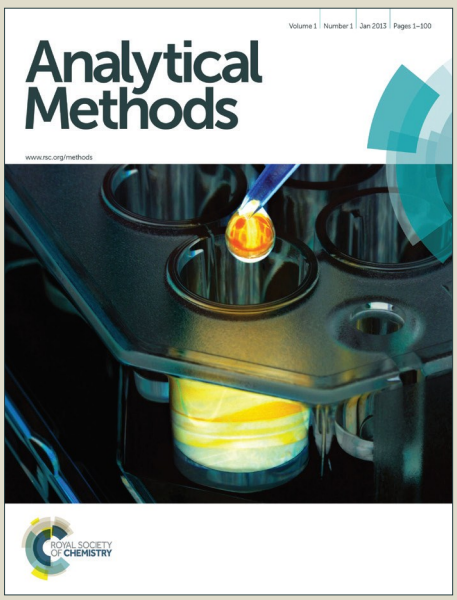
Analytical Methods

Accepted Manuscript



This is an *Accepted Manuscript*, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. We will replace this *Accepted Manuscript* with the edited and formatted *Advance Article* as soon as it is available.

You can find more information about *Accepted Manuscripts* in the Information for Authors.

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard Terms & Conditions and the Ethical guidelines still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this *Accepted Manuscript* or any consequences arising from the use of any information it contains.

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/methods

Rapid determination of total polyphenols content and antioxidant activity in *Dendrobium officinale* by near-infrared spectroscopy

Longhui Ma[a], Zhimin Zhang[a], Xingbing Zhao[b], Sufeng Zhang[b], Hongmei Lu[a*]

[a]*College of Chemistry and Chemical Engineering, Central South University, Changsha 410083, PR China*

[b]*Hunan Longshishan Dendrobium Candidum Wall.ex Lindl Base Co., Ltd, Changsha 410205, PR China*

Abstract: Despite their popularity and extensive use, some herbs have not been officially recognized in most countries. The main reason is the lack of comprehensive research data and methods. In this paper, a rapid approach based on near-infrared spectroscopy (NIR) was developed for the determination of total polyphenols content (TPC) and antioxidant activity (AA) in *Dendrobium officinale* (*D. officinale*), an important Chinese herb. Adopting the Folin-Ciocalteu (FC) assay and 2,2-diphenyl-1-picrylhydrazyl radical (DPPH) free radical scavenging activity as the reference methods, TPC and AA in *D. officinale* samples (n=83) collected from different locations in China were analyzed. Spectra generated by NIR were pretreated with different preprocessing methods and analyzed with partial least-square (PLS). To obtain robust and predictive quantitative model, competitive adaptive reweighted sampling (CARS) was applied to screen the key variables. The correlation coefficient of prediction (R_{pre}^2) and root mean square error of prediction (RMSEP) by competitive adaptive reweighted sampling - partial least-square (CARS-PLS) were 0.8412, 0.2905 for TPC and 0.9062，0.1028 for AA, respectively. The results show that the combination of NIR spectroscopy with CARS-PLS provides a rapid and precise alternative to existing chemical analysis for the determination of TPC and AA in *D. officinale*.

Keywords: Polyphenols; Antioxidant activity; Near-infrared spectroscopy; *Dendrobium officinale*

* Corresponding author.*E-mail address:* hongmeilu@csu.edu.cn

1. Introduction

Dendrobium officinale (D. officinale) is ranked "the first of the nine Chinese fairy herbs" and has been widely used as pharmaceuticals or functional foods for thousands of years in China.[1, 2] Despite its popularity and extensive use during the last decade, it has not been officially recognized in most countries. It is far from sufficient to meet the criteria needed to support its use world-wide. The main reason is the lack of comprehensive research data and methods about component, activity and safety etc. Previous studies on *D. officinale* mainly focused on polysaccharides, due to the content of polysaccharides used as one of the quality assessment criteria (no less than 0.2500 g of glucose per g dry weight) in Chinese pharmacopoeia.[3] Barely researches focus on determining the polyphenol, another kind of the main components in *D. officinale*, and its activity.[4] It is well-known that polyphenol compounds are associated with reducing risk of developing chronic diseases, such as aging, chronic gastric, various cancer and cardiovascular.[5, 6] Moreover, many studies have demonstrated that polyphenol compounds also contribute to the antioxidant activity.[7, 8] It is important and necessary to qualitatively and quantitatively study polyphenol compounds and antioxidant activity in *D. officinale*. However, existing analytical methods such as colorimetric measurement[9, 10] and HPLC measurement[11] involve tedious sample preparations, intensive labor and expensive solvents,[12] which make them unsuitable for routine analysis.

