and microscopic features. However, the operation of liquid chromatography-tandem mass spectrometry is complicated, and it is impossible to realize rapid and nondestructive detection; the anti-interference ability and sensitivity of near-infrared spectroscopy are insufficient; and the sample surface is ablated during the detection process of LIBS, reducing the intensity and stability of the spectral lines (Zheng et al. 2015). A microscope is used to directly observe the surface characteristics of the sample in the detection process of microscopic identification, it is expensive and slow, and it also cannot meet the requirements of rapid detection (Zhao et al. 2017).

THz is a tiny band of the electromagnetic spectrum between the microwave and infrared region (Yan et al. 2018). Many low-energy physical phenomena (e.g., vibrational modes, phonons, rotational and vibrational energy levels) and biomolecular activities (e.g., nucleic acids, amino acids, carbohydrates, peptides, and proteins) can be observed in the terahertz frequency range; in recent decades, the THz-TDS technology has been widely used in the field of food safety detection (Yun et al. 2020), such as adulteration detection (Bin et al. 2021), Chinese herbal medicine identification (Long et al. 2020), the detection of fat in food (Ge et al. 2014; Bin et al. 2021), and the detection of food additives (Baek et al. 2014). At the same time, the energy of terahertz wave photons is so small that terahertz waves do not have enough energy to break molecular bonds, the operation is simple and it is enough to obtain a wide range of information about the sample, and the terahertz spectrum also has a high signal-to-noise ratio and high penetration power. Based on the above advantages, THz-TDS is used to detect adulteration of Panax ginseng powder for research purposes in this paper.

Currently, the existing detection technology cannot realize rapid, nondestructive, and accurately identifying the different parts of *P. notoginseng* powder adulteration, so the combined THz-TDS and chemometrics method is proposed to identify the four parts of *P. notoginseng* powder adulteration, and the optimal model of qualitative identification of different parts of *P. notoginseng* is established.

Materials and methods

Sample preparation

The materials used in the experiment are purchased from wen-shan Kang-wan-Jia agricultural development limited company. The samples contain taproot, scissor, rib, and hairy root of *P. notoginseng*. The grinder smashes the experimental samples, and they are filtered by a 200-mesh sieve and are dried by the dryer, and the powder samples are pressed for 1 min under the pressure of 10 MPa to make the samples in a diameter of 13 mm. The thickness of the sample is measured before spectral collection, and the thickness of the samples is from 0.8 mm to1.1 mm. Each sample is measured four times at four different locations. The samples' spectral data are collected by the THz-TDS instrument of TAS7400TS which is developed by Edwin Company of Japan. The quantity information of experimental samples is shown in Table 1.

Spectral pretreatment

There are three primary sources of noise in the THz spectrum (Peng et al. 2018): (1) When the particle size of the sample material is equivalent to the wavelength of THz, the scattering effect of the particle results in the visible loss of the amplitude of the THz, producing in a banked baseline; the baseline breaks the linear relationship between the mixture and its components, which increases the established difficulty of the model; (2) the purity of the sample also introduces noise, the purity of the sample is lower, and the accuracy of the model is lower; and (3) the system and environmental noise of the THz spectral system appear in the sample collection process randomly, and the existence of these noises also increases the difficulty of spectral analysis. Therefore, spectral pretreatment algorithms are needed to be used to eliminate the noise to improve the accuracy of the model. In this paper, baseline correction, multiplicative scatter correction (MSC), and normalization are used to preprocess the spectral data, respectively. Baseline correction algorithm can eliminate spectral baseline

Table 1 Number of samples prepared from different parts of Panax notoginseng

Sample name	Number of experimental samples	Acquisition times	Number of spectra
Taproot	30	4	120
Scissor	35	4	140
Rib	35	4	140
Hairy root	33	4	132

deviation (Zhou et al. 2016), and MSC algorithm can eliminate the scattering effect caused by the uneven distribution of solid particles (Zhang et al. 2019). Normalization can calibrate spectral changes caused by minor optical path differences. All data processing in this paper are performed in The Unscrambler X (CAMO Software Inc., USA) and MATLAB 2019a (The MathWorks Inc., Natick, USA).

Modeling method

The successive projections algorithm (SPA) is a forward variable selection method, and it is used to reduce the collinearity of different variables (Liu et al. 2020). In the SPA process, the maximum number of selected variables is first set, and an initial vector (M is the original variable) is chosen in the m-dimensional space. Then, a high projection vector is selected as the new starting vector in the orthogonal subspace. The support vector machine (SVM) algorithm, whose main idea is to find the optimal separating hyperplane, is a supervised learning model, and the nonlinear mapping function is used to map the training data set to get a high-dimensional space, and the distance between samples of different classes is most significant (Shi et al. 2020). The SVM has a good generalization ability in the classification of samples of other styles (Cao et al. 2020). In the process of establishing the SVM model, the values of penalty factor c and kernel parameter g are the key (Cao et al. 2018); the values of optimal c and g parameters are obtained through grid search in this paper.

