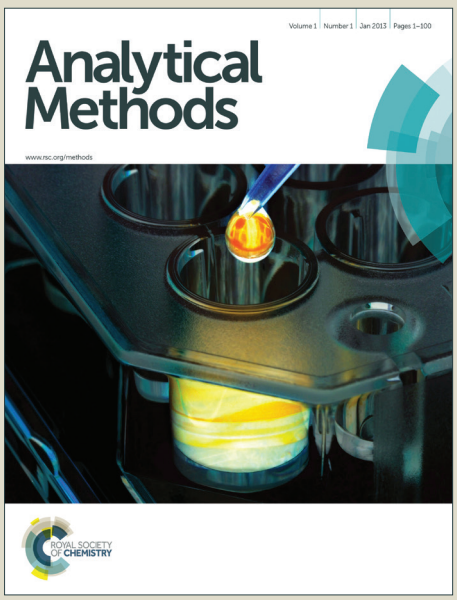
Analytical Methods

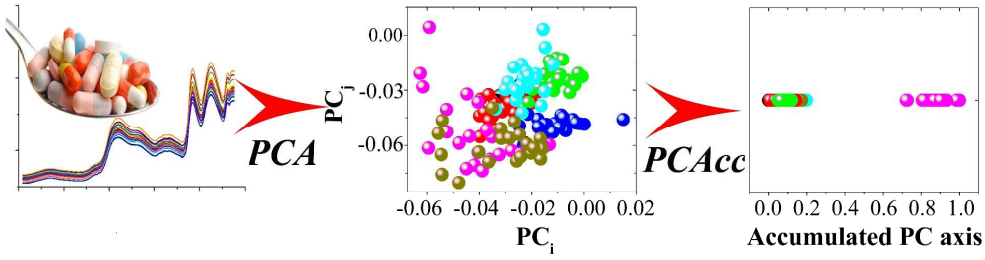
Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/methods

Graphical abstract



Near-infrared (NIR) spectroscopy combined with principal component accumulation (PCAcc) method was used to identify 12 classes of different Chinese patent medicines.

Discrimination of Chinese patent medicines using near-infrared spectroscopy and principal component accumulation method

Ruifeng Shan[a], Zhiyi Mao[a], Lihui Yin[b], Wensheng Cai[a] and Xueguang Shao[a*]

[a]Collaborative Innovation Center of Chemical Science and Engineering (Tianjin), State Key Laboratory of Medicinal Chemical Biology, Research Center for Analytical Sciences, College of Chemistry, Nankai University, Tianjin, 300071, P. R. China

[b]National Institutes for Food and Drug Control, Beijing, 100050, P. R. China

Corresponding address:

College of Chemistry,

Nankai University,

Tianjin, 300071, P. R. China

Tel: +86-22-23503430

Fax: +86-22-23502458

E-mail: xshao@nankai.edu.cn

[*] Corresponding author.

Abstract

Discrimination of pharmaceutical products has been an important task in pharmaceutical industry and pharmaceutical safety. In this study, principal component accumulation (PCAcc) method was investigated for discrimination of Chinese patent medicines. In PCAcc method, an accumulation strategy is utilized to combine the classification information contained in multiple PC subspaces by using a rotation, a projection and a summation operation. To improve the performance of classification, continuous wavelet transform (CWT) is applied as the pretreatment method to eliminate the background. The results show that, among the 12 classes of Chinese patent medicines, 8 classes are correctly classified, and a total of ten samples are misclassified for the other four classes. Compared with the results obtained by principal component analysis (PCA), radial basis function artificial neural network (RBF-ANN) and partial least squares discriminant analysis (PLSDA), PCAcc produces the best classification.

Introduction

Near-infrared (NIR) spectroscopy is a fast and nondestructive analytical technique and has been widely used in food industry, agriculture, petroleum industry, and etc.[1,2] In the last decades, the technique has attracted considerable attention in pharmaceutical industry for quantitative analysis, qualitative analysis and on-line control of pharmaceutical products.[3-5] Due to its flexibility in measurement, NIR spectroscopy is suitable for analysis of samples in different pharmaceutical forms. The technique has been extensively studied for quantifying active principal ingredients (API),[6,7] excipients,[8-10] and water content[11-13] in pharmaceutical products. Moreover, because NIR spectroscopy has advantages of rapid and nondestructive analysis, it has been used to monitor the production process of pharmaceutical products, e.g., assessing tableting process,[14-16] monitoring blend uniformity of solid dosage forms[17] or API concentration in powder mixing process.[18] On the other hand, NIR spectroscopy has a characteristic that could capture both chemical and physical information of the samples. The parameters of pharmaceutical products such as hardness, particle size, compaction force and dissolution rate can be determined by the technique. NIR spectroscopy was also used to provide the information of polymorphic form,[19-21] which affects dissolution property of the pharmaceutical products.