Recently, our group have successfully applied near-infrared spectroscopy (NIR) for rapid authentication or identification of some herb and food[13, 14] and precise

prediction of component content.[15, 16] NIR has several desirable advantages over existing analytical methods: e.g. non-invasion, cost-efficiency, less laboring and high efficiency.[17] As the absorption of NIR spectra correspond to molecular overtone and combination vibrations, the absorption peaks are broad and strongly overlapped, which is hard to interpret. In such case, partial least-squares (PLS) regression has been frequently employed to make a calibration model with spectral data .[18] Since some spectral regions contain noise from environment and interference variables, better calibration model can be obtained by selecting effective variables instead of the full-spectrum.[19, 20] With the rapid calculation process, competitive adaptive reweighted sampling (CARS) variable selection method has been proposed by Li et al lately.[21] And lots of literatures have successfully proved that CARS is a powerful and high-performance tool in complex analytical system.[22, 23]

In this study, NIR spectroscopy together with CARS variables selection and PLS regression was used to simultaneous determination of total polyphenols content (TPC) and antioxidant activity (AA) in *D. officinale*. To the best of our knowledge, this approach has never been tried before to determinate the *D. officinale* TPC and AA. The specific procedure was outlined as follows: (1) adopting existing chemical analysis methods as the reference methods to analyze TPC and AA in *D. officinale* samples; (2) comparing different spectral data preprocessing methods and screening the key variables using CARS algorithm; (3) establishing the quantitative models for determination of TPC and AA based on NIR spectra.

2. Materials and methods

2.1. Materials and Reagents

A total of 83 *D. officinale* samples were collected from different locations in China during the period from April 2012-April 2014. It provided a representative set of *D. officinale* consumed in domestic market, which comprised enough variation to make the quantitative model robust. Folin-Ciocalteu reagent was purchased from Sigma Aldrich (St. Louis, Mo, USA). 2,2-diphenyl-1-picrylhydrazyl radical (DPPH), gallic acid (99% purity) and 6-hydroxy-2,5,7,8-tetramethyl-2-carboxylic acid (Trolox) were obtained from the National Institution for Food and Drug Control (Beijing, China). Dehydrated alcohol and anhydrous sodium carbonate (Na_2CO_3) were purchased from Sinopharm Chemical Reagent Co.Ltd (Shanghai, China). Deionized water was purified with a Milli-Q system (Millipore, Bedford, MA, USA).

2.2. Reference analysis

The dried samples were ground and passed through a 60 mesh sieve. An accurate weight (1.0 g) of each powder sample was mixed with 30 mL 80% anhydrous ethanol. The mixture was then sonicated for 30 min at 35 ℃. After the ethanol extract was filtered, the filtrate was moved into 50 mL volumetric flask. Prepared extracts were stored at 4 ℃. TPC was determined by an improved Folin-Ciocalteu colorimeter[24] and expressed as mg of gallic acid equivalent per 1g of dry weight sample. Reference measurement of AA in *D. officinale* was assayed by the improved DPPH radical scavenging activity.[25] The UV-Vis spectra of *D. officinale* were acquired on a BTT miniature array spectrophotometer (B&W Tek, Newark, DE, USA) equipped with

glass or quartz cells of 1 cm path length in the range 200-800 nm.

2.3. NIR spectra collection

NIR spectra of sample powders were acquired by a diffuse reflectance mode using the Antaris Ⅱ Fourier transform-NIR System (Thermo Scientific Inc, Madison, USA). The number of scans was 32 and the spectral resolution was 8 cm^{-1}. The range of spectra was from 10000 to 4000 cm^{-1} and the data were measured in 3.9 cm^{-1} interval, which resulted in 1557 variables. Each sample was scanned three times in a ring cup and the average spectra were collected for subsequent analysis.

2.4. Chemometric methods

2.4.1. Spectral preprocessing and outlier detection

Since physical variations, such as particle size and shape, sample packing and sample surface, could impact on spectra measurement, the raw spectra inevitably consist of systematic noises or background information.[26] Accordingly, a proper spectral preprocessing method is necessary to reduce the unwanted spectral variations. In this study, different kinds of spectral preprocessing methods were compared, including smoothing, standard normal variable (SNV), multivariate scatter correction (MSC), Savitzky-Golay first-derivative (SG1), and the combinations of SNV (or MSC) with the derivative. Smoothing is an averaging algorithm that used to reduce the noise and to enhance the signal-noise ratio (SNR). MSC or SNV method is always performed to remove slope variation and to modify scatter effects. By calculating SG1 derivative, the baseline drift are eliminated and small spectral differences are enhanced.

