Journal of Innovative Optical Health Sciences
Vol. 7, No. 4 (2014) 1350061 (11 pages)
© The Authors
DOI: 10.1142/S1793545813500612

Wavelet-based classification and influence matrix analysis method for the fast discrimination of Chinese herbal medicines according to the geographical origins with near infrared spectroscopy

Wenlong Li and Haibin Qu*
Pharmaceutical Informatics Institute
Zhejiang University
No. 866, Yuhangtang Road
Hangzhou 310058, P. R. China
**quhb@zju.edu.cn*

Received 10 July 2013
Accepted 29 September 2013
Published 9 January 2014

A discriminant analysis technique using wavelet transformation (WT) and influence matrix analysis (CAIMAN) method is proposed for the near infrared (NIR) spectroscopy classification. In the proposed methodology, NIR spectra are decomposed by WT for data compression and a forward feature selection is further employed to extract the relevant information from the wavelet coefficients, reducing both classification errors and model complexity. A discriminant-CAIMAN (D-CAIMAN) method is utilized to build the classification model in wavelet domain on the basis of reduced wavelet coefficients of spectral variables. NIR spectra data set of 265 *salviae miltiorrhizae radix* samples from 9 different geographical origins is used as an example to test the classification performance of the algorithm. For a comparison, k-nearest neighbor (KNN), linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) methods are also employed. D-CAIMAN with wavelet-based feature selection (WD-CAIMAN) method shows the best performance, achieving the total classification rate of 100% in both cross-validation set and prediction set. It is worth noting that the WD-CAIMAN classifier also shows improved sensitivity, selectivity and model interpretability in the classifications.

Keywords: Discriminant analysis; near infrared spectroscopy; Chinese herbal medicines; variable selection; wavelet analysis.

1. Introduction

Near infrared (NIR) spectroscopy has the advantage of being fast, robust, nondestructive and especially suitable for the online application. With the development of modern instruments and chemometrics, NIR spectroscopy has been widely applied for the quantitative and qualitative analysis in large areas, such as agriculture, pharmaceuticals, food, textiles and polymer production.[1] However, the physical and chemical information cannot always be extracted straightforwardly from the spectra due to the existence of band overlapping, multicollinearity, poor signal-to-noise ratio, baseline fluctuations, and so on. Thus, in order to overcome these difficulties in NIR spectral analysis, chemometrics has to be used for preprocessing, modeling, validation, etc. A main part of chemometrics is multivariate data analysis, which is essential for qualitative and quantitative assays based on NIR spectroscopy. Statistical classification with NIR data has been used in a number of scientific publications and practical applications.[1,2]

Thanks to the modern techniques of analysis; objects described by a large number of variables (i.e., absorbance at defined wavelengths or wavenumbers) can be easily measured in a short time. However, for high-dimensional data, singularity problems arise if the variables are highly correlating or if the number of variables exceeds the number of samples available for analysis. In this situation, the most common classifier, one based on conventional linear discriminant analysis (LDA),[3] is of limited use. Then, multivariate chemometric techniques with a dimension reduction, e.g., principal component regression (PCR) and partial least squares (PLS),[4] become necessary in order to extract the most relevant information from spectroscopic data. And a group of classification methods, such as principal component analysis (PCA) or partial least squares discriminant analysis (PLS-DA), soft independent modeling of class analogy (SIMCA), and so on, have developed and received attention recently.[5–9] However, one will obtain the discriminant difficultly or reach an unreliable model when the spectrum is dominated by a varying background or some nonlinear relations exist with the response.[10]

The redundant information in the spectral data and information irrelevant to the response will worsen the quality of the model and the precision of the prediction. Various signal preprocessing techniques, including variable selection,[7] orthogonal signal correction (OSC),[11] uninformative variable elimination (UVE),[12] frequency-domain processing using wavelet transformation (WT),[13] have been employed to eliminate background and noise to improve the robustness and reliability of the model in both calibration and classification. WT has been proven to be a powerful tool for dimension reduction and noise removal.[14] Vannucci *et al.*[15] have used wavelet-based feature selection technique to classify the NIR and mass spectra. In combination with WT, a lot of algorithms, such as wavelet transformation- uninformative variable elimination (WT-UVE),[16] wavelet transformation- modified uninformative variable elimination (WT-MUVE),[17] wavelet orthogonal signal correction (WOSC),[10] wavelet packet transform for efficient pattern recognition of signals (WPTER),[18] etc., have also been proposed and successfully used for NIR calibration and classification.

Classification and influence matrix analysis (CAIMAN)[19] is a new classification technique based on the influence matrix (also called leverage matrix). Since proposed, it has shown excellent performances with several classification data sets[19] and applications of geographical classification.[20] Recently, it has been developed further by Forina *et al.*[21] to obtain a family of powerful classification and class modeling techniques, by adding distances and leverages to original variables. The method uses a simple mathematical approach to get the leverage matrix, the diagonal elements of which provide information on the influence of each sample within the model. In comparison with other approaches mentioned above, there is no assumption on the multinormal distribution of the data in CAIMAN and the classification model has a good interpretability as the results can be easily interpreted by analyzing the leverage and hyper-leverage values. As CAIMAN still requires the number of class objects to be significantly greater than the number of variables, wavelet-based feature selection will be employed as a reliable dimension reduction technique to overcome this limit.