Discrimination of geographic origins or manufacturer and identification of counterfeit drugs have been an important task in pharmaceutical industry. However, in some cases, e.g., the same pharmaceutical product from different manufacturers,

there is no significant difference. Therefore, efficient methods are needed to classify the similar samples by exploring the tiny difference between the products. Pattern recognition techniques combined with NIR spectroscopy have attracted considerable attention for variety discrimination. Principal component analysis (PCA) is one of the most popular and straightforward pattern recognition methods, and has been used to taxonomic discrimination, quality assessment and discrimination of geographic origins of pharmaceutical products, and etc.[22] For example, PCA was employed to discriminate three types of Indigowoad Root samples from different origins[23] and to identify counterfeit drugs.[24] In the method, the relation between the samples can be directly observed by the plot of principal components (PCs). In our recent work, classification of azithromycin tablets from four manufacturers was studied by PCA and the effect of morphology was examined by preparing the samples in different forms.[25] The results show that both the samples from different manufacturers and the samples in different forms can be satisfactorily classified with the help of chemometric methods. Moreover, least-squares support vector machine (LS-SVM) was adopted for discrimination of Rhizoma Corydalis and mint tea samples from different sources. The results demonstrated that the method can find the non-linear relation between the spectra and predicted properties.[26,27] Furthermore, K-means method has been used to discriminate tablets from different manufacturers.[28]

The aim of this work is to establish an approach for rapid identification of Chinese patent medicines. NIR spectroscopy was used as a tool for fast and destructive analytical technique to obtain the information of the samples, and

chemometric methods, including continuous wavelet transform (CWT) and principal component accumulation (PCAcc), were employed to explore the small difference between the spectra. 12 classes of different Chinese patent medicines or the same medicine from different manufacturers were studied to demonstrate the performance of the method in discrimination of Chinese patent medicines.

Experiment and data description

NIR spectral dataset of Chinese patent medicine is supplied by National Institutes for Food and Drug Control. The dataset includes NIR spectra of five Chinese patent medicines produced by different manufacturers. Table 1 summarized the information of the samples. The samples of one medicine from one manufacturer were taken as a class. The capital letters A, B, C, D, and E were used to denote the five medicines and a number following the letter was used to code the manufacturer.

Table 1

The spectra are divided into calibration and prediction set by Kennard-Stone (KS) method.[29] In order to use the same number of spectra in calibration set for the 12 classes of medicines, 22 spectra of each class (a total of $12 \times 22 = 264$ spectra) were used and the remaining spectra as listed in Table 1 were taken as prediction set.

All the spectra were recorded on an MPA FT-NIR spectrometer (Bruker, Germany) in the wavenumber range 3999.7-11995.3 cm^{-1} with the digitization interval 3.857 cm^{-1}. In the calculations, the variables from 4246.6 to 8913.7 cm^{-1} (1211 data points) were used. Fig. 1(a) displays the measured spectra of the samples. It can be

seen that most of the spectra are similar and highly overlapped. The reason is that Chinese patent medicines are mixture of several herbs in composite formulae. Thus, there is no significant difference between the medicines. On the other hand, the chemical constituents in component herbs may vary with harvest season, geographic origin, drying processes and other factors. This may cause the difference between the samples of the same medicine from different manufacturers.

Figure 1

Generally, signal preprocessing methods such as multiplicative scattering correction (MSC),[30] standard normal variate (SNV)[31] and derivative are used for correcting the scattering effect and background removal. In our previous works, CWT has been proved to be an efficient tool for removing the variant background and noise.[32,33] Therefore, CWT is applied as the pretreatment method to eliminate the background in this work. In the calculation of CWT, Haar wavelet with a scale parameter 20 was used. Fig. 1(b) shows the preprocessed spectra. It can be seen that the variant background is removed compared with the spectra in Fig. 1(a). However, the spectra are still overlapped. Therefore, it is impossible to distinguish the 12 classes of the medicines directly by the spectra, although there are differences between these spectra.