In general, outliers are incorrect or abnormal ones in some sense compared to the majority of the data. Outliers in the calibration set would lead to severe errors on the model, while outliers in the prediction set would obtain misleading results to evaluate the model performance. Considering the calibration model is so sensitive to outliers, Monte-Carlo (MC) method[27] was applied to eliminate outliers in spectral dataset before quantitative analysis. The core idea of the MC outlier detector is to develop a Monte-Carlo procedure for detecting outlier by studying the distribution of prediction errors of each sample obtained from original data set. The detailed description of scheme and procedure for this method based on predictive errors and Monte-Carlo sampling has been shown in the Ref.[27]

2.4.2. PLS model and evaluation

The quantitative models of TPC and AA in *D.officinale* were developed by PLS regression. As a well-known method, it is used to establish relationships between spectra data matrix (**X**) and reference concentration of elements matrix (**y**) with a small number of latent variables (LV).[18, 28] PLS follows a two-step strategy to establish a functional relationship between **X** and **y**. Firstly, the input variable matrix **X** (n, m) and output variable matrix **y** (n, 1) are decomposed as followed:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathbf{T}} + \mathbf{E} \quad (1)$$

$$\mathbf{y} = \mathbf{T}\mathbf{q}^{\mathbf{T}} + \mathbf{f} \quad (2)$$

where **P** (m, k) and **q** (k, 1) are the loadings, **T** (n, k) is scores matrix, **E** (n, m) and **f** (n, 1) are error terms which are not explained by the model and k is the number of LV used in the PLS model.

Subsequently, a least squares regression is performed on the extracted orthogonal latent variables/score vectors. Input and output scores vectors are related by a multiple lineal regression model:

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} + \mathbf{f} \quad (3)$$

where b (m,1) is an inner coefficient.

All of the samples were divided into two sets by Duplex algorithm: one for calibrating model and the other for prediction ability. Duplex algorithm is a method for the selection of a representative set of samples, in which calibration and prediction objects are selected alternately, starting with the inclusion of the most distant pair of objects into the calibration set. [29]

The root mean square error of calibration (RMSEC), correlation coefficient of calibration (R_c^2), the root mean square error of prediction (RMSEP) and correlation coefficient of prediction (R_{pre}^2) were calculated to evaluate the performance of the final quantitative models. Generally, a robust and accurate model should have low values of RMSEC and RMSEP and high values of R_c^2 and R_{pre}^2. Besides, these parameters are defined as follows:

$$\text{RMSEC or RMSEP} = \sum_{i=1}^{n} \sqrt{\frac{(\hat{y} - y_i)^2}{n}} \quad (4)$$

$$R_c^2 \text{ or } R_{pre}^2 = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}} \quad (5)$$

where y_i is the measured value and $\hat{y}_i$ is the prediction value; $\bar{y}$ is the average measurement.

2.4.3. Key variables selection

CARS algorithm was applied to locate an optimal combination of the variables for accurately determination of TPC and AA. The major idea of CARS is to employ the mechanic of "survival of the fittest" based on Darwin's Evolution.[21] Absolute values of regression coefficients of PLS model are used as an index for evaluating the importance of each variable. Then, according to the importance level of each variable, CARS sequentially selected N subsets of variables from N Monte Carlo (MC) sampling run in an iterative and competitive manner. Next, the exponentially decreasing function (EDF) and adaptive reweighted sampling (ARS) are employed to eliminate the variables, which are of relatively small absolute regression coefficients by force. Finally, the subset with the lowest root mean square error of cross validation (RMSECV) is considered as optimal combination of the variables.

2.5. Software

All algorithms were implemented with MATLAB for Windows (Version 2013A, the MathWorks, Inc). The code of CARS was available as open source software in the website: http://www.libpls.net/.

3. Results and discussion

3.1. TPC and AA measure results using the reference methods

The TPC and AA of 83 samples were determined using the reference methods (see Section 2.2). The content ranges of TPC (mg/g) and AA (mg/g) are shown in Table 1. Obviously, different geographical origins and harvest seasons lead to considerable variability of TPC and AA in *D. officinale*. The content ranges of

reference values in calibration set are 0.3148-1.5922 (mg/g) for TPC and 1.3606-3.8962 mg/g for AA. The content ranges of TPC and AA in prediction set are 0.2332-1.4835 mg/g and 1.1995-4.3435 mg/g, respectively. There are no significant dissimilarities to be observed in means, ranges and standard deviations (SD) between calibration set and prediction set. This indicates that the Duplex algorithm enable the same diversity in both sets.