In the manuscript, we introduced a discriminant classification method named discriminant-CAIMAN (D-CAIMAN) to simultaneously classify the NIR spectral data of samples. A wavelet-based feature selection was first used to compress while retaining information useful for classification, and a new algorithm, WD-CAIMAN, was proposed. NIR spectra data set of 265 *salviae miltiorrhizae radix* samples from nine different geographical origin was used as an example to show the classification performance achieved by the algorithm. For the comparison, methods of k-nearest neighbor (KNN),[22] LDA and quadratic discriminant analysis (QDA) are also used.

2. Theory

2.1. *Influence matrix and leverage*

A commonly used linear model can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{X}$ is the data matrix, $\mathbf{y}$ is a vector of the response. By solving the model using linear least squares, the regression coefficients $\boldsymbol{\beta}$ are

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y}. \tag{2}$$

So the estimated values are

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y} = \mathbf{H}\mathbf{y}, \tag{3}$$

$$\mathbf{H} = \mathbf{X} \cdot (\mathbf{X}^{T}\mathbf{X})^{-1} \cdot \mathbf{X}^{T}. \tag{4}$$

In statistics, the matrix $\mathbf{H}$ is called hat matrix, or influence matrix, or leverage matrix, which describes the influence each observed variable has on response variable. The influence matrix has a number of useful algebraic properties and some practical applications in regression analysis.[23,24] The diagonal elements of $\mathbf{H}$, h_{ii}, are defined as leverages which are key elements of the CAIMAN method.

CAIMAN has been developed with the aim of representing each class space by using the class influence matrix and calculating the leverage value of each object from each class influence matrix. Provided that the size of data matrix $\mathbf{X}$ is $n \times p$, collecting n objects belonging to one of defined G classes and each object has p variables (wavelengths or wavenumbers for NIR spectra). The basic idea is to model each class by the corresponding dispersion matrix estimated by the training objects belonging to the class. For each class $g (g = 1, \ldots, G)$, the leverage of the ith object $\mathbf{x}_i$ is calculated as:

$$h_{ig} = (\mathbf{x}_i - \mathbf{x}_g^*)^T (\mathbf{X}_g^T \mathbf{X}_g)^{-1} (\mathbf{x}_i - \mathbf{x}_g^*), \tag{5}$$

where $\mathbf{x}_g^*$ is the centroid of class g and $\mathbf{X}_g$ is the submatrix of size $n_g \times p$, collecting the n_g centered objects assigned to the gth class. For the objects in training set of the class, the leverage are replaced by the leave-one-out (LOO) estimate, which can be directly derived from the leverage h_{ig}:

$$h_{ig}^* = h_{ig}/(1 - h_{ig}). \tag{6}$$

For the meaning of the leverage, it should be expected that typical characteristic objects of the class have low leverages, while objects far from the class have high leverages.[19] A classification rule based on the minimum leverage is the simplest leverage-based classifier, which exploits the information given by a single leverage calculated independent of the other classes. This classifier is useful in several cases, but it is not able to properly deal with nonlinear class separability, and this limit of the minimum leverage makes it unsuitable to solve classification problems characterized by more complex class structure. Therefore, the CAIMAN approach has been further developed defining a new mathematical concept called hyper-leverage. For each ith object, the hype-leverage hh_{ig} for class g is calculated as:

$$hh_{ig} = (\mathbf{h}_i - \mathbf{h}_g^*)^T (\mathbf{H}_g^T \mathbf{H}_g)^{-1} (\mathbf{h}_i - \mathbf{h}_g^*), \tag{7}$$

where $\mathbf{h}_i = [h_{i_1}, \ldots, h_{ig}, \ldots, h_{iG}]^T$ is the vector of the G leverages of object i, $\mathbf{h}_g^*$ is the centroid of leverages of class g and $\mathbf{H}_g$ is the matrix of class centered leverages, both computed with only the objects in the training set of the gth class.

For the objects belonging to training set of the gth class, the LOO estimation of hyper-leverage is computed as:

$$hh_{ig}^* = hh_{ig}/(1 - hh_{ig}). \tag{8}$$

By projecting the objects in the $\mathbf{H}$-space, the hyper-leverage makes the relationships among the classes more apparent by taking into account all the leverages simultaneously for each object. Therefore, the hyper-leverage is very useful for classifying objects especially when there is a nonlinear class separability in the $\mathbf{X}$-space, which becomes linear in the $\mathbf{H}$-space. By combining information from leverages and hyper-leverages, the leverage score is computed as:

$$w_{ig} = (1 - a)h_{ig} + a\, hh_{ig} \quad 0 \le a \le 1, \tag{9}$$

where a is called trade-off parameter. When the value of a is near 0, it indicates that good class discrimination is directly obtained from the $\mathbf{X}$-space, while a value near 1 results from more complex class structures.

2.2. *D-CAIMAN method*

Discriminant analysis is a supervised classification method which is primarily used to build classification rules for a number of pre-specified subgroups. These rules are later used for allocating new and unknown samples to the most probable subgroup. Another important application of discriminant analysis is to help in interpreting differences between groups of samples.[3] For the purpose of discriminant analysis, one of the CAIMAN methods, called D-CAIMAN, can be used.

In D-CAIMAN, an object is assigned to one of the predefined class by minimizing the leverage score:

$$i \to g \quad \text{if } w_{ig} = \min_{j=1,\ldots,G}\{w_{ij}\}. \tag{10}$$

With the D-CAIMAN classifier, all the objects are always classified to one of the defined classes.