Calculations

PCA has been the basic method for classification or discrimination analysis. In PCA, the information contained in first two or three PCs are generally used for

inspection of classification. However, high-order PCs may contain the classification information. In order to use the information sufficiently, PCAcc was proposed in our previous works.[34,35] The essential of PCAcc method is an accumulation strategy to accumulate the classification information contained in multiple PC subspaces. Therefore, the difference between the spectra of the samples in any PC subspaces is used for the classification. A rotation, projection and summation operations are included in the calculations. For building a PCAcc classification model, PCA is applied on the calibration set. In order to explore the information contained in different PC subspaces, a large number of PCs can be used. By using the information in PC subspaces, a decision tree can be obtained, in which each node has two branches. One branch contains the samples of one class and the other one contains the samples of remaining classes. For each decision node, the class with the largest difference from the others is separated out. The process is repeated until only one class remains. The classifier in each node is built with the accumulation, which includes the following operations: (1) finding the axis maximizing the distance between one class and the other classes using Fisher criterion[36] in each PC subspace; (2) rotating all the PC subspaces to the same direction; and (3) accumulating the information of the effective subspaces. The effective subspace is defined as the one producing a "minimal increase"[34] to the classification. In the end, an accumulation sequence (of the PC subspaces) and a threshold to produce the best classification can be obtained as the classifier of the node in the decision tree.

For predicting an unknown sample, a decision can be obtained by testing the

sample along the decision tree. In each node, the spectrum of the sample is projected into the PC subspaces, rotated, and then an accumulation is performed according to the sequence. The classification or discrimination will end when the sample is classified to the end node, i.e., the node containing only one class, using the classifier and the threshold.

Detail procedures of PCAcc method can be found in our previous works.[34,35] In this paper, resolution is still employed as a quantitative measure of the difference between classes, which is defined by[34]

$$R_s = \frac{|m_A - m_B|}{s_A + s_B} \quad (1)$$

where R_s is the resolution between class A and the other classes (denoted as B), m_A, m_B, s_A and s_B are the mean values and standard deviations of the two classes, respectively.

Results and discussion

Discrimination using PCA

PCA is the most commonly used unsupervised pattern recognition method. In this work, PCA is employed to investigate the classification of the medicines. The result shows that the first four PCs explain more than 98% of the variance. Therefore, most information of the spectra is included in the first four PC subspaces. Fig. 2(a) and (b) shows the distribution of the calibration samples in PC1-PC2 and PC3-PC4 subspace, respectively. It can be seen that, in PC1-PC2 subspace, 5 classes of the medicines can be separated, including E1, E2, C1, A3 and A4, and in Fig. 2(b), A3 is

almost separable from the others in PC3-PC4 subspace. For the samples of the other classes, however, it is too difficult to separate them by the four PCs. This result evidently indicates that the difference between the samples may be contained in the high order PCs although explaining very small variance in the spectral data. Furthermore, even for the separable classes, it is obvious that the in-class variance is much larger than the between-class variance. The result clearly demonstrates the difficulty in classification of Chinese patent medicines by using PCA.

Figure 2

Building PCAcc model

In order to use the comprehensive information contained in the PCs to improve the classification, PCAcc model was studied for discrimination of the medicines using the spectra in the calibration set. In the method, all the possible information for the classification contained in multiple PCs is used. In order to use more information, 12 PCs were used in the calculations. Therefore, a total of 66 PC subspaces are included in the accumulation.

To demonstrate the effect of the accumulation, the variation of R_s in the accumulation process for the last node to separate the class D1 and D2 is shown in Fig. 3(a). Clearly, 21 PC subspaces that increase the separation of the two classes are accepted and the R_s value increases from 0.97 (the best PC subspace) to 1.60 (the accumulated value). The result means that 21 of the 66 PC subspaces contain the effective information for the separation and acceptable classification of the two classes is obtained. Statistically, when the value of R_s is above 1.5, it can be known as

a complete separation.

Figure 3

Fig. 3(b) shows the R_s values of the accepted PC subspaces in the accumulation. Clearly, even for the accepted subspaces, the R_s values vary significantly. The largest value can be as high as 0.97 and the smallest one can be as low as 0.12. All the values are lower than 1.0 and more than half of the values are lower than 0.5. This indicates that the discriminating information in an individual PC subspace is limited. Therefore, the accumulation is necessary. Moreover, a sharp increase can be seen in Fig. 3(a) when the PC subspace No. 9, 10, 13 and 14 was accumulated. This indicates that the PC subspaces with a higher R_s value may have a significant contribution to the accumulation. However, there are also cases that the accumulation of the PC subspaces with higher R_s value does not produce significant contribution to the separation, e.g., the PC subspaces No. 16, 18, and 20 in the figure. This maybe accounted for by the fact that the information contained in these PCs is similar with that in the previously accepted PCs.