Insert Table 1

Insert Fig. 1

3.2. Selection of preprocessing methods

Fig.1 shows the raw NIR spectra of 83 *D. officinale* samples at wavenumbers 10000-4000 cm^{-1}. NIR spectra in this region contain some intensive spectral peaks. These intensive peaks correspond to the vibration of some groups such as the combination of C-H and C-C (4000 cm^{-1}), the second overtone vibration of the carbonyl group (5350 cm^{-1}), the first overtone of O-H and N-H (6900 cm^{-1}), stretching and deformation vibrations of C–H (7200 cm^{-1}). It is apparently difficult to discriminate samples just by visually examining the raw average spectra. Hence, preprocessing methods are critical to enhance the quality of spectra. Five spectral preprocessing methods were applied and compared on the basis of RMSECV. The results are showed in Table 2. According to the model evaluation standard, different preprocessing methods are chosen for TPC and AA. Smoothing + MSC is the best choice for the developing of TPC model, while SNV + SG1 derivative is the best one for the developing of AA quantitative model. The reason is that weights of

wavenumber are different when building PLS models for TPC and AA. SG1derivative algorithm for TPC calibration model may increase noise, whereas this pretreatment method may reflect more information corresponding to AA.

Insert Table 2

3.3. Outlier detection for robust models

The method based on Monte-Carlo was performed to detect outliers.[27] Three types of outliers should be considered, namely outliers in X, outliers in y, as well as outliers towards the model. Fig. 2a and Fig. 2b show the diagnostic plots of the outlier detection for TPC and AA respectively. As shown in Fig. 2, a threshold of standard deviation was set on the top left area. The samples were outliers in X if their standard deviation is larger than the threshold. In the example, object 28 in Fig. 2a and four objects (10, 11, 27, 28) in Fig. 2b are X outliers. The *y* outliers appear on the lower right area, namely objects (39, 42) in Fig. 2a and objects (9, 12) in Fig. 2b, which could cause a large error sum of squares. In addition, it can be seen that object 29 on top right area of the two figures is outlier towards the model. Clearly, both objects 28 and 29 decrease accuracy of the model and affect the subsequent ability to determine the TPC and the AA. These two outliers may be caused by measurement errors when collected the NIR spectra. Hence, 4 samples (28, 29, 39, 42) and 6 samples (10, 11, 12, 27, 28, 29) are removed for TPC and AA, respectively. After samples outlier detection, there are 79 and 76 samples for TPC and AA analysis, respectively. Then the remaining samples are divided using the Duplex algorithm.

Insert Fig. 2

3.4. Variable selection

As stated above, quantitative models are first established based on full spectrum (10000-4000 cm^{-1}). The results are satisfying (Table 3**).** RMSEP are 0.3242, 0.1232 and R_{pre}^2 are 0.8023, 0.8653 for TPC and AA, respectively. However, the input variables are too much, which affect the robust of the quantitative models and increase the calculation time. Thus, CARS has been applied to screen key variables prior to application of PLS and improve the model performance.

In this study, the number of MC sampling runs was set to 100 as default during the calculation process. In order to guarantee the reliability of the model, the CARS procedure was conducted 100 times and RMSECV values were recorded. The optimal variable subsets were selected for further analysis according to the minimum RMSECV. Finally, the obtained numbers of key variables are decreased from 1557 to 26 for TPC analysis and from 1557 to 34 for AA analysis, respectively. Under the selected key variables and spectral preprocessing method, PLS models for TPC and AA are established. The results of both full spectrum-PLS models and competitive adaptive reweighted sampling - partial least-square (CARS-PLS) are listed in Table 3. The R_c^2 values of two methods for TPC are very close. However, CARS-PLS has smaller RMSEP (0.2905) and higher R_{pre}^2 (0.8412), indicating that CARS-PLS models for TPC has better predictive ability than full spectrum-PLS. In terms of CARS-PLS model for determining the AA, the R_c^2 and R_{pre}^2 are significantly increased to 0.9854 and 0.9062; RMSEC and RMSEP are reduced to 0.0294 and 0.1028. In short, CRAS could eliminate uninformative variables effectively and improve the predictive precision of the model to a certain extent.

Fig. 3 shows the relationship between NIR predicted values and measure values for TPC and AA obtained by PLS calibration models combining CARS method. The diamond marked points referred to calibration set, and round marked points referred to prediction set. With a close observation, TPC model is a little worse compared with the AA model, because some points in Fig. 3a fall off the bisectrix line. The reason may be that Folin–Ciocalteu assay for TPC is a less accurate or precise method than DPPH radical scavenging assay for AA. Therefore, it is difficult to achieve a more robust model for determination of TPC. In general, the results suggest that PLS models based on CARS for determination TPC and AA are accurate compared with full spectrum-PLS models.