The linear discriminant analysis (LDA) algorithm is a supervised classification method. The basic idea of LDA classification is to extract the best recognizable low-dimensional features from the high-dimensional parts, and these selected features are used to classify samples, so similar samples can be gathered together as much as possible; when the LDA is optimized, the variance of inter-class is most extensive, and the conflict of intra-class is most minor (Liu et al. 2015, 2018). The distribution covariance matrix of inter-class and intra-class is defined as:

$$\boldsymbol{S_B}=\sum_{i=1}^{k} N_i(\bar{\boldsymbol{x}}_i-\bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_i-\bar{\boldsymbol{x}})^T \tag{1}$$

$$\boldsymbol{S_W}=\sum_{i=1}^{k}\sum_{j=1}^{N_i} N_i(\boldsymbol{x}_{ij}-\bar{\boldsymbol{x}}_i)(\boldsymbol{x}_{ij}-\bar{\boldsymbol{x}}_i)^T \tag{2}$$

where $\boldsymbol{x_{ij}}$ is the spectral vector of the j sample in the i class sample, k is the total of class samples, $\bar{\boldsymbol{x}}_i$ is the average spectrum of the i class sample, N_i is the total of variable in the i class sample, and $\bar{\boldsymbol{x}}$ is the average spectrum of all the samples. The purpose of LDA is to find a transformation matrix to get the maximum inter-class variance and the minimum intra-class variance. The formula of the transformation matrix is as follows:

$$\boldsymbol{J(w)}=\frac{\boldsymbol{w}^T\boldsymbol{S_B}\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{S_A}\boldsymbol{w}} \tag{3}$$

where $\boldsymbol{S_A}$ is the nonsingular matrix, $\boldsymbol{J}$ is the transformation matrix, and $\boldsymbol{w}$ is the characteristic matrix of $\boldsymbol{S_A^{-1}S_b}$.

Principal component analysis

Principal component analysis (PCA) is a multivariate statistical method that transforms the original high-dimensional data into linearly uncorrelated low-dimensional feature variables through an orthogonal transformation (Abdi and Williams 2010), and the transformed variables are called principal components (PCs). PCA is a linear algorithm, and thus, it cannot be used to explain complex polynomial relationships between features. In general, the original data can be replaced by the first n PCs when the cumulative variance contribution of the current n PCs is sufficiently large (typically 85%), and the process of principal component analysis is as follows (Li et al. 2020):

Standardize the original spectral data $\boldsymbol{A}_i$ and calculating the covariance matrix $\boldsymbol{S}$ for n samples by formula (4) and formula (5):

$$\boldsymbol{A}_{\mathrm{i}}^{*}=\frac{\boldsymbol{A}_i-mean(\boldsymbol{A}_i)}{std(\boldsymbol{A}_i)}(i=1,2,3,\cdots,n) \tag{4}$$

$$\boldsymbol{S}=\frac{\boldsymbol{A}^{*T}\boldsymbol{A}^{*}}{n-1} \tag{5}$$

Calculate the eigenvalues of the covariance matrix $\boldsymbol{S}$, and correlation coefficient matrix $\boldsymbol{R}$ by formula (6) and formula (7):

$$r_{ij}=\frac{\sum_{k=1}^{n} A_{ki}^{*}A_{kj}^{*}}{n-1},(i,\ j=1,2,3,\cdots,m) \tag{6}$$

$$\boldsymbol{R}=(r_{ij})_{m\times n} \tag{7}$$

where $r_{ii}=1$, $r_{ij}=r_{ji}$, r_{ij} is the correlation coefficient between the sample i and the variable j, m is the number of eigenvalues, and k is the k-th standardized spectral data. The eigenvalues are presented accurately and in the correct sequence, $\lambda_1\geq\lambda_2\geq\lambda_3\geq\lambda_m\geq 0$.

Based on the cumulative variance contribution, choose the suitable number of principal components and build the model.

In this study, four different parts of Panax ginseng roots, including main root, cuttings, tendons, and hairy roots, are investigated using THz time-domain spectroscopy in the following steps:

(1) The MSC, baseline correction, and normalization are used to preprocess the absorption coefficient spectra;
(2) The SPA is used to select feature variables from absorption coefficient spectra;
(3) The PCA is applied for characteristic analysis of spectral variables;
(4) Based on the selected bands, the SVM and LDA models are established;
(5) The results of the two models are compared and analyzed, and the best model is selected.

Results and discussion

Spectral analysis

The THz-TDS system used in this experiment is developed by Advantest, Japan. The spectral measurements are performed in the transmission mode of the time domain, with a spectral acquisition range of 0.5–7 THz, the resolution of it is 7.6 GHz, the center wavelength of the laser is 1560 nm, and the power of the laser is 400 μW. Due to the strong influence of moisture on terahertz spectra, the spectra are acquired in a closed chamber with a continuous pumping of dry air to keep the air humidity below 10% and the temperature is kept around 25 °C. All THz-related data required for this experiment are collected by this system.

Figure 1 shows the average spectra of *P. notoginseng* powder samples from four different parts. Figure 1a shows the average time-domain spectra of four samples, the time-domain spectra in the range of 16-20 ps are selected, and those higher than 20 ps or lower than 16 ps are regarded as invalid bands with no practical information; for the four samples, the time and amplitude of the peak of the hair root and scissor of *P. notoginseng* are similar; and the peak amplitude of the rib and taproot of *P. notoginseng* is lower than that of the hair root and scissor. Figure 1b shows the average absorption coefficient spectrum of the four samples. The absorption coefficient spectrums of 0.5–2.0 THz are taken for analysis, and there is no prominent absorption peak, which is similar to the results obtained by Xiao-Li et al. (Xiao-Li and Jiu-Sheng 2011). With the increase in frequency, the absorption coefficients between samples increase, and each of the absorption coefficients of four samples is different. Taking the absorption coefficient at 1.9 THz as an