Figure 4

Figs. 4(a) - (k) shows the discrimination sequence of 12 classes of medicines. The balls in the bottom line display the situation of the calibration samples. The long vertical line in the center denotes the threshold of the classification, and two short vertical lines denote the mean values of the two classes. The position of the long line, i.e., the threshold, is determined by the two short lines, locating at the middle of the two lines. Moreover, the accumulated R_s values are labeled in the figure.

Fig. 4 shows the sequence of the separation in the training process. Clearly, C1 is the first class to be separated. This is because the R_s value between C1 and the other 11 classes is the largest one. After removing the samples of C1, B2 is selected as the second class to be separated with the reason that the R_s value between B2 and the other 10 classes is the largest one. The remaining classes are separated in an order of B1, A2, A4, A3, E2, E1, A5, A1, D1 and D2. The sequence forms a decision tree with 11 nodes. From the R_s values labeled in the figure, it can be known that all the values are bigger than 1.5, indicating a good classification. With detail examination of the figures it can be found that all the samples are correctly discriminated except for one sample in class D1 and three samples in class D2, as shown in Fig. 4 (g) and (k), respectively.

Discrimination of the prediction samples

In order to validate the efficiency of the PCAcc model for the 12 class medicines, discrimination was performed using the spectra of the prediction set. Along the decision tree, a spectrum is repeatedly identified with the classifier in the node until the sample is classified into a class. The operation for the identification in each node includes the projection into the accepted PC subspaces in the node, the rotation, the accumulation, and then identification with the threshold. The results for the samples in the prediction set are shown by the balls in the upper line in Figs. 4(a) - (k). From the figures, it can be seen that all the samples are correctly classified except for one sample in class E2, one sample in class A5 and eight samples in class D2, as shown in

Figs. 4 (g), (i) and (k), respectively. Further investigation shows that all the 10 samples are misclassified from class D1. Therefore, the large diversity of the samples in class D1 is the reason for the misclassification. From Table 1 it can be seen that the samples in class D1 and D2 are same medicine from different manufacturers. This is the reason for explanation of the eight samples misclassified from D1 into D2.

Table 2

To further investigate the performance of the method, the values of the true positive (TP) and false positive (FP) obtained by PCAcc, PCA, radial basis function artificial neural network (RBF-ANN) and partial least squares discriminant analysis (PLSDA) were summarized in Table 2. The two parameters are generally used to evaluate the performance of a classifier, which are defined as the ratio of the number of correctly classified and misclassified samples, respectively, to the total number of the samples in the class. From the table, it can be seen that PCAcc method produces the best result for the prediction set. Among the 12 classes of Chinese patent medicines, 8 classes are correctly classified, the true positive accuracies for the other four classes (E2, A5, D2 and D1) are 100%, 100%, 100%, 67.7%, and the false positive accuracies are 10%, 9.1%, 100%, 0.0%, respectively. Clearly, the results for all the 12 classes are acceptable except for the true positive accuracies of class D1 and the false positive accuracies of class D2. However, only five classes can be classified by PCA, and it is difficult to obtain acceptable results by RBF-ANN and PLSDA, because the true positive accuracies for some classes are lower than 50% and the false positive accuracies are even higher than 60%.

Conclusions

Discrimination of the 12 classes of Chinese patent medicines was studied using NIR spectroscopy and PCAcc method. CWT was adopted to eliminate the variant background in the NIR spectra. Because PCAcc method uses the accumulation of the information contained in multiple PC subspaces, an acceptable classification was achieved for different medicines or the same medicine from different manufacturers. Due to the advantage of the PCAcc in exploring as much as the classification information in the NIR spectra of the samples, PCAcc produced the best classification compared with the results of PCA, RBF-ANN and PLSDA.

Acknowledgements

This study was supported by National Natural Science Foundation of China (No. 21175074).

References

1. Y. Roggo, P. Chalus, L. Maurer, C. Lema-Martinez, A. Edmond and N. Jent, *J. Pharm. Biomed. Anal.*, 2007, **44**, 683–700.