Insert Fig.3

Insert Table 3

3.5. Interpretation of key variables

The positions of selected key variables are illustrated by marked points in Fig. 4. The polyphenol compounds contain abundant hydrogenous bonds (i.e. C-H and O-H groups, etc) and exhibit antioxidant activity. The information contained in selected spectral regions plays a crucial role in determination of TPC and AA. The selected variables for both TPC and AA prediction are mainly concentrated in the region of 7200-4000 cm^{-1} and 9990-8333 cm^{-1}. The former is assigned to the combination bands of the functional groups –C=O, N-H, C-H and C-C. The latter is the region of C-H third overtone and combination tone.[17, 23] For example, the selected region of around 4204 cm^{-1} is related to the combination tone of C-H and C-C stretching vibration, and

the region of 6940-7140 cm^{-1} is the first overtone of O-H.[30] One can see that the selected variables are associated with needful chemistry feature of the TPC and AA in *D. officinale*. These indicate that CARS method has the ability to screen the key and effective variables.

Insert Fig.4

4. Conclusion

NIR spectroscopy coupled with chemometric methods was successfully utilized for rapid quantification of the total polyphenols content (TPC) and antioxidant activity (AA) in *D. officinale*. The most suitable data preprocessing methods were smoothing + MSC and SNV + SG1 derivative for TPC and AA model, respectively. Combined with the corresponding preprocessing method, CARS was used for screening informative variables and reducing uninformative variables. Twenty-six variables were picked out of 1557 wavenumbers by CARS for the prediction TPC, and 34 variables for the prediction AA. CARS-PLS models achieved the optimal performance with R_{pre}^2 and RMSEP were 0.8412, 0.2905 for TPC and 0.9062, 0.1028 for AA, respectively. The results show that NIR spectroscopy combined CARS with PLS algorithm is able to determine TPC and AA in *D. officinale* in a fast, accurate and reliable way. Based on this study and previous studies, NIR spectroscopy coupled with chemometric methods is a rapid potent approach for simultaneous quantification of chemical constituent content and activities in herbs and foods. These results also provide some fundamental research data and method for *D. officinale*. In the future, more samples from different geographical areas will be studied to obtain a model with

the greatest applicability.

Acknowledgments

The authors gratefully thank the National Natural Science Foundation of China for support of the projects (No. 21175157, 21375151 and 21305163).

References

1. X. Lin, P.-C. Shaw, S. C.-W. Sze, Y. Tong and Y. Zhang, *International immunopharmacology*, 2011, **11**, 2025-2032.
2. G. Ding, G. Xu, W. Zhang, S. Lu, X. Li, S. Gu and X.-Y. Ding, *European Food Research and Technology*, 2008, **227**, 1283-1286.
3. Chinese Pharmacopoeia Committee, *Chinese Pharmacopoeia,* People's Medical Publishing House, Beijing, 2010.
4. Y. Li, C.-L. Wang, Y.-J. Wang, S.-X. Guo, J.-S. Yang, X.-M. Chen and P.-G. Xiao, *Chemical and Pharmaceutical Bulletin*, 2009, **57**, 218-219.
5. X. Zhang, J.-K. Xu, J. Wang, N.-L. Wang, H. Kurihara, S. Kitanaka and X.-S. Yao, *Journal of Natural Products*, 2007, **70**, 24-28.
6. Y. Cai, Q. Luo, M. Sun and H. Corke, *Life sciences*, 2004, **74**, 2157-2184.
7. M. P. Kähkönen, A. I. Hopia, H. J. Vuorela, J.-P. Rauha, K. Pihlaja, T. S. Kujala and M. Heinonen, *Journal of agricultural and food chemistry*, 1999, **47**, 3954-3962.
8. J. Švarc-Gajić, Z. Stojanović, A. S. Carretero, D. A. Román, I. Borrás and I. Vasiljević, *Journal of Food Engineering*, 2013, **119**, 525-532.
9. L. M. Magalhães, M. A. Segundo, S. Reis and J. L. Lima, *Analytica chimica acta*, 2008, **613**, 1-19.
10. S. Georgé, P. Brat, P. Alter and M. J. Amiot, *Journal of Agricultural and Food Chemistry*, 2005, **53**, 1370-1373.
11. M.-S. Lee, Y.-S. Hwang, J. Lee and M.-G. Choung, *Food chemistry*, 2014, **158**, 351-357.
12. H. Yan, B.-x. Han, Q.-y. Wu, M.-z. Jiang and Z.-z. Gui, *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 2011, **79**, 179-184.
13. Y. Wei, W. Fan, X. Zhao, W. Wu and H. Lu, *Analytical Letters*, 2014, **48**, 817-829.
14. X. Wang, J. Huang, W. Fan and H. Lu, *Analytical Methods*, 2015, **7**, 787-792.
15. Y. Yun, Y. Wei, X. Zhao, W. Wu, Y. Liang and H. Lu, *RSC Advances*, 2015.
16. Q. Luo, Y. Yun, W. Fan, J. Huang, L. Zhang, B. Deng and H. Lu, *RSC Advances*, 2015, **5**, 5046-5052.
17. X. B. Zou, J. W. Zhao, M. J. Povey, M. Holmes and H. P. Mao, *Analytica chimica acta*, 2010, **667**, 14-32.
18. S. Wold, M. Sjöström and L. Eriksson, *Chemometrics and intelligent laboratory systems*, 2001, **58**, 109-130.