2.3. *Feature selection in wavelet domain*

Data preprocessing, involving data transformation and/or data reduction can dramatically influence the final results of recognition. This is particularly true for spectral data, which contain hundreds or thousands of highly correlated variables, noise and irrelevant information. Removing nonrelevant variability and extracting relevant features from the initial data set is of vital importance.

The main goal of this paper is to present an effective procedure for classification of NIR spectral data. In order to obtain a good estimate of the class dispersion matrix, the number of class objects should be significantly greater than the number of variables, which is a basic condition for all methods based on class covariance matrices,[3] including CAIMAN. However, the NIR spectral data are severely ill-conditioned (with thousands of variables, i.e., wavelengths), so the classification methods cannot be directly applied. This problem can be alleviated by using a compression technique to reduce the dimensionality of the data prior to the variable selection procedures, after that, the selected features are employed as the input of the classifiers. In the present work, a wavelet-based feature selection method is adopted for this purpose.

The wavelet transform is a multiresolution signal processing tool which has been successfully utilized for spectral data analysis in background removing, denoising, feature extraction and compression.[10,14–18,25] With the transformation by wavelet, the information contained in original spectra data can be represented by the wavelet coefficients. Compression is achieved by suppressing the wavelet coefficients that do not hold valuable information. Finally, the original data can be explained by only a small amount of wavelet coefficients. The number of wavelet coefficients to be retained is determined by a threshold value, which results from a preset compression ratio (CR), a key parameter in wavelet compression.[26] A high CR is always expected, but useful information will be lost if too many wavelet coefficients are suppressed. On the contrary, keeping much of the information with a lower CR leads to a low object/variable ratio which will make classification methods to fail. Furthermore, a coefficient that is retained only means it is important to represent the main information of the original signals, it still contains irrelevant information for multivariate calibration. Therefore, a compromising CR combined with a further suppression of the less informative ones from the retained wavelet coefficients for compensation, called wavelet-based feature selection, is an intelligent procedure. In this paper, the forward variable selection technique[27] is adopted for further extraction of the relevant information from the wavelet coefficient. This method starts with no variables and adds one variable at a time to the model, and the inclusion of a variable is based on the $\mathrm{ER}_{\mathrm{LOO}}$ (error rate leave-one-out) value, i.e., the variables will be entered into the model if $\mathrm{ER}_{\mathrm{LOO}}$ is minimized.

2.4. *WD-CAIMAN method*

A wavelet-based feature selection was first used for data compressing while retaining information useful for classification, and a new algorithm, WD-CAIMAN, was proposed. The detailed procedure of the WD-CAIMAN method can be described as follows:

(1) WT data compression. For a given wavelet function and a maximum decomposition level J, the wavelet decomposition of each spectrum is performed to calculate wavelet coefficients $c\mathbf{W} = [c\mathbf{A}_J, c\mathbf{D}_J, c\mathbf{D}_{J-1}, \ldots, c\mathbf{D}_1]$. Then, small coefficients in $c\mathbf{W}$ will be suppressed by thresholding method.[28] The number of wavelet coefficients to be stored is going to be determined by a threshold value, which is calculated

by a preset CR. CR is generally defined by N/N', where N is the data point number of the original signal and N' is the number of the coefficients to be retained. For a given CR, different wavelet functions are compared with maximum number of decomposition level in order to achieve optimum compression.

(2) Variable selection in wavelet domain. The forward variable selection technique is adopted for further extraction of the relevant information from the wavelet coefficients. Compared to forward selection, other variable selection strategies including backward selection, forward-backward selection, subset selection and block-addition and block-deletion selection are normally more complex, and they havenot showed to obtain better results.[29] The forward selection method adds coefficients to the model one at a time. The inclusion of a coefficient is based on a F_{enter} (*F to enter value*).[30] The coefficients will be entered into the classification model if their respective F value is larger than the specified F_{enter}. The F values can be calculated and tested based on the Wilk's lambda statistic,[27] which is a measure of the quality of the separation among the classes. LOO validation procedure will give the final classification model, and the coefficients will be kept if ER_{LOO} is minimized.

(3) Build D-CAIMAN model by using the suppressed wavelet coefficients. First, the leverages and hyper-leverages are calculated for each class g by training samples using Eqs. (5)–(8). Then, leverage score is obtained by Eq. (9) with a leave-more-out (LMO) cross-validation for the optimization of parameter a (between 0 and 1), which gives information about the trade-off between the leverage and hyper-leverage role in the final classification rule. The procedure is performed as following: The samples are split into different cross-validation groups; Once at a time, each validation group is removed from the training set and the percentage of wrong assignments in the cross-validation groups (ER_{CV}) is calculated; the parameter a is chosen on the basis of the results by minimizing the error rate. Finally, the D-CAIMAN classifier is built with the retained wavelet coefficients and parameter a.

(4) Discriminate unknown samples by WD-CAIMAN model. Obtain the suppressed wavelet coefficients of the prediction data set by steps (1) and (2) using the same parameters, and perform prediction by using the D-CAIMAN classifier built in step (3). WD-CAIMAN makes all the objects always be classified to one of the defined classes.