2. W.J. Dong, Y.N. Ni and S. Kokot, *J. Agric.Food Chem.,* 2013, **61**, 540–546.

3 L. Tomuta, L. Rus, R. Lovanov and L. L. Rus, *J. Pharm. Biomed. Anal.*, 2013, **84**, 285–292.

4 O. Scheibelhofer, N. Balak, P. R. Wahl, D. M. Koller, B. J. Glasser and J. G. Khinast, *AAPS Pharmscitech*, 2013, **14**, 234-244.

5 D. S. Bu, B. Y. Wan and G. McGeorge, *Chemom. Intell. Lab. Syst.*, 2013, **120**, 84-91.

6 M. Boiret, L. Meunier and Y-M. Ginot, *J. Pharm. Biomed. Anal.*, 2011, **54**, 510–516.

7 S. S. Rosa, P. A. Barata, J. M. Martins and J.C. Menezes, *Talanta*, 2008, **75**, 725–733.

8 M. Blanco and A. Peguero, *TRAC-Trend. Anal. Chem*., 2010, **29**, 1127–1136.

9 A. D. Karande, P. W. S. Heng and C. V. Liew, *Int. J. Pharm.*, 2010, **396**, 63–74.

10 M. Blanco, R. Cueva-Mestanza and A. Peguero, *Talanta*, 2011, **85**, 2218–2225.

11 H. Grohganz, D. Gildemyn, E. Skibsted, J. M. Flink and J. Rantanen, *Anal. Chim. Acta*, 2010, **676**, 34–40.

12 C. C. Corredor, D. Bu and D. Both, *Anal. Chim. Acta*, 2011, **696**, 84–93.

13 X. B. Zhang, Y. C. Feng and C. Q. Hu, *Anal. Chim. Acta*, 2008, **630**, 131–140.

14 J. Moes, M. M. Ruijken, E. Gout, H. W. Frijlink and M. Ugwoke, *Int. J. Pharm.*,

2008, **357**, 108–118.

15 M. Blanco, M. Alcala, J.M. Gonzales and E. Torras, *J. Pharm. Sci.* 2006, **95**, 2137–2144.

16 M. Blanco and A. Peguero, *J. Pharm. Biomed. Anal.*, 2010, **52**, 59–65.

17 Y. Sulub, B. Wabuyele, P. Gargiulo, J. Pazdan, J. Cheney, J. Berry, A. Gupta, R. Shah, H. Wu and M. Khan, *J. Pharm. Biomed. Anal.*, 2009, **49**, 48–54.

18 A.U. Vanarase, M. Alcalà, J. I. Jerez Rozo, F .J. Muzzio, R. J. Romanach, *Chem. Eng. Sci.*, 2010, **65**, 5728–5733.

19 C. M. McGoverin, L. C. H. Ho, J. A. Zeitler, C. J. Strachan, K. C. Gordon and T. Rades, *Vib. Spectrosc.*, 2006, **41**, 225–231.