19. H.-D. Li, Y.-Z. Liang, X.-X. Long, Y.-H. Yun and Q.-S. Xu, *Chemometrics and Intelligent Laboratory Systems*, 2013, **122**, 23-30.
20. B.-C. Deng, Y.-H. Yun, P. Ma, C.-C. Lin, D.-B. Ren and Y.-Z. Liang, *The Analyst*, 2015, **140**, 1876-1885.
21. H. Li, Y. Liang, Q. Xu and D. Cao, *Analytica chimica acta*, 2009, **648**, 77-84.
22. D. Wu and D.-W. Sun, *Talanta*, 2013, **116**, 266-276.
23. X. Wei, N. Xu, D. Wu and Y. He, *Food and Bioprocess Technology*, 2013, **7**, 184-190.
24. K. Tawaha, F. Q. Alali, M. Gharaibeh, M. Mohammad and T. El-Elimat, *Food chemistry*, 2007, **104**, 1372-1378.
25. R. Scherer and H. T. Godoy, *Food chemistry*, 2009, **112**, 654-658.
26. J. W. Jin, Z. P. Chen, L. M. Li, R. Steponavicius, S. N. Thennadil, J. Yang and R. Q. Yu, *Analytical chemistry*, 2012, **84**, 320-326.
27. D. S. Cao, Y. Z. Liang, Q. S. Xu, H. D. Li and X. Chen, *Journal of computational chemistry*, 2010, **31**, 592-602.
28. P. Geladi and B. R. Kowalski, *Analytica chimica acta*, 1986, **185**, 1-17.
29. R. D. Snee, *Technometrics*, 1977, **19**, 415-428.
30. J. Workman Jr and L. Weyer, *Practical guide to interpretive near-infrared spectroscopy*, CRC press, Boca Raton, 2007.

Tables:

Table 1. TPC and AA measured with the existing chemical methods and the number of *D. officinale* samples used in dataset.

Constituent	Calibration set				Prediction set			
	S.N.[a]	Mean	S.D.[b]	Range	S.N.[a]	Mean	S.D.[b]	Range
TPC(mg/g)	60	2.2802	0.5634	1.3606-3.8962	19	2.4886	0.7177	1.1995-4.3435
AA(mg/g)	59	0.9535	0.2452	0.3148-1.5922	17	0.8594	0.3456	0.2332-1.4835

[a]S.N: sample number; [b]S.D: standard deviation.

Table 2. Comparison of different spectral preprocessing methods on performance of PLS calibration models.

Properties	Methods	PLS results			
		RMSECV	RMSEP	RMSEC	R_c^2
TPC	raw data	0.3916	0.4228	0.2440	0.8358
	Smoothing+ SNV	0.3705	0.4158	0.1906	0.8659
	Smoothing+ MSC	0.3287	0.3922	0.1837	0.8766
	SNV+SG1st	0.3639	0.4052	0.2148	0.8509
	MSC+SG1st	0.3617	0.4049	0.1977	0.8522
AA	raw data	0.2558	0.3109	0.1083	0.8641
	Smoothing+ SNV	0.2343	0.2386	0.1009	0.8761
	Smoothing+ MSC	0.2464	0.2841	0.1072	0.8626
	SNV+SG1st	0.2280	0.2291	0.0846	0.9182
	MSC+SG1st	0.2350	0.2394	0.0853	0.9094

Table 3. PLS model results based on full-spectrum and selected wavelengths using CARS.

properties	methods	Variable number	Calibration set		Prediction set	
			R_c^2	RMSEC	R_{pre}^2	RMSEP
TPC(mg/g)	PLS	1557	0.9121	0.1656	0.8023	0.3242
	CRAS-PLS	26	0.9185	0.1596	0.8412	0.2905
AA(mg/g)	PLS	1557	0.9488	0.0550	0.8653	0.1232
	CRAS-PLS	34	0.9854	0.0294	0.9062	0.1028