3. Experimental Methods

Herbal medicines, with their effective pharmacological activities and low toxicity have been playing an important role in clinical therapy in China. The quality and efficacy of herbal medicines vary with geographical areas of production. It is urgently needed to develop a reliable method for the geographical origin identification which is also of great importance to the quality control of herbal medicines. Compared with some traditionally used identification methods such as chromatography, NIR spectroscopy technique is simple, rapid and no sample preparation is needed. It has proved to be feasible and effective for qualitative and quantitative analysis of herbal medicines.[31] In this work, NIR spectroscopy technique coupled with chemometrics tools was employed for the fast discrimination of *salviae miltiorrhizae radix* according to geographical origins.

3.1. *Sample preparation*

A total of 265 samples of *salviae miltiorrhizae radix* from different geographical origins in China were collected. The details of samples are shown in Table 1. The samples were dried at 60°C for 10 min, then cut and ground into powder. The final powder samples were prepared by passing the ground powder through an 80-mesh sieve.

Table 1. Class, origin and number of the 265 *salviae miltiorrhizae radix* samples.

Origin	Number	Class
Bozhou, Anhui	30	C1
Zhongjiang, Sichuan	29	C2
Pingyi, Shandong	29	C3
Yuncheng, Shanxi	30	C4
Yancheng, Jiangsu	30	C5
Anguo, Hebei	28	C6
Xingtang, Hebei	30	C7
Lushi, Henan	29	C8
Shangluo, Shanxi	30	C9

3.2. *Spectra collection and data processing*

NIR spectra of the samples were collected in the reflectance mode at 8 cm^{-1} intervals over the spectral region $4000 \sim 10,000$ cm^{-1} with an Antaris FT-NIR System (Thermo scientific, USA) equipped with an optical fiber. Each sample spectrum was obtained by averaging 64 scans, and all the spectra were recorded as the logarithm of the reciprocal, log $(1/R)$. The mean of the three spectra which were collected from the same sample was used in the following analysis steps.

All spectra were randomly divided into a calibration set and a prediction set: 75% of the samples were the calibration set while the remaining samples were utilized for prediction set. Each sample was only used once in one data set. A classification model was developed in the calibration set, and the developed model was evaluated by the prediction set. MATLAB 7.8 (The Math Works Inc., Natick, MA) was used for all data analysis. The free CAIMAN toolbox[32] was applied with MATLAB to derive the D-CAIMAN models.

3.3. *Classification validating*

To evaluate the classification performance of classifier, three parameters, ER, Sn and Sp, are used.[33] Error Rate (ER) is the percentage of misclassified objects, $\mathrm{ER}\% = 1 - \mathrm{NER}\%$. Non-Error Rate (NER) is defined as the percentage of objects correctly classified. This parameter is commonly called sensitivity (Sn), i.e., probability (percentage) of predicting "yes" given true state is "yes", which reflects the ability of a classifier to correctly capture all the objects of the classes. Specificity (Sp) is a parameter which characterizes the ability of the gth class to reject objects of the other classes after the application of a classifier, which indicates the probability (percentage) of predicting "no" given true state is "no".

4. Results and Discussion

The mean spectra of each class for the original data were shown in Fig. 1. This work aims at exploring the possibility of using NIR spectra to assign unknown samples to the correct class. As can be seen, the NIR spectra of different classes are very similar: two water absorption bands around 5155 and 7000 cm^{-1}, other intensive bands in the spectrum are contributed by the vibration of the second overtone of the carbonyl group (5352 cm^{-1}), the C–H stretch and C–H deformation vibration (7212 cm^{-1}). Owing to the complexity of the spectra, the discrimination of nine classes of samples on the basis of NIR spectra cannot be obtained straightforwardly, which is apparently shown in the PC score plots presented in Fig. 2. The C2 and C8 samples are isolated in the PCs space and can be well separated along PC2. However, even if the NIR spectra seem to contain some information useful for distinguishing the different classes, there remains a considerable overlap among the other seven classes.

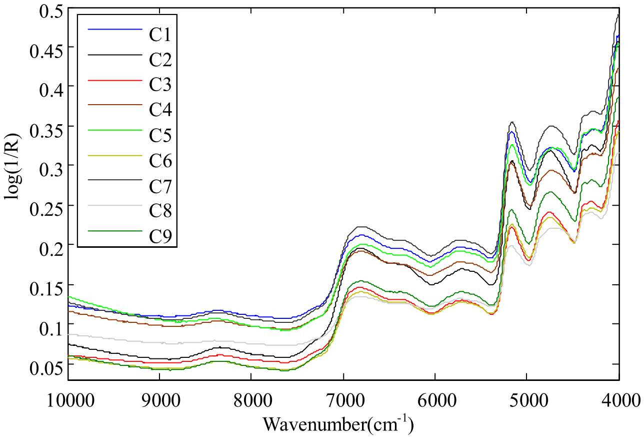


Fig. 1. The mean spectra of each class for the original data.