20 Y. Hu, A. Erxleben, A. G. Ryder and P. McArdle, *J. Pharm. Biomed. Anal.*, 2010, **53**, 412–420.

21 A. Heinz, M. Savolainen, T. Rades and C. J. Strachan, *Eur. J. Pharm. Sci.*, 2007, **32**, 182–192.

22 H. A. Gad, S. H. El-Ahmady, M. I. Abou-Shoer and M. M. Al-Azizi, *Phytochem. Anal.* 2013, **24**, 1-24.

23 Y. N. Ni, R. M. Song and S. Kokot, *Spectrochim. Acta A*, 2012, **96**, 252-258.

24 S. H. F. Scafi and C. Pasquini, *Analyst*, 2001, **126**, 2218–2224.

25 P. Li, G. R. Du, W. S. Cai and X. G. Shao, *J. Pharm. Biomed. Anal.*, 2010, **70**, 288–294.

26 Y. H. Lai, Y. N. Ni and S. Kokot, *Vib. Spectrosc*., 2011, **56**, 154-160.

27 W.J. Dong, Y.N. Ni and S. Kokot, *Appl. Spectros.*, 2014, **68**, 245-254.

28 P. Gemperline, *CRC/Taylor & Francis, Boca Raton.*, 2006.

29 R. W. Kennard and L. A. Stone, *Technometrics*, 1969, **11**, 137–148.

30 R. J. Barnes, M. S. Dhanoa and S. J. Lister, *Appl. Spectrosc*, 1989, **43**, 772-777.

31 P. Geladi, D. MacDougall and H. Martens, *Appl. Spectrosc*, 1985, **39**, 491-500.

32 X. G. Shao, A. K. M. Leung and F. T. Chau, *Acc. Chem. Res.*, 2003, **36**, 276–283.

33 X. G. Shao, G. R. Du, M. Jing and W. S. Cai, *Chemom. Intell. Lab. Syst*, 2012, **114**, 44-49.

34 J. J. Liu, W. S. Cai and X. G. Shao, *Sci. China: Chem.*, 2011, **54**, 802–811.

35 Y. Wang, X. Ma, Y. D. Wen, J. J. Liu, W. S. Cai and X. G. Shao, *Anal. Methods*, 2012, **4**, 2893-2899.

36 R. A. Fisher, *Ann. Hum. Genet.*, 1936, **7**, 179–188.

Figure captions

Fig. 1 Measured (a) and preprocessed (b) spectra of 12 classes of Chinese patent medicines.

Fig. 2 Distribution of the calibration samples in PC1-PC2 (a) and PC3-PC4 (b) subspaces for 12 classes of medicines.

Fig. 3 Resolution parameter (R_s) of the accepted subspaces (a) and their accumulated effect (b) in the discrimination of D1 and D2.

Fig. 4 Distribution of the samples along the accumulated PC axis for the calibration and prediction set of the medicines.