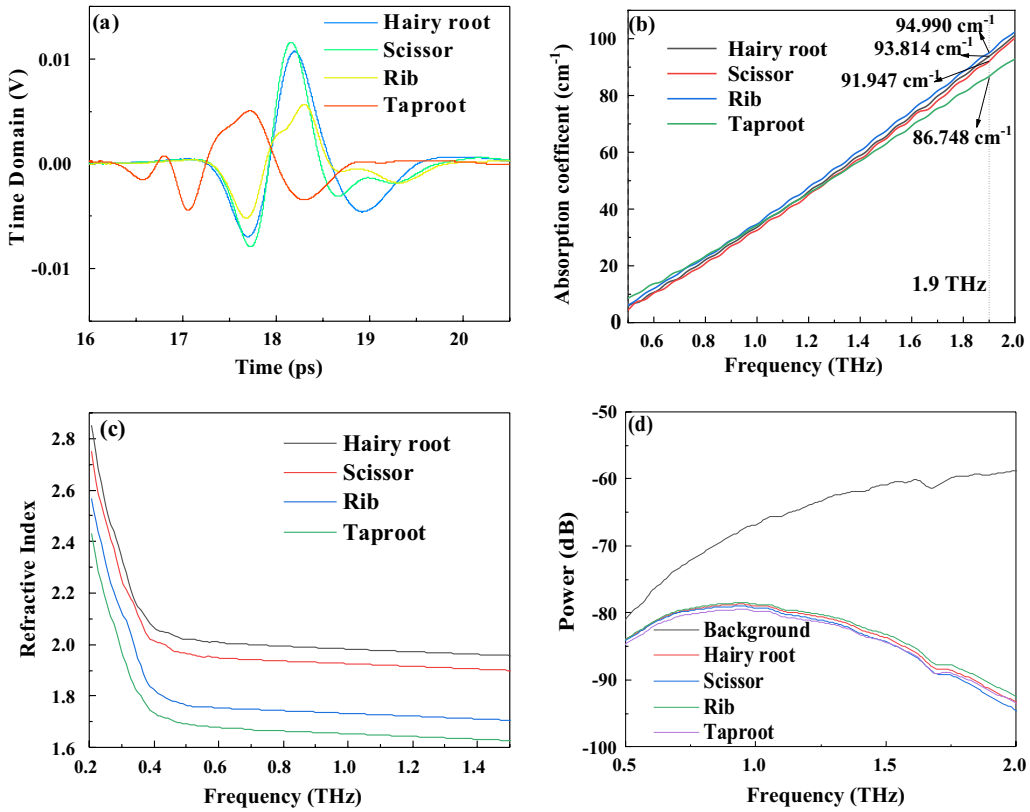

Fig. 1 THz spectra of four different parts, **a** is time-domain spectroscopy; **b** is absorption coefficient spectrum in 0.5–2 THz; **c** is refractive Index in 0.2–1.5 THz; and **d** is frequency domain in 0.5–2 THz

example, it is found that the absorption coefficients of the main root, shear, hairy root, and tendon increase gradually and there are obvious differences, so we think that the absorption coefficient chosen for further analysis may obtain satisfactory results.

Figure 1c shows the average refractive index spectrum of the four samples in 0.5–1.5 THz. With the increase in frequency, the refractive index between samples decreases, and each of the absorption coefficients of the four samples is different. At the same frequency, the refractive index of the hair root is highest, and that of the taproot is lowest. Figure 1d shows the frequency-domain spectra of four samples. Compared with the background spectrum, the spectral energy decreases significantly after the THz range passing through the samples. The attenuation of the spectral power of four samples becomes more evident with the increase in frequency. In this paper, the THz absorption coefficient spectrum is used to establish the model.

Further analysis of the above figures, for the time-domain signals, the spectra of the main root, and tendon signals are significantly different from the hairy root and shear signals, so we can identify the main root and tendon samples. However, their spectral variation patterns are similar for the hairy root and shear samples, which can easily lead to misclassification. For the frequency-domain signal, the four types of sample power spectra have a similar evolutionary pattern and we can classify them by the differences in spectral lines. However, the scissor samples are blended with the hairy root samples at 0.5–0.75 THz, and then with the main root samples at 1.4–1.7 THz, which is highly susceptible to misclassification. In the case of refractive index, the magnitude is influenced by the homogeneity of the pressed sheet samples, but in practice, it is difficult to achieve absolute homogeneity within the pressed sheet samples, which introduces a slight bias.

All samples have their own intrinsic frequency. When the terahertz waves through a sample, the sample absorbs more radiation of the intrinsic frequency and resonates at this frequency, and the external energy is converted into energy within the molecule through resonance, in which case the amount of transmission is relatively small. Conversely, if a substance absorbs less radiation at an intrinsic frequency, the amount transmitted is greater. As the thickness of the pressed sample changes, the amount of transmission also changes, but the nature of the sample, which resonates at the intrinsic frequency (the point where absorption shows its strongest peak), does not change, as this is the nature of matter. As in Fig. 1b, we have added a reference line at 1.9 THz, and it is found that the absorption coefficient is 94.990, 93.814, 91.947, and 86.748 for hairy root, scissor, rib, and taproot, respectively. Therefore, by using the frequency of the terahertz wave corresponding to the absorption peak, we can identify the types of sample, so this paper uses the THz absorption coefficient to build the model.