4.1. *Wavelet compression*

Determination of right threshold value is a key step for compression procedures. In this paper, only hard

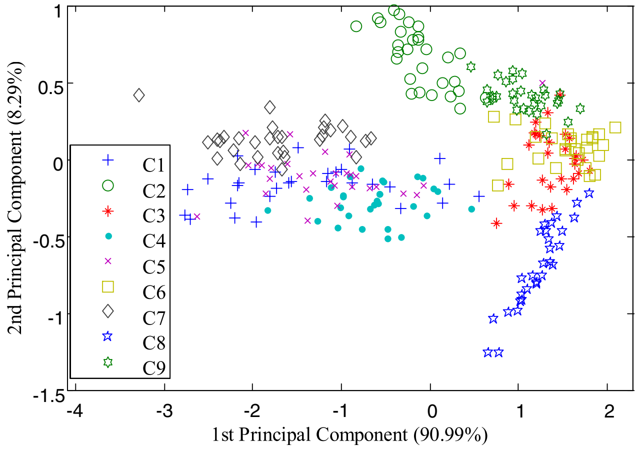


Fig. 2. Plot of first two principal components (PCs) obtained by PCA of NIR spectra.

Table 2. RMS between original NIR spectrum and reconstructed spectra by WT with different wavelets and a maximum decomposition level (CR = 16).

Wavelet	Max level	RMS ($\times 10^{-4}$)	Wavelet	Max level	RMS ($\times 10^{-4}$)	Wavelet	Max level	RMS ($\times 10^{-4}$)
Db1	10	18	Db8	6	6.99	Sym8	6	7.15
Db2	9	7.53	Db9	6	8.22	Sym9	6	7.51
Db3	8	6.19	Db10	6	8.81	Sym10	6	7.79
Db4	7	5.97	Sym4	7	6.10	Coif1	8	7.55
Db5	7	6.08	Sym5	7	5.62	Coif2	9	7.11
Db6	7	7.12	Sym6	7	6.85	Coif3	6	8.00
Db7	6	6.55	Sym7	6	6.83	Coif4	6	9.44

thresholding was used because the aim of compression was to remove the small coefficients. For this purpose, three thresholds, CR = 8, 16 and 20, with which 195, 97 and 78 coefficients were retained respectively, were set to investigate the effect at different CR. The root mean square (RMS) errors between the original spectrum and reconstructed spectra were $3.89 \times 10^{-4}, 7.53 \times 10^{-4}$ and 9.58×10^{-4}, which indicated that there was almost no difference between the original and reconstructed spectra. As stated in Sec. 2.3, a compromising CR = 16 was the best choice for the next feature selection procedure.

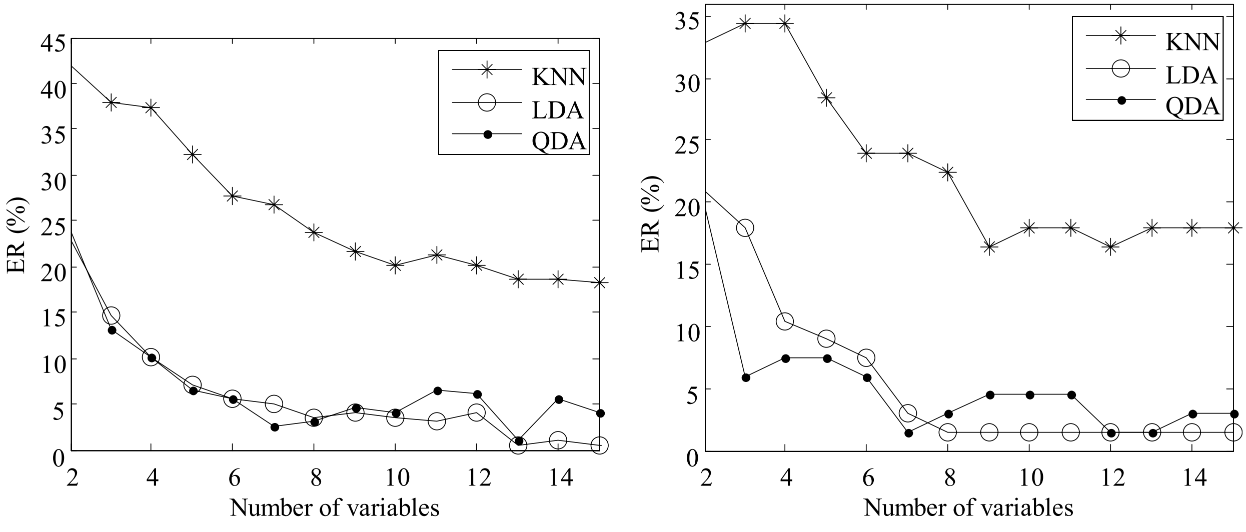
Different wavelet functions and decomposition level will result in different effects of compression. In order to obtain optimal compression, the RMS errors of the original measured spectrum and reconstructed signal were investigated with different wavelets and maximum number of decomposition levels for each wavelet. Table 2 summarizes the RMS by 21 different wavelets (Daubechies 1–10, Symlet 4–10 and Coieflet 1–4) with maximum decomposition level for each wavelet and CR = 16. It can be seen that the "sym5" and "db4" give a slightly better results for compression, yet the difference among most of these wavelets are not significant. Finally, "db4" wavelet was adopted in the following studies.

4.2. *Classification*

After compression by WT, number of variables was considerably reduced from 1557 to 97. However, the sample/variable ratios in nine classes are all still smaller than 0.3, which does not meet the request of sample/variable ratio for classification algorithms. The forward selection has been applied to the compressed wavelet coefficients in order to increase the sample/variable ratio. As explained before, each variable was sorted according to the Wilk's lambda value and selected to model a variable once until the sample/variable ratio larger than 1. KNN, LDA, QDA and WD-CAIMAN were used on the data, in order to compare the classification performances.