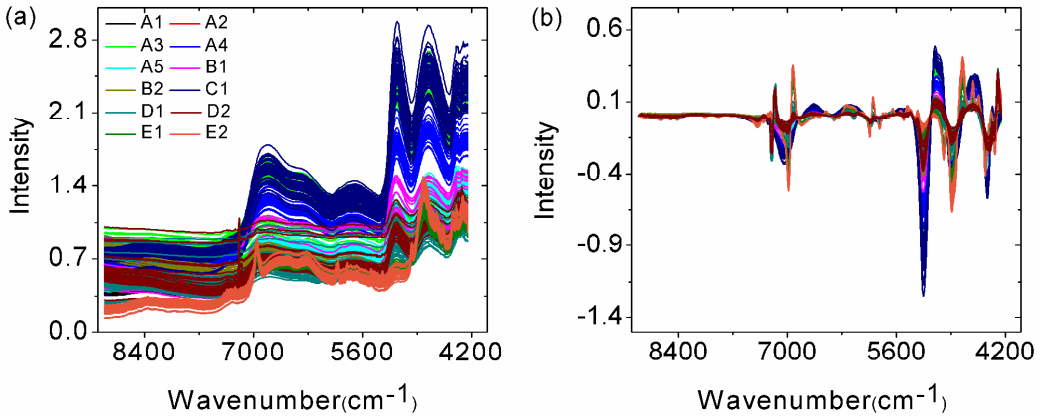


Figure 1

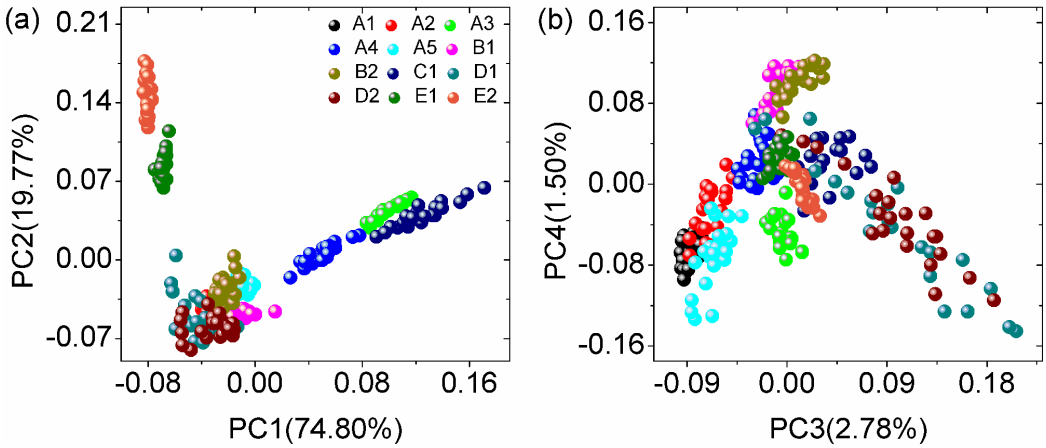


Figure 2

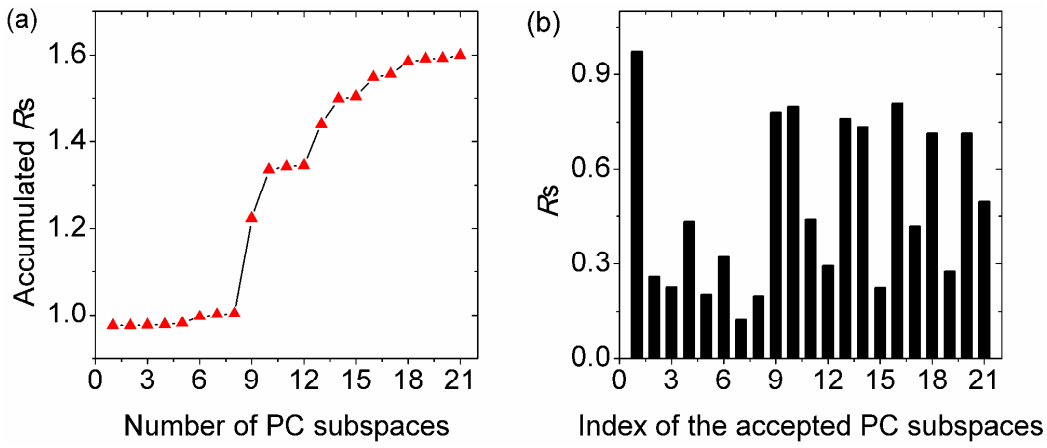


Figure 3

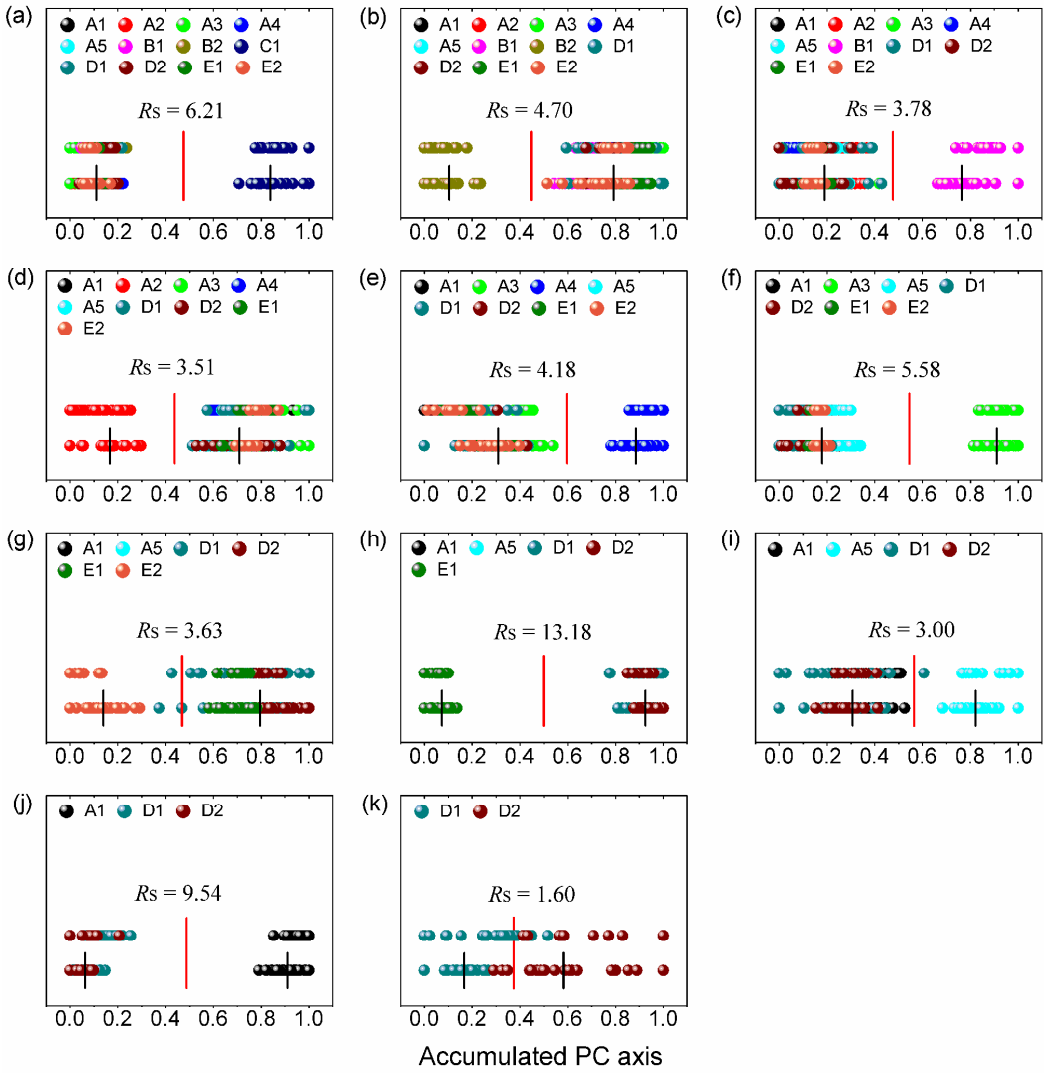


Figure 4

Table 1 Information of the samples in the calibration and prediction set

Medicine	Manufacturer	Class Label	Number of samples	Calibration set	Prediction set
A	1	A1	40	22	18
	2	A2	45	22	23
	3	A3	45	22	23
	4	A4	40	22	18
	5	A5	33	22	11
B	1	B1	42	22	20
	2	B2	47	22	25
C	1	C1	52	22	30
D	1	D1	53	22	31
	2	D2	30	22	8
E	1	E1	39	22	17
	2	E2	32	22	10
Total			498	264	234

Table 2 Comparison of the classification accuracy by different methods.

Class	Method[a]	Accuracy (TP, %)		Accuracy (FP, %)	
		Calibration set	Prediction set	Calibration set	Prediction set
C1	PCAcc	100.0	100.0	0.0	0.0
	PCA	100.0	96.7	0.0	3.3
	RBF-ANN	77.3	76.7	22.7	30.0
	PLSDA	86.4	46.7	31.8	33.3
B2	PCAcc	100.0	100.0	0.0	0.0
	RBF-ANN	77.3	92.0	22.7	24.0
	PLSDA	72.7	80.0	27.3	68.0
B1	PCAcc	100.0	100.0	0.0	0.0
	RBF-ANN	90.9	80.0	22.7	0.0
	PLSDA	59.1	80.0	22.7	25.0
A2	PCAcc	100.0	100.0	0.0	0.0
