Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 124 (2014) 52–58

Contents lists available at ScienceDirect

Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy

journal homepage: www.elsevier.com/locate/saa

Discrimination of three dimensional fluorescence spectra based on wavelet analysis and independent component analysis

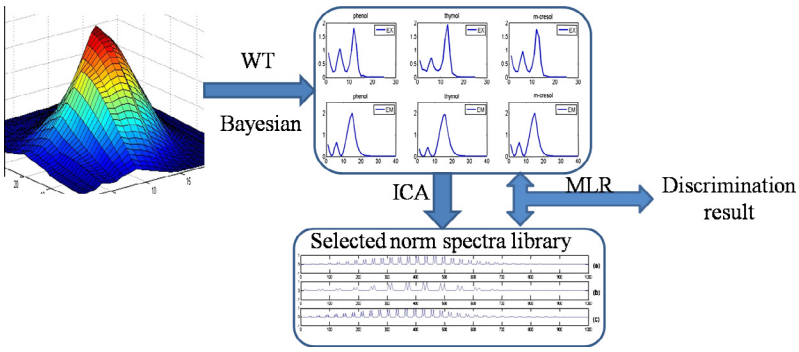
Xiaoya Yu, Yujun Zhang *, Gaofang Yin, Nanjing Zhao, Xue Xiao, Changhua Lu, Yanwei Gao, Wei Zhang

Key Laboratory of Environmental Optics & Technology, Chinese Academy of Sciences, Anhui Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Hefei 230031, China

HIGHLIGHTS

- Wavelet analysis and ICA are applied for recognition of overlapped spectra.
- Wavelet analysis extracts the features of the spectra and amplifies differences.
- ICA analysis is used to separate single component before linear regression.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 5 September 2013
Received in revised form 17 November 2013
Accepted 5 December 2013
Available online 21 December 2013

Keywords:
Blind signal separation
Three dimensional fluorescence spectra
Wavelet analysis
Independent component analysis

ABSTRACT

Fluorescence spectroscopy is a rapid and non-destructive method for monitoring water quality. In this work, wavelet analysis, together with independent component analysis (ICA), was applied for component recognition of seriously overlapped, multi-component, three dimensional fluorescence spectra. Wavelet analysis extracts the features of the spectra and amplifies differences among phenolic homologs. ICA analysis in blind signal separation was used to separate single component before multiple linear regression (MLR). The proposed method increases the correct classification rate and enriches the spectra library. As such, it is a useful alternative to traditional techniques in component recognition.

Introduction

Fluorescence spectroscopy is a rapid and non-destructive method used in vivo and in situ water quality monitoring because of its high sensitivity and good expression of features [1,2]. Methods for spectra analysis based on the three-linear model have received significant attention over the past several years. Among known methods, parallel factor analysis (PARAFAC) is the most classical one [3–5]. However, non-multi-linear problems are very common in this field, and such problems may be attributed to [6]: (1) The non-linear relationship between signals and analyte concentrations, (2) the non-multi-linear of signals, and (3) the variation in component profiles across different samples.

Models that allow deviations of multi-linearity in one way or another include: parallel profiles with linear dependencies (PARALIND) [7], multivariate curve resolution couple to ALS (MCR-ALS) [8], non-bilinear rank annihilation (NBRA), [9] unfolded partial least-squares (U-PLS), [10] multi-way PLS (N-PLS), [11] and artificial neural networks (ANN) [12,13]. Models must be selected according to the cause of deviation. For example, U-PLS can be used for three dimensional fluorescence spectra, whereas ICA can be used to analyze second order data, regardless of whether the data is in accordance with the three-linear model or not [14,15].

* Corresponding author. Tel.: +86 551 65593691; fax: +86 551 65593530.
E-mail address: yjzhang@aiofm.ac.cn (Y. Zhang).

1386-1425/$ - see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.saa.2013.12.033

Moreover, the non-negativity constraint is unnecessary when using the ICA [16]. When the extracted proportions are negative, the proportions should be multiplied by −1.