Feature information extraction

There are 200 frequency points in 0.5–2.0 THz. However, some frequency points do not correlate with the quality of the samples, and the irrelevant information reduces the model efficiency (Hu et al. 2017). In this paper, the SPA is used to select spectral characteristic variables to simplify the model. In the variable selection program, the number interval of variable selection is set from 10 to 50. As shown in Fig. 2, the frequency points are selected by SPA. As shown in Fig. 2a, a total of 19 characteristic frequency points are selected in the spectra without pretreatment. As shown in Fig. 2b, a total of 20 characteristic frequency points are selected in the spectra by baseline correction. As shown in Fig. 2c, a total of 21 specific frequency points are selected in the spectra by MSC. As shown in Fig. 2d, a total of 20 wavelength points are selected in the spectra by normalization.

Characteristic analysis of spectral variables

The contribution rate of each component can be obtained by PCA, and the cumulative contribution rate of each component is shown in Fig. 3. It is found that for the validation set, the cumulative contribution rate of PC3 is 97.9465%, the three components can reflect the characteristics of the sample good, and the contribution rate of PC4 has less change compared to PC3, while the cumulative contribution rate of PC5 is 98.4056%, so the PC3 and PC5 are used for further analysis.

The value of cumulative variance contribution of the absorption coefficient of each variable spectrum is proposed to assess whether the variable is good or bad to model. The value of cumulative variance contribution is more considerable, and the variable is more suitable to model. The principal component analysis (PCA) is used to calculate the value of incremental variance contribution of the absorption coefficient of each variable spectrum. Firstly, PCA is used for each principal component variable, and then, the values of cumulative variance contribution of three main components and five principal components are calculated, respectively. As shown in Fig. 4, the values of incremental variance contribution are more significant in the low-frequency band and the high-frequency band than others. The changing trend of the value of cumulative variance contribution with variable is similar between three principal components and five principal components, the result of three principal components shows that the value of incremental variance contribution is meager in the interval of 160–190

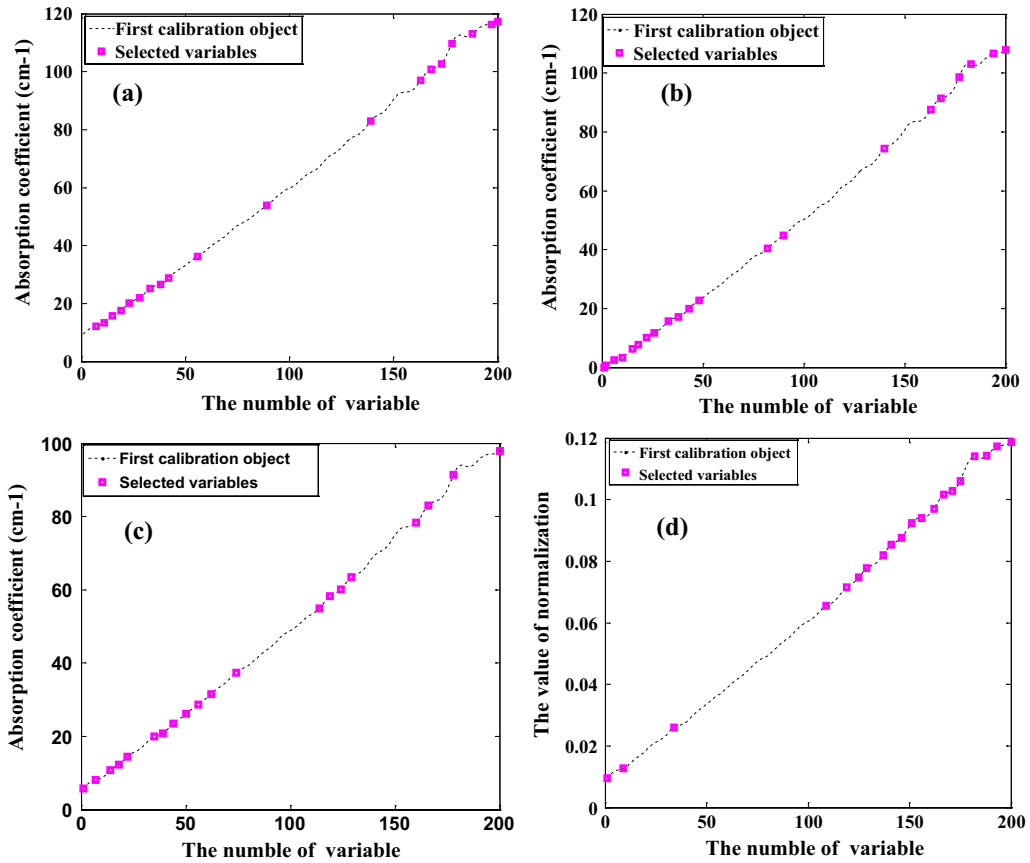

Fig. 2 The characteristic variable selection results of SPA, **a** is the variable selection without pretreatment; **b** is the variable selection with baseline correction; **c** is the variable selection with MSC; and **d** is the variable selection with normalization