As a sample/variable ratio greater than 2 or 3 is usually suggested, the number of retained variables should be less than 15 for the NIR data used in this paper. Classification error curves for the first 15 retained variables for the cross-validation of three classifiers, KNN, LDA and QDA, are depicted in Fig. 3(a). A 10-fold cross-validation for samples in calibration set was used and the best classifiers were obtained for both three methods. As shown in Fig. 3(a), ER for the KNN classifier is much larger than the others, reaching a minimum error of 18.18% at 15 variables. LDA and QDA almost have the same classification performances, both requiring 13 variables, with a minimum error of 0.51% and 1.01%, respectively. Figure 3(b) shows the external prediction classification errors from all three classifiers by using prediction samples. Determined by cross-validation, KNN, LDA and QDA classifiers reach optimal classification errors of 17.91%, 1.49% and 1.49%, respectively.

In WD-CAIMAN procedures, the LMO technique was utilized to optimize a value. Specifically, for each a value (between 0 and 1, with step 0.1), 300 iterations excluding 20% of objects for each class were performed. The best a value and number of variables are obtained (see Table 3) when $\mathrm{ER}_{\mathrm{LMO}}$ reached to a minimum value 3.1%. In order to compare the results with other classification methods, the LOO validation technique was employed with $a = 0.5$ and 13 variables obtained by LMO technique. Moreover, the models have been tested also with the external prediction set of samples (see Table 3). As it can be seen, the results obtained by WD-CAIMAN are satisfactory with $\mathrm{ER}_{\mathrm{LOO}}$ and on the



(a) (b)

Fig. 3. The curves of total classification error versus retained variables for KNN, LDA and QDA methods: (a) Cross-validation and (b) prediction.

external prediction set (ER_{EXT}) both equal to 0%. By considering all the performance parameters, i.e., ER_{LOO}, ER_{cv}, ER_{EXT}, Sn and Sp, the best approach seems to be WD-CAIMAN on the selected variables, and KNN on the selected variables gives the worst results.

4.3. *Model interpretation with leverage (or hyper-leverage) plot*

The classification model obtained by WD-CAIMAN method has a good interpretability as the results can be easily interpreted by the leverage plot or hyper-leverage plot. The leverage plot is basically a scatter plot obtained by projecting the samples in the space defined by the leverages referring to two chosen classes.[19,20] Each axis represents a class and farer the objects are from the axis origin, greater their distance is from the corresponding class. Investigating the distance between each two classes depicted in the PCA plot (see Fig. 2), it can be seen that the classes in group G1 = {C1, C4, C5, C7} and group G2 = {C3, C6, C9} are hard to identify from each other. Therefore, the leverage plot between those classes will be plotted to better understand how the WD-CAIMAN model works.

Define L_{ij} as the leverages (or hyper-leverages) of Ci samples from C$j(i, j = 1, 2, \ldots, 9)$. The C$i$ samples can be successfully identified from C$j(j \neq i)$ if the following conditions are satisfied:

Cd.1: Small values of L_{ii} $(i = j)$;
Cd.2: Large values of L_{ij} $(i \neq j)$;
Cd.3: No overlap between ranges of L_{ii} and each $L_{ij}(i \neq j)$.

Table 3. Classification results obtained by WD-CAIMAN, KNN, LDA and QDA with retained wavelet coefficients.

		Calibration			Prediction				
Methods	Number of variables	ER^a(%)	Sn	Sp	ER^b(%)	Sn	Sp	ER^c_{LMO}(%)	a^c
WD-CAIMAN	13	0	100	100	0	100	100	3.1	0.5
KNN	15	18.18	54.55	94.89	17.91	75	94.92	—	—
LDA	13	0.51	100	99.43	1.49	100	98.31	—	—
QDA	13	1.01	95.45	99.43	1.49	100	98.31	—	—

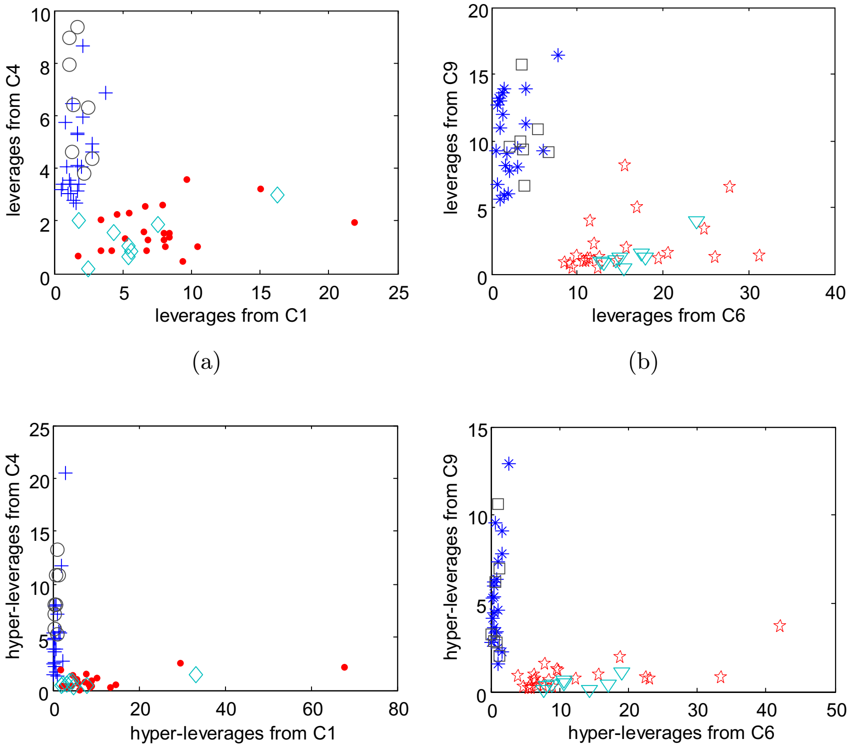
[a] For WD-CAIMAN, leave-one-out cross validation was used (ER_{LOO}), 10-fold cross validation (ER_{cv}) was for KNN, LDA and QDA methods.
[b] ER was calculated by external validation samples in prediction set (ER_{EXT}).
[c] LMO validation was only used in WD-CAIMAN method, by which parameter a was optimized.