Zhang et al. converted three dimensional spectra into two dimensional spectra [17,18], extracted fluorescence features from harmful algal bloom(HAB) species and discriminated the algae at the division and genus level. Wavelet analysis combined with Bayesian discriminant analysis, a method established by MLR, was used to determine discriminant spectra. This method discriminates the algae in the wavelet domain, thereby avoiding the errors caused by tri-linearity deviations. However, this method only discriminates the species in the spectral library. Even worse, operations containing all spectra increase the risk of error. Therefore, blind separation must be executed before discrimination. Thus MLR only takes specific components into account. When the result of blind separation contains the spectra that are not included in the spectral library, these spectra are considered new species and added to the library. Therefore, the proposed process not only increases the correct classification rate, but also enriches the spectral library.

In the present work, wavelet analysis and ICA are used to analyze three dimensional fluorescence spectra and facilitate the monitoring of water quality. In the next section, the theoretical bases of our method are introduced in detail. The section thereafter verifies the feasibility of the proposed method by experiments. Finally, a concise conclusion and some remarks are given.

Theories

Wavelet analysis

Wavelet multi-scale decomposition, also called "mathematics microscope" [19], can refine the intrinsic information of data and extract inner relations. Local representation information in terms of both time and frequency can be extracted by wavelet analysis. The wavelet features of the spectra are the projections of original spectra in the wavelet space [20–22].

The wavelet analysis theory of fluorescence spectra is as follows:

The fluorescence spectra of organic matter is: $H(f), f = 1, 2, \ldots, F$, where F is the number of measured points. First, wavelet analysis refines $H(f)$ into multi-scale signals by selecting an appropriate orthogonal scale base $\Phi_{j,n}(f)$ and corresponding wavelet base. The relationship of scale space Φ_j and wavelet space Ψ_j is as follows:

$$\Phi_j \perp \Psi_j \tag{1}$$

$$\Phi_{j+1} = \Phi_j \oplus \Psi_j \tag{2}$$

Then, the scale component $b_{j,n}$ and wavelet component $e_{k,n}$ can be obtained:

$$b_{j,n} = \sum_{f=1}^{F} H(f)\Psi^*_{j,n}(f) \tag{3}$$

$$e_{k,n} = \sum_{f=1}^{F} H(f)\Phi^*_{k,n}(f) \quad k = 1, 2, \ldots, j \tag{4}$$

The scale component (also called the low-frequency component) represents most of the measured points and contains large scale information. The wavelet component (also called the high-frequency component) reveals few findings in several measured points and contains small scale information. Thus, the scale component was used as standard spectral feature in this paper. Daub4 (Daubechies wavelet with four filter coefficients) [23] was employed as the mother wavelet because it is the most local in terms of time domain [21].

Independent component analysis

ICA is a signal processing technique that aims to recover underlying source signals from a set of mixed signals, based on the assumption that source signals are statistically independent [15]. ICA has been applied to spectroscopic data [24], speech recognition [25], blind signal separation [26], fault detection [27], statistical process monitoring [28], and batch process monitoring [29].

The ICA of a random vector is used to search the linear transformation that minimizes the statistical dependence among its components [15,30]. To design a practical optimization criterion, the expansion of mutual information is used as the function of cumulants.

The following linear statistical model is assumed as:

$$y = Px + v \tag{5}$$

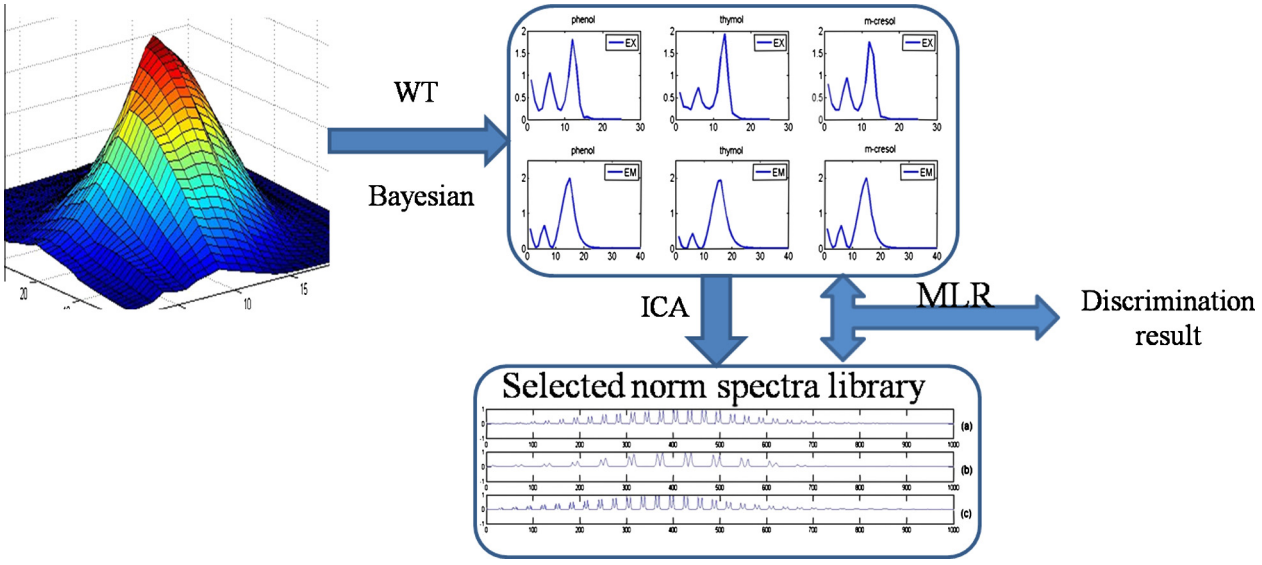


Fig. 1. Main steps of dealing with spectra data.

where x, y and v are random vectors with zero mean and finite covariance, and P is a regular square matrix. The ICA problem is that both P and its corresponding x can be calculated from a given y.

A fast fixed-point algorithm for ICA is used in this paper [31]. The number of independent components (ICs) is evaluated by finding the break of statistical value of the factor analysis functions which are the eigenvalue (EV), the logarithm of eigenvalue (Log EV), and ratio of the ith eigenvalue (EVR) to its previous one [32].

Main steps of dealing spectra

N mixture samples of M components are measured by fluorescence spectrophotometer, and three dimensional multi-component fluorescence spectra $Y_n \in R^{I\times J}$ $(n = 1, \cdots, N)$ are obtained; here, I is the number of excitation wavelengths and J is the number of emission wavelengths. Based on the Beer–Lambert law and the add-sum principle of multi-component spectra, the linear separation model can be described as follows:

$$Y_n = a_{n,1}X_1 + \cdots + a_{n,M}X_m + E_n \ (n = 1, \ldots, N) \tag{6}$$

where X_i $(i = 1, \ldots, M) \in R^{I\times J}$ are the original signal of the spectra, $a_{n,i}$ $(i = 1, \ldots, M) \in R^{I\times J}$ are the corresponding concentrations, and E_n indicates noise residuals. When $Y_n \in R^{I\times J}(n = 1, \ldots, N)$ and X_i $(i = 1, \ldots, M) \in R^{I\times J}$ are unfolded, the linear separation model above has the equivalent expression:

$$Y = AX + E \tag{7}$$

where $Y \in R^{N\times(I\times J)}$ $(n = 1, \ldots, N)$ are the known measured spectra, $A \in R^{N\times M}$ and $X \in R^{M\times(I\times J)}$ are unknown hybrid matrix and source spectra signal respectively, and $E \in R^{N\times(I\times J)}$ is the noise.

The main steps applied in dealing with spectral data, as shown in Fig. 1 are as follows:

Step 1. Analyze two dimensional spectra data by wavelet transformation.
Step 2. Find the reference spectra by Bayesian discrimination from the scale vectors and establish a reference spectral library by cluster analysis [17,18].
Step 3. Blindly separate the wavelet scale vectors of the mixed spectra by ICA using a fast fixed-point algorithm, where the