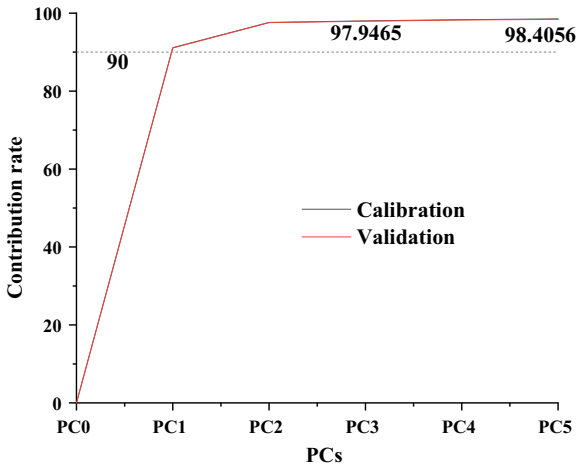

Fig. 3 Contribution rate of cumulative variables

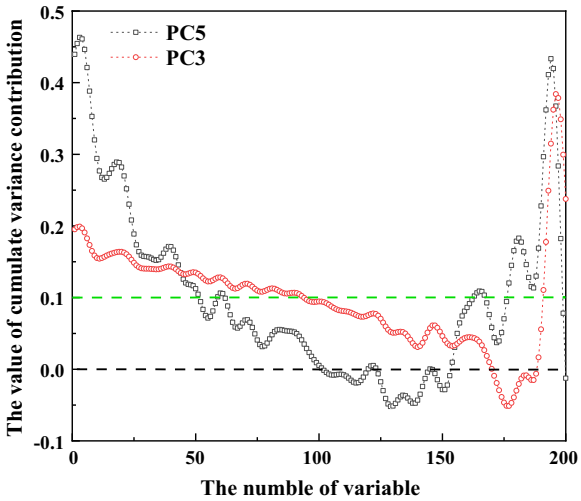

Fig. 4 Calculation of cumulative variance contribution of variables

variables, and the influence of five main features shows that in the variable interval of 110–160, the value of cumulative variance contribution is also merger. Next, the variables were selected by SPA which is input into SVM and LDA to establish models to verify whether the value of cumulative variance contribution of the absorption coefficient can be used to assess the model.

Qualitative analysis

In this paper, the SVM and LDA algorithm is used to establish the qualitative analysis models. Firstly, the modeling set and test set are divided at the ratio of 3:1. Then, the characteristic wavelength points of the THz absorption coefficient spectrum of 0.5–2.0 THz are used to establish the models. The model is validated by cross-validation, and some test samples are set aside for external validation of the model.

SVM classification model

The SVM algorithm is used to establish the qualitative analysis models. Firstly, the modeling set and test set are divided at the ratio of 3:1. Then, the characteristic wavelength points of the THz absorption coefficient spectrum of 0.5–2.0 THz selected by SPA are input to SVM to establish the models of cluster analysis. The grid search method is used to find the optimal parameters, which are the penalty factor c and kernel function g. The model is validated by cross-validation, and some test samples are set aside for external validation of the model. In this paper, the pretreatment methods, which are MSC, baseline correction, and normalization, are input into the SVM algorithm. The two kernel functions in SVM are used to build the model. The results of the models under different pretreatment methods and different kernel functions are compared. The results are shown in Table 2. In the LIN-SVM model, the model by baseline correction has the highest precision. In contrast, the model accuracy only improves a little compared with the LIN-SVM model without pretreatment, and the accuracy of LIN-SVM models with normalization and MSC pretreatment is lower than the LIN-SVM model without pretreatment. The correct rate of the model test set by baseline is 99.25%, the c is 3.162, the training accuracy of the model is 99.50%, the verification accuracy is 97.26%, and the correct rate is 99.25%. In the RBF-SVM models, the RBF-SVM model by MSC has the highest precision, the c is 21.544, g is 0.0681, the training accuracy of the RBF-SVM model by MSC is 99.75%, the accuracy of the verification set is 85.32%, and there are four classifications errors in 134 test samples, so the classification accuracy is 97.01%. Although the accuracy of the test set of the RBF-SVM model by baseline is 100%, the accuracy of the verification set is too low. Both two kernel functions can get good modeling results, but the performance of the LIN-SVM model is more stable and robust than the RBF-SVM model. At the same time, it is found that the effect of normalization pretreatment is unsatisfactory. Table 2 shows that the normalization pretreatment may cause severe loss of characteristic spectral information, so it is not suitable for the pretreatment of this batch absorption coefficient spectrum.

Figure 5 shows the optimal test set results of the SVM model test set under two kinds of kernel functions. Figure 5a shows the result of the optimal LIN-SVM model, it can be seen that one rib sample is wrongly divided into a hairy root sample, and the classification accuracy of it is 99.25%. Figure 5b shows the result of the optimal RBF-SVM model, it can be seen that four samples in the test set are misclassified, and the classification accuracy is 97.01%.