Table 4. The ranges of each class leverage L_{ij} in group G1 and G2.

Group	Leverages	C1	C4	C5	C7	Cd.1[b]	Cd.2[b]	Cd.3[b]
G1	C1[a]	0.55–3.78	1.66–21.82	75.92–170.27	23.51–80.01	√		
	C4[a]	2.70–9.34	0.18–3.58	53.54–93.68	18.80–41.63	√		
	C5[a]	31.76–66.50	38.23–65.93	0.19–7.93	35.95–73.23	√	√	√
	C7[a]	11.26–31.08	17.96–47.33	14.36–85.95	0.61–5.99	√	√	√
G2	Leverages	C3	C6	C9				
	C3[a]	0.26–5.34	9.22–28.07	8.52–31.25		√	√	√
	C6[a]	37.31–151.33	0.58–7.74	8.49–31.35		√		√
	C9[a]	11.54–39.20	5.66–16.44	0.44–8.17		√		

[a] Raw i represents the ranges of each L_{ij} with j changes.
[b] Three conditions used for discrimination.

Table 4 lists the ranges of class leverages L_{ij} in two groups. It is clearly visible that samples have the smallest leverages L_{ij} $(i = j)$ from the membership class, i.e., Cd.1 is satisfied. However, overlapping happens between class leverages ranges, such as L_{11} and L_{14}, L_{96} and L_{99}, which will be due to class overlaps in classification. Therefore, class C5 and C7 can be successfully discriminated from G1, and class C3 can be separated from G2. For classes with overlapping range of leverages, the leverage and hyper-leverage plots are shown in Fig. 4. As the value of a is set to 0.5, the leverages and hyper-leverages are both used in the WD-CAIMAN classification model. Samples falling in the left top and right bottom corner are very close to the class represented by the horizontal axis and vertical axis, respectively, and far from the other one. As shown in Fig. 4, class C1 and C6 can be discriminated, respectively, from



(a)

(b)

(c)

(d)

Fig. 4. The leverages and hyper-leverages plot: (a) leverages from C1 and C4; (b) leverages from C6 and C9; (c) hyper-leverages from C1 and C4; (d) hyper-leverages from C6 and C9. The (+, ○), (•, ◇), (∗, □), (☆, ▽) denote (calibration, prediction) samples of C1, C4, C6, C9, respectively.

C4 and C9 by a combination of the leverages and hyper-leverages. This letsus achieve the best result by WD-CAIMAN classification model.

5. Conclusion

A new algorithm, WD-CAIMAN, was proposed for NIR spectra classification based on the wavelet-based feature selection and D-CAIMAN discrimination methods. NIR spectra data set of 265 *salviae miltiorrhizae radix* samples from different geographical origin (nine classes) was used as an example to validate the algorithm. The proposed wavelet-based feature selection procedure was useful to compress the spectra data and make the classification model brief and clear. With a comparison with KNN, LDA and QDA methods, WD-CAIMAN shows the best performance, achieving the total classification rate of 100% in both cross-validation and prediction set. It is worth noting that the WD-CAIMAN classifier also shows improved sensitivity and selectivity in these classifications. Moreover, the WD-CAIMAN classification model has a good interpretability as the results can be easily interpreted by analyzing the leverage and hyper-leverage values and no assumption on the multi-normal distribution is needed.

Acknowledgments

Financial support from China Postdoctoral Science Foundation Special Funded Project (2013T60604) and Zhejiang Provincial Public Welfare Application Project of China (2012C21102) are gratefully acknowledged.