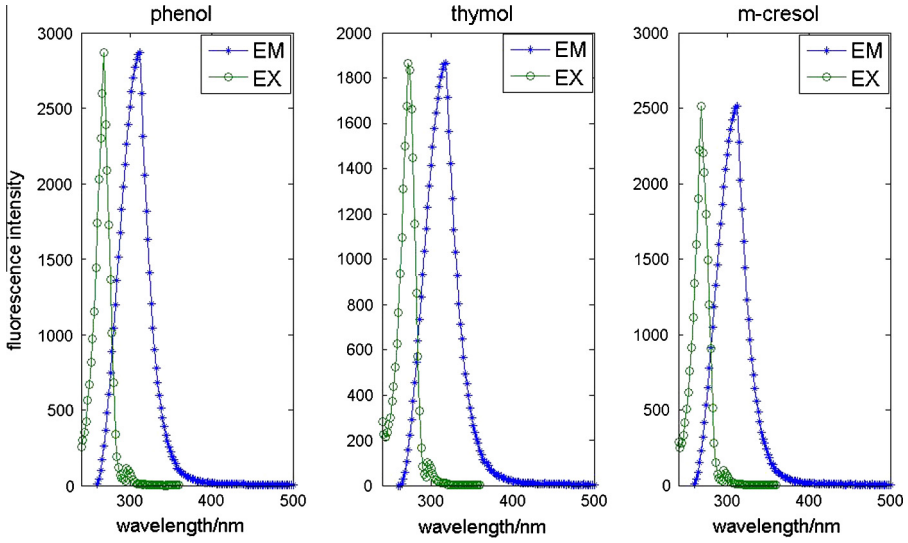


Fig. 2. Emission and excitation spectra of three phenolic compounds.

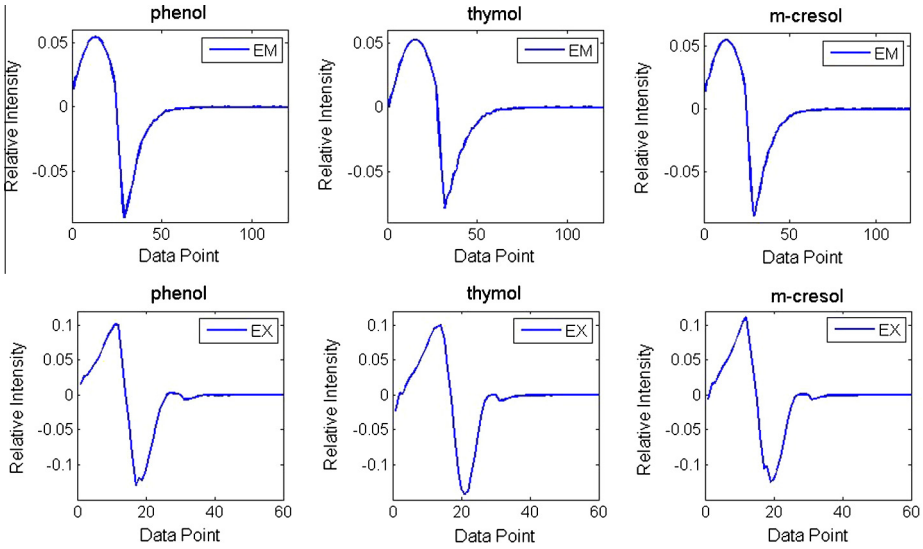


Fig. 3. Differential emission and excitation spectra of three phenolic compounds.

input signals are wavelet scale vectors of the mixed spectra and the output signal is the solution after mixing.

Step 4. Based on the reference spectra database, establish a fluorometric quantify method by MLR.

Experiment and discussion

Experiment description

The reagents [phenol (Shanghai Chemical Reagent Co., Ltd., PR China), thymol (Tianjin Guangfu Fine Chemical Research Institute, PR China), *m*-cresol(Shanghai Chemical Reagent Co., Ltd., PR China), and their mixtures] were prepared. Stock solutions (100 mg L^{-1} for each phenolic compound) were prepared by dissolving the appropriate phenolic compound in HPLC-grade ethanol, and the solutions were stored for use. Standard solutions (100 μg L^{-1} for each phenolic compound) were made by diluting the stock solutions in ultrapure water, and stored at the same conditions as the stock solutions.

The reagents were scanned by a Hitachi F-7000 fluorescence spectrophotometer (Hitachi High-Technologies Corporation, Tokyo, Japan) with excitation wavelengths of 260–500 nm and emission wavelengths ranging from 240 nm to 360 nm. During measurement, the scanning intervals were taken as 2.0 nm for both excitation and emission wavelengths.