Table 2 The SVM model of results after band selection of absorption coefficient spectrum

Kernel function type	Pretreatment type	Parameter	Training accuracy / %	Verification accuracy / %	Number of test sets	Prediction accuracy / %
LIN	No	c = 0.46415	98.76	95.27	134	98.50
	Baseline correction	c = 3.1622	99.50	97.26	134	99.25
	MSC	c = 0.4641	92.04	84.83	134	97.01
	Normalization	c = 3.1622	44.30	44.03	134	25
RBF	No	c = 3.1622 g = 0.01	100	60	134	92.53
	Baseline correction	c = 21.5443 g = 0.01	100	68.65	134	100
	MSC	c = 21.5443 g = 0.0681	99.75	85.32	134	97.01
	Normalization	c = 21.5443 g = 146.7799	44.03	44.28	134	25

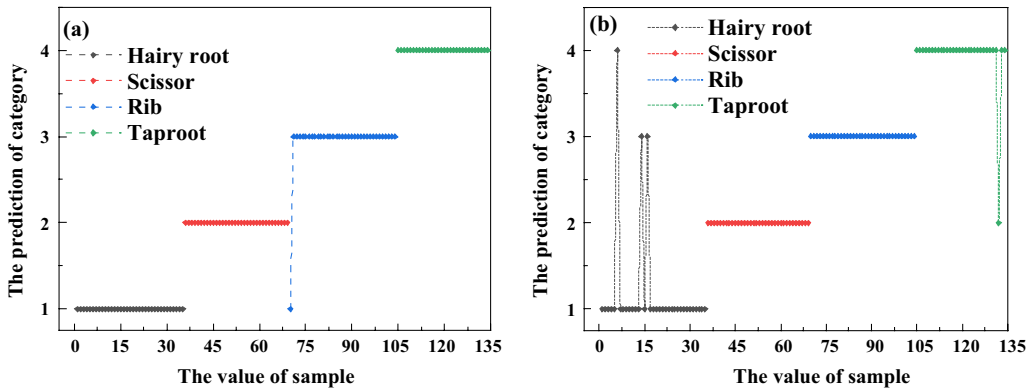

Fig. 5 Test results of SVM, **a** is the result in LIN-SVM with baseline correction, and **b** is the result in RBF-SVM with MSC

LDA classification model

The variables are selected, and LDA is used to establish classification models. After the models are established, the reserved test set samples are imported into the LDA classification models to evaluate them. Figure 6 shows the plane classification diagrams drawn by the first two discriminant functions (DF1, DF2) of modeling set samples. It is found that the distribution of modeling set samples of different classes all has obvious classification boundaries. As shown in Fig. 6, the distribution of the boundary of the model by baseline correction is more evident than others, and the accuracy of the modeling set is 99.25% without pretreatment; the accuracy of the modeling set is 99% by baseline correction, the accuracy of the modeling

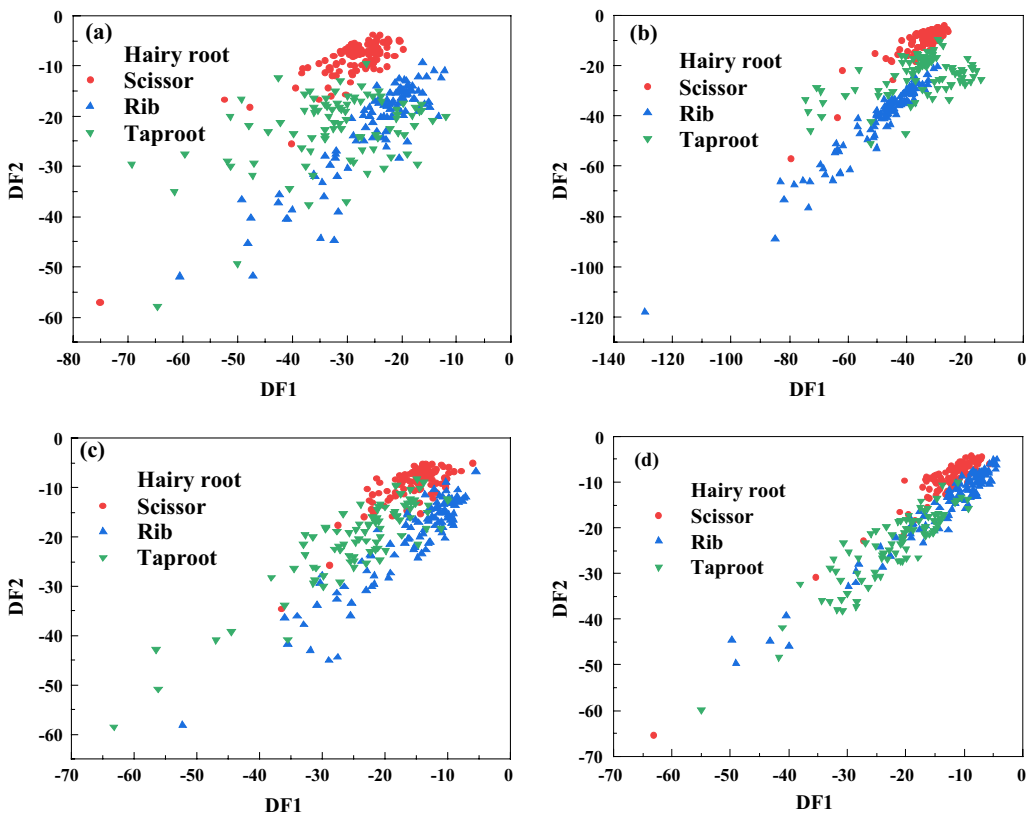

Fig. 6 Plane classification diagram of LDA model for modeling set samples, **a** is the model without pretreatment, **b** is the model with baseline correction, **c** is the model with MSC, and **d** is the model with normalization

set is 91.79% by MSC, and the accuracy of the modeling set is 82.84% by normalization.