References

1. D. A. Burns, E. W. Ciurczak, *Handbook of Near-Infrared Analysis*, 3rd Edition, CRC Press, New York (NY) (2008).
2. T. Naes, T. Isaksson, T. Fern, T. Davies, *A User Friendly Guide to Multivariate Calibration and Classification*, NIR Publications, Chichester (2002).
3. G. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, New York (NY) (1992).
4. E. K. Kemsley, "Discriminant analysis of high-dimensional data: A comparison of principal components analysis and partial least squares data reduction methods," *Chemometr. Intell. Lab. Syst.* **33**, 47–61 (1996).
5. S. Wold, "Pattern recognition by means of disjoint principal components models," *Pattern Recogn.* **8**, 127–139 (1976).
6. I. E. Frank, "DASCO — a new classification method," *Chemometr. Intell. Lab. Syst.* **4**, 215–222 (1988).
7. D. Ballabio, T. Skov, R. Leardi, R. Bro, "Classification of GC-MS measurements of wines by combining data dimension reduction and variable selection techniques," *J. Chemometr.* **22**, 457–463 (2008).
8. M. Forina, P. Oliveri, H. Jäger, H. Römisch, J. Smeyers-Verbeke, "Class modeling techniques in the control of the geographical origin of wines," *Chemometr. Intell. Lab. Syst.* **99**, 127–137 (2009).
9. C. Tan, X. Qin, M. Li, "Comparison of chemometric methods for brand classification of cigarettes by near-infrared spectroscopy," *Vibr. Spectrosc.* **51**, 276–282 (2009).
10. W. D. Ni, S. D. Brown, R. L. Man, "Wavelet orthogonal signal correction-based discriminant analysis," *Anal. Chem.* **81**, 8962–8967 (2009).
11. S. Wold, H. Antti, F. Lindgren, J. Ohman, "Orthogonal signal correction of near-infrared spectra," *Chemometr. Intell. Lab. Syst.* **44**, 175–185 (1998).
12. V. Centner, D. L. Massart, O. E. De Noord, S. De Jong, B. M. Vandeginste, C. Sterna, "Elimination of uninformative variables for multivariate calibration," *Anal. Chem.* **68**, 3851–3858 (1996).
13. X. G. Shao, W. S. Cai, "Wavelet analysis in analytical chemistry," *Rev. Anal. Chem.* **17**, 235–285 (1998).
14. X. G. Shao, A. K. M. Leung F. T. Chau, "Wavelet: A new trend in chemistry," *Acc. Chem. Res.* **36**, 276–283 (2003).
15. M. Vannucci, N. J. Sha, P. J. Brown, "NIR and mass spectra classification: Bayesian methods for wavelet-based feature selection," *Chemometr. Intell. Lab. Syst.* **77**, 139–148 (2005).
16. X. G. Shao, F. Wang, D. Chen, Q. D. Su, "A method for near-infrared spectral calibration of complex plant samples with wavelet transform and elimination of uninformative variables," *Anal. Bioanal. Chem.* **378**, 1382–1387 (2004).
17. X. J. Chen, D. Wu, Y. He, S. Liu, "Detecting the quality of glycerol monolaurate: A method for using Fourier transform infrared spectroscopy with wavelet transform and modified," *Anal. Chim. Acta* **638**, 16–22 (2009).
18. G. Foca, M. Cocchi, M. L. Vigni, R. Caramanico, M. Corbellini, A. Ulrici, "Different feature selection strategies in the wavelet domain applied to NIR-based quality classification models of bread wheat

flours," *Chemometr. Intell. Lab. Syst.* **99**, 91–100 (2009).

19. R. Todeschini, D. Ballabio, V. Consonni, A. Mauri, M. Pavan, "CAIMAN (classification and influence matrix analysis): A new approach to the classification based on leverage-scaled functions," *Chemometr. Intell. Lab. Syst.* **87**, 3–17 (2007).
20. D. Ballabio, A. Mauri, R. Todeschini, S. Buratti, "Geographical classification of wine and olive oil by means of classification and influence matrix analysis (CAIMAN)," *Anal. Chim. Acta* **570**, 249–258 (2006).
21. M. Forina, M. Casale, P. Oliveri, S. Lanteri, "CAIMAN brothers: A family of powerful classification and class modeling techniques," *Chemometr. Intell. Lab. Syst.* **96**, 239–245 (2009).
22. B. R. Kowalski, C. F. Bender, "K-Nearest neighbor classification rule (pattern recognition) applied to nuclear magnetic resonance spectral interpretation," *Anal. Chem.* **44**, 1405–1411 (1972).
23. D. C. Hoaglin, R. E. Welsch, "The hat matrix in regression and ANOVA," *Am. Stat.* **32**, 17–22 (1978).
24. N. R. Draper, H. Smith, *Applied Regression Analysis*, Wiley-Interscience, New York (NY) (1998).
25. M. J. C. Pontes, J. Cortez, R. K. H. Galvao, C. Pasquini, M. C. U. Araújo, R. M. Coelho, M. K. Chiba, M. F. de Abreu, B. E. Madari, "Classification of Brazilian soils by using LIBS and variable selection in the wavelet domain," *Anal. Chim. Acta* **642**, 12–18 (2009).
26. F. T. Chau, Y. Z. Liang, J. B. Gao, X. G. Shao, *Chemometrics: From Basics to Wavelet Transform*, Wiley-Interscience, New York (NY) (2004).
27. R. J. Jennrich, K. Einslein, A. Ralston, Eds., Statistical Methods for Digital Computers, Wiley, New York (NY) (1977).
28. M. Jansen, *Noise Reduction by Wavelet Thresholding*, Springer-Verlag, New York (NY) (2001).
29. L. J. Herrera, G. Rubio, H. Pomares, B. Paechter, A. Guillen, I. Rojas, *Proc. Conf. Artificial Neural Networks*, C. Alippi, Ed., Lecture Notes in Computer Science, Springer-Verlag, Heidelberg (2009).
30. G. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, New York (NY) (1992).
31. Y. Jiang, B. David, P. F. Tu, Y. Barbin, "Recent analytical approaches in quality control of traditional Chinese medicines- A review," *Anal. Chim. Acta* **657**, 9–18 (2010).
32. CAIMAN Toolbox Matlab code available at http://michem.disat.unimib.it/chm/.
33. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, Springer, New York (NY) (2001).