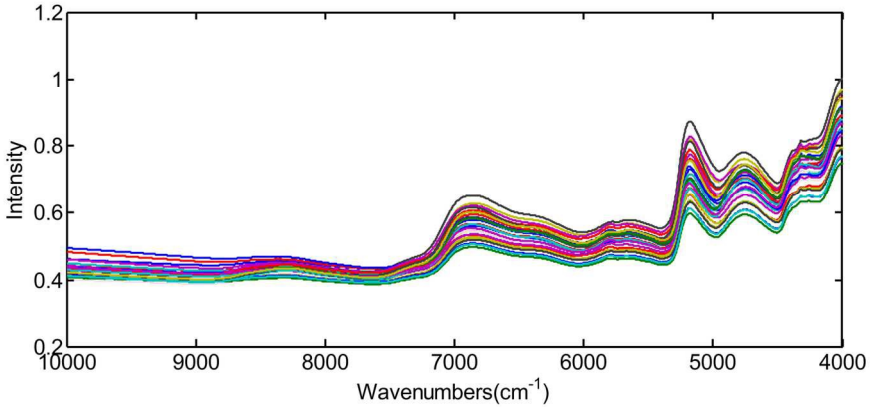
Figure captions

Fig. 1. NIR spectra of the *D.officinale* samples (n=83).

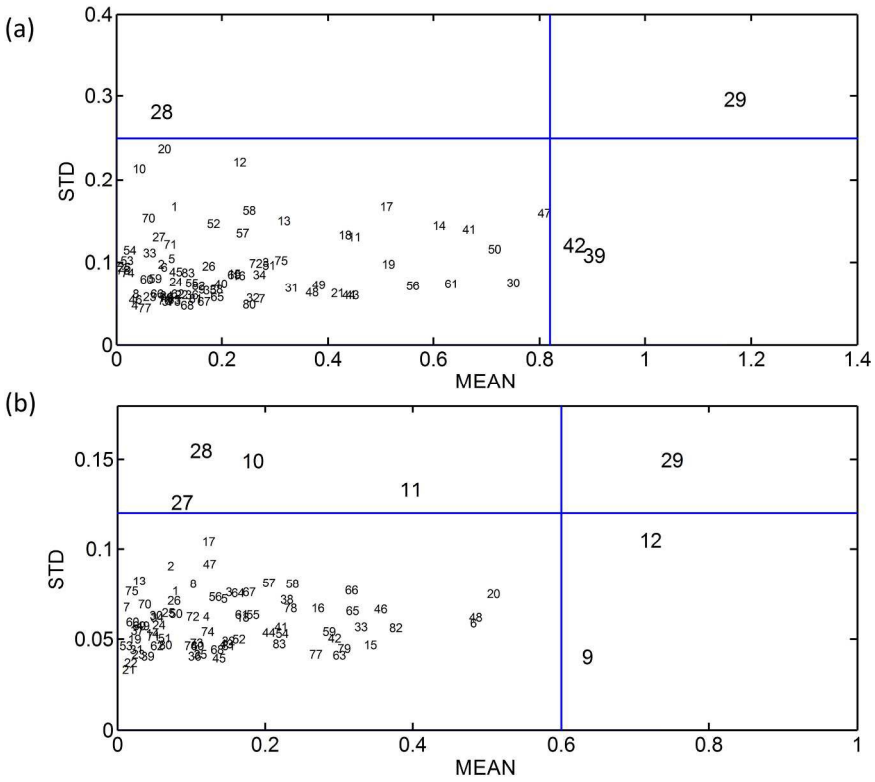
Fig. 2. Diagnostic plots for outlier detection based on MCS. (a) TPC; (b) AA.

Fig. 3. Correlation diagrams between the predicted values and measured values based on CARS-PLS method. (a) TPC; (b) AA.