To evaluate models, some of the samples are set aside. They are used for external model validation, 134 samples are used as a testing model for the experiment, and the test results are shown in Table 3; in the case of without pretreatment, the accuracy model of the test set is 99.25%; the accuracy of the test set by baseline correction is 100%; the accuracy of the test set by MSC is 25%; and the accuracy of the test set by normalization is 97.01%, but the accuracy of the modeling set is far lower than that of the test set, the phenomenon of overfitting is easy to exist.

As shown in Fig. 7, a sample is misclassified in the LDA model without pretreatment, the rest of the samples are classified correctly, and the classification accuracy of it is 99.25%; no sample is misclassified in the LDA model by baseline, and the classification accuracy of it is 100%.

Among the four pretreatment methods, the variables selected by SPA without pretreatment and the variables chosen by SPA with baseline correction are more consistent with the results in Fig. 4, Tables 2 and 3 show that the classification accuracies of baseline correction and no pretreatment are better than others, and this is that because MSC and normalization can choose some negative frequency points. Those results show that the value of cumulative variance contribution can be used to assess whether the variable is good or bad to model.

Table 3 The LDA model of results after band selection of absorption coefficient spectrum

Model type	Pretreatment type	Modeling accuracy / %	Number of test sets	Prediction accuracy / %
LDA	No	99.25	134	99.25
	Baseline correction	99	134	100
	MSC	91.79	134	25
	Normalization	82.84	134	97.01

Model evaluation

In this paper, the qualitative discrimination models for different parts of samples are established. The accuracy of the test set is used to evaluate the models. The optimal results of SVM and LDA are shown in Tables 2 and 3. It is found that both SVM and LDA models can achieve suitable identification, which shows that the THz-TDS technology is feasible to identify the different parts of *P. notoginseng*. The accuracy of the LIN-SVM model test set is 99.25%, and the accuracy of the LDA model test set is 100%, so the performance of the LDA model is slightly better than the SVM model.

Conclusion

The rapid identification of different parts of *P. notoginseng* is studied by THz spectroscopy. The THz transmission system is used to obtain the THz spectrum of the samples. The absorption coefficient spectrum of different *P. notoginseng* in 0.5–2.0 THz is used to establish classification models. Firstly, MSC, baseline correction, and normalization are used to pretreat the absorption coefficient spectrum, and then, SPA is used to select the characteristic variables from the absorption coefficient spectrum. The value of cumulative variance contribution is proposed to assess whether the variable is good or bad to model. Then, the selected bands are input into SVM and LDA to establish models, and the optimal parameters c and g of SVM are selected by grid search. In the SVM model, the model performance of LIN-SVM

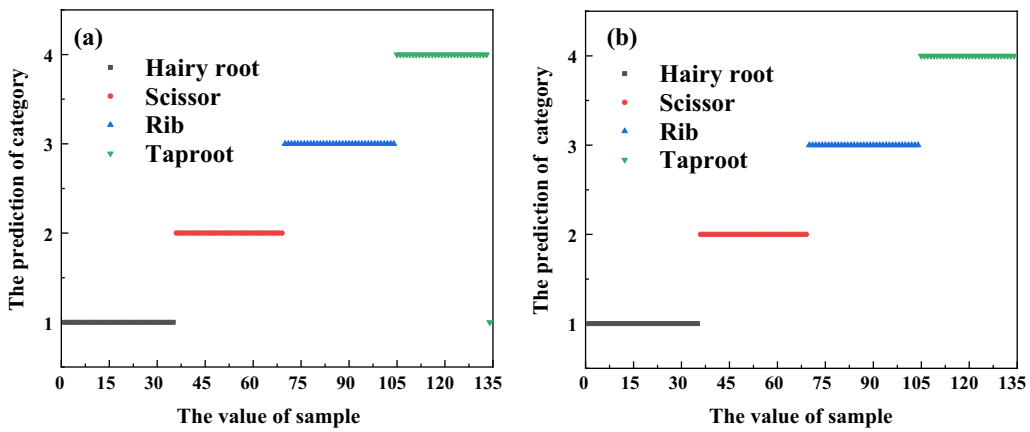

Fig. 7 Test results of the LDA model, **a** is the LDA model without pretreatment; **b** is the LDA model with baseline correction

with baseline correction is best, and the accuracy of it is 99.25%. In the LDA model, the result with baseline correction is best, the accuracy of the training set is 100%, and the accuracy of the test set is 100%; this shows that the four different parts of the samples can be separated by LDA well. Therefore, the combined THz-TDS and chemometrics method can be used to identify different parts of *P. notoginseng* rapidly, accurately, and nondestructively. This study provides a reference for the application of THz-TDS in the rapid detection of food.

Acknowledgements
The authors gratefully acknowledge the financial support provided by the National Natural Science Foundation of China (Project No. 12103019).

Author contributions
BL, ZH, and HC contributed to writing—original draft, writing—review and editing, and experiment; BL, ZH, and OYA were involved in experimental scheme design; and ZH and AY contributed to review and editing and experiment.

Funding
This study received financial support provided by the National Natural Science Foundation of China (Project No. 12103019).

Availability of data and materials
The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participation
This article does not contain any studies with human participants or animals by any of the authors.

Competing interests
Li Bin declares that he has no conflict of interest. Han Zhao-yang declares that he has no conflict of interest. Cai Hui-zhou declares that he has no conflict of interest. Ou Yang Ai-guo declares that he has no conflict of interest.