	RBF-ANN	45.5	52.2	45.5	21.7
	PLSDA	54.6	65.2	54.6	43.5
A4	PCAcc	100.0	100.0	0.0	0.0
	PCA	90.9	100.0	0.0	0.0
	RBF-ANN	72.7	83.3	45.5	44.4
	PLSDA	72.7	66.7	50.0	50.0
A3	PCAcc	100.0	100.0	0.0	0.0
	PCA	100.0	100.0	9.1	0.0
	RBF-ANN	90.9	100.0	63.6	43.5
	PLSDA	81.8	56.5	50.0	43.5
E2	PCAcc	100.0	100.0	4.6	10.0
	PCA	100.0	100.0	0.0	0.0
	RBF-ANN	95.5	100.0	0.0	0.0
	PLSDA	86.4	100.0	18.2	30.0
E1	PCAcc	100.0	100.0	0.0	0.0
	PCA	100.0	100.0	0.0	5.9
	RBF-ANN	95.5	100.0	4.6	0.0
	PLSDA	68.2	82.4	31.8	11.8
A5	PCAcc	100.0	100.0	0.0	9.1
	RBF-ANN	45.5	18.2	18.2	45.5
	PLSDA	54.6	54.6	31.8	18.2
A1	PCAcc	100.0	100.0	0.0	0.0
	RBF-ANN	54.6	72.2	13.6	22.2
	PLSDA	54.6	72.2	13.6	27.8
D1	PCAcc	95.5	67.7	13.6	0.0
	RBF-ANN	59.1	54.8	27.3	6.5
	PLSDA	50.0	48.4	31.8	9.7
D2	PCAcc	86.4	100.0	0.0	100.0
	RBF-ANN	81.8	100.0	27.3	75.0
	PLSDA	59.1	75.0	36.4	50.0

[a] PC1-PC2 was used in PCA method, and the same latent variable number as in PCAcc was used in PLSDA. 33 hidden neurons were used in RBF-ANN. For the classes that can not be discriminated by PCA, the result was not listed.