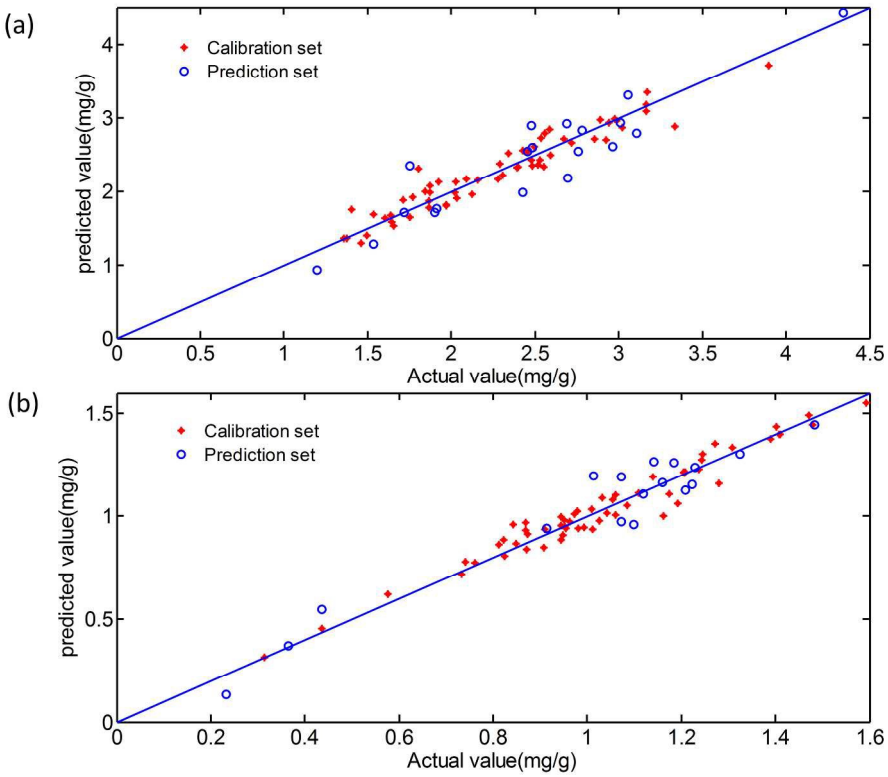
Fig. 4. Positions of variables selected by CARS-PLS for prediction of TPC (round marked points) and AA (diamond marked points) on the full spectra.



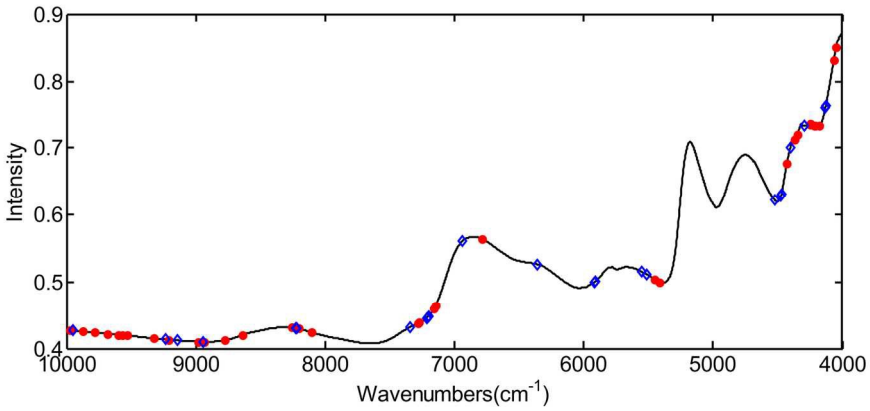
152x64mm (300 x 300 DPI)



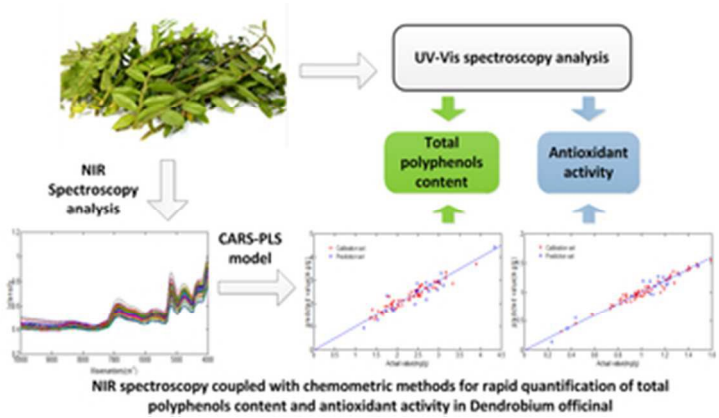
Diagnostic plots for outlier detection based on MCS. (a) TPC; (b) AA.
231x184mm (300 x 300 DPI)



Correlation diagrams between the predicted values and measured values based on CARS-PLS method.
(a)TPC; (b) AA
231x184mm (300 x 300 DPI)



Positions of variables selected by CARS-PLS for prediction of TPC (round marked points) and AA (diamond marked points) on the full spectra
152x64mm (300 x 300 DPI)



NIR spectroscopy coupled with chemometric methods for rapid quantification of total polyphenols content and antioxidant activity in Dendrobium officinale
36x17mm (300 x 300 DPI)