Received: 22 February 2022 Accepted: 19 May 2022
Published online: 27 May 2022

References

Abdi H, Williams LJ. Principal component analysis. Wiley Interdiscip Rev Comput Stat. 2010;2:433–59.
Baek SH, Lim HB, Chun HS. Detection of melamine in foods using terahertz time-domain spectroscopy. J Agric Food Chem. 2014;62(24):5403–7.
Bin L, Bing L, Yan-de L, et al. Detection of adulteration of kudzu powder by terahertz time-domain spectroscopy. J Food Meas Charact. 2021;15(5):4380–7.
Cao B, Li H, Fan M, et al. Determination of pesticides in a flour substrate by chemometric methods using terahertz spectroscopy. Anal Methods. 2018;10(42):5097–104.
Cao C, Zhang Z, Zhao X, et al. Terahertz spectroscopy and machine learning algorithm for non-destructive evaluation of protein conformation. Opt Quant Electron. 2020;52(4):1–18.
Ge H, Jiang Y, Xu Z, et al. Identification of wheat quality using THz spectrum. Opt Express. 2014;22(10):12533–44.
Hong LQ, Wang SQ. Identification standard and pharmacological analysis of *Panax notoginseng*. J Froniers Med. 2018;8(18):339–339.
Hu X, Liu W, Liu C, et al. Rapid identification of producing area of coffee bean based on terahertz spectroscopy and support vector machine. Trans Chin Soc Agric Eng. 2017;33(9):302–7.
Huang LQ, Yuan Y, Yuan QJ, et al. Key problems in development of molecular identification in traditional Chinese medicine. China J Chin Materia Med. 2014;39(19):3663–7.
Li WT, Chen M. Simultaneous determination of 508 pesticide residues in *Panax notoginseng* by liquid chromatography-tandem mass spectrometry. J Instrum Anal. 2019;38(7):761–74.
Li C, Li B, Ye D. Analysis and Identification of rice adulteration using terahertz spectroscopy and pattern recognition algorithms. IEEE Access. 2020;8:26839–50.
Liu J, Li Z, Hu F, et al. Method for identifying transgenic cottons based on terahertz spectra and WLDA. Optik Int J Light Electron Opt. 2015;126(19):1872–7.
Liu J, Xie H, Zha B, et al. Detection of genetically modified sugarcane by using terahertz spectroscopy and chemometrics. J Appl Spectrosc. 2018;85(1):119–25.
Liu W, Zhang Y, Li M, et al. Determination of invert syrup adulterated in acacia honey by terahertz spectroscopy with different spectral features. J Sci Food Agric. 2020;100(5):1913–21.
Long Z, Chun L, Tianying L, et al. Classification of calculus bovis and its confounding substances based on terahertz time-domain spectroscopy. Laser Optoelectron Prog. 2020;57(23):233001.
Meng XS, Jiang L, Yu YY, et al. Micro-macroscopical Identification of *Panax Notoginseng* Flower and Its Adulterants. Chin J Exp Tradit Med Formulae. 2018;24(11):39–43.
Peng Y, Shi C, Xu M, et al. Qualitative and quantitative identification of components in mixture by terahertz spectroscopy. IEEE Trans Terahertz Sci Technol. 2018;8(6):696–701.
Shen T, Li W, Zhang X, et al. High-sensitivity determination of nutrient elements in *Panax notoginseng* by laser-induced breakdown spectroscopy and chemometric methods. Molecules. 2019;24(8):1525.
Shi C, Zhu J, Xu M, et al. An approach of spectra standardization and qualitative identification for biomedical materials based on terahertz spectroscopy. Sci Program. 2020. https://doi.org/10.1155/2020/8841565.
Xiao-Li Z, Jiu-Sheng L. Terahertz spectroscopic investigation of Chinese herbal medicine. J Phys Conf Ser. 2011;276(1):012233.
Yan L, Liu C, Qu H, et al. Discrimination and measurements of three flavonols with similar structure using terahertz spectroscopy and chemometrics. J Infrared Millim Terahertz Waves. 2018;39(5):492–504.
Yun LI, Zhang J, Fei LIU, et al. Prediction of total polysaccharides content in P. notoginseng using FTIR combined with SVR. Spectrosc Spectr Anal. 2018;38(6):1696.
Yun LI, Zhang J, Hang JIN, et al. Study of the Underground Parts Identification and Saponins Content Prediction of *Panax Notoginseng* Based on FTIR Combined with Chemometrics. Spectrosc Spectr Anal. 2019;39(1):103.
Yun W, Qin J, Jia S, et al. Detection of Trichlorfon in soil using THz-FDS. Spectrosc Spectr Anal. 2020;40(6):1791.
Zhang D, Xu L, Wang Q, et al. The optimal local model selection for robust and fast evaluation of soluble solid content in melon with thick peel and large size by vis-NIR spectroscopy. Food Anal Methods. 2019;12(1):136–47.
Zhao Y, Han B, Peng H, et al. I dentification of "H uoshan shihu" Fengdou: comparative authentication of the Daodi herb Dendrobium huoshanense and its related species by macroscopic and microscopic features. Microsc Res Tech. 2017;80(7):712–21.
Zheng P, Shi M, Wang J, et al. The spectral emission characteristics of laser induced plasma on tea samples. Plasma Sci Technol. 2015;17(8):664.
Zhou ZL, Li J, Huang SQ, et al. Development of chemometric modelling in the application of NIR to the quality control of Chinese herbal medicine: literature review and future perspectives. Chem Ind Eng Prog. 2016;35(6):1627–45.

Publisher's Note