The similarity coefficient p between spectra s_i and spectra s_j is calculated by:

$$p = \left| \frac{s_i \times s_j^T}{\|s_i\| \|s_j\|} \right| \tag{8}$$

where s_i, s_j are spectral components. Based on this definition, $0 \leqslant p \leqslant 1$ can be obtained; here, the larger the value of p is, the more similar the spectra are. The value of p can be used as a recognition threshold during spectral analysis.

Table 2
Reagents of different concentrations (unit: mg/L).

	No. 1	No. 2	No. 3	No. 4	No. 5
Phenol	0.167	0.15	0.025	0.165	0.125
Thymol	0.2	0	0.08	0.044	0.08
m-Cresol	0	0.28	0.1	0.168	0.025

Spectra pre-processing

The feature excitation and emission spectra of phenol, thymol, and *m*-cresol are shown in Fig. 2. These organic compounds are priority pollutants in water, and their spectra overlap extensively, thus, these reagents were selected for detection in this paper.

Given that the spectra of the compounds are very similar to each other, as shown in Fig. 2 analysis of Formula (5) is inapplicable when the spectra are operated directly. Differential and wavelet analyses are used to improve the resolution of the spectra and reduce correlations.

Figs. 3 and 4 show that the correlations among the spectra are significantly reduced; this decrease is even more obvious after wavelet analysis. This result may be attributed to the fact that the component selected for the frequency domain has the largest difference among the phenolic compounds. In this paper, the third scale component is selected as the feature wavelet spectra. In practice, two or more components can be combined as feature spectra according to requirement synthesize effects among different wavelet components. The strength of the first peak differs significantly among the emission spectra of the three phenolic compounds in Fig. 4 and the rate of change between the 5th and 10th data point of thymol is lower than of two other phenolic compounds because these data points contain distinctive frequency components. Wavelet analysis results in more spectral peaks, more incisive spectral band and higher resolutions than the original spectra.

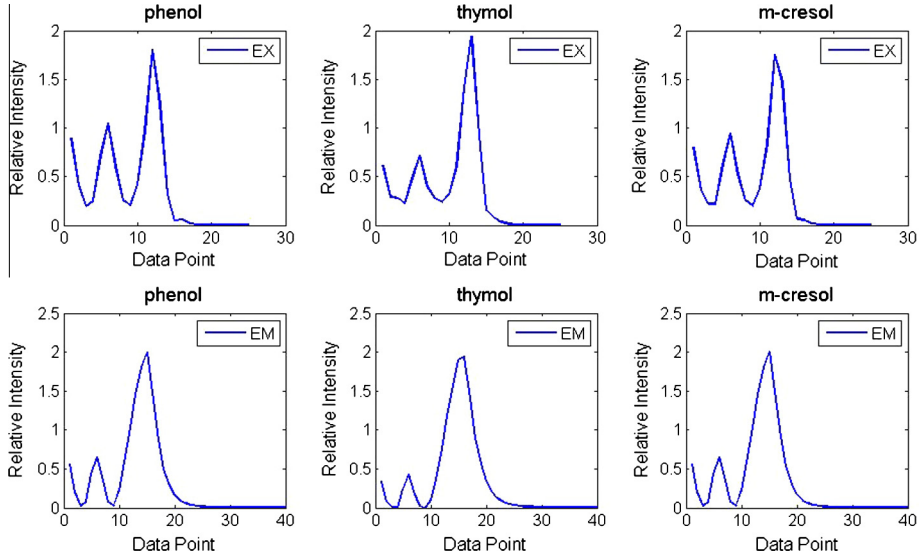


Fig. 4. Emission and excitation frequency domain spectra of three phenolic compounds.

Table 1
Similarity coefficients of the three phenolic compounds after transformation.

Original/diff/wavelet	Phenol	Thymol	*m*-Cresol
Phenol	–	0.9907/0.8213/0.7867	0.9568/0.8311/0.7374
Thymol	0.9907/0.8213/0.7867	–	0.9238/0.9007/0.7089
m-Cresol	0.9568/0.8311/0.7374	0.9238/0.9007/0.8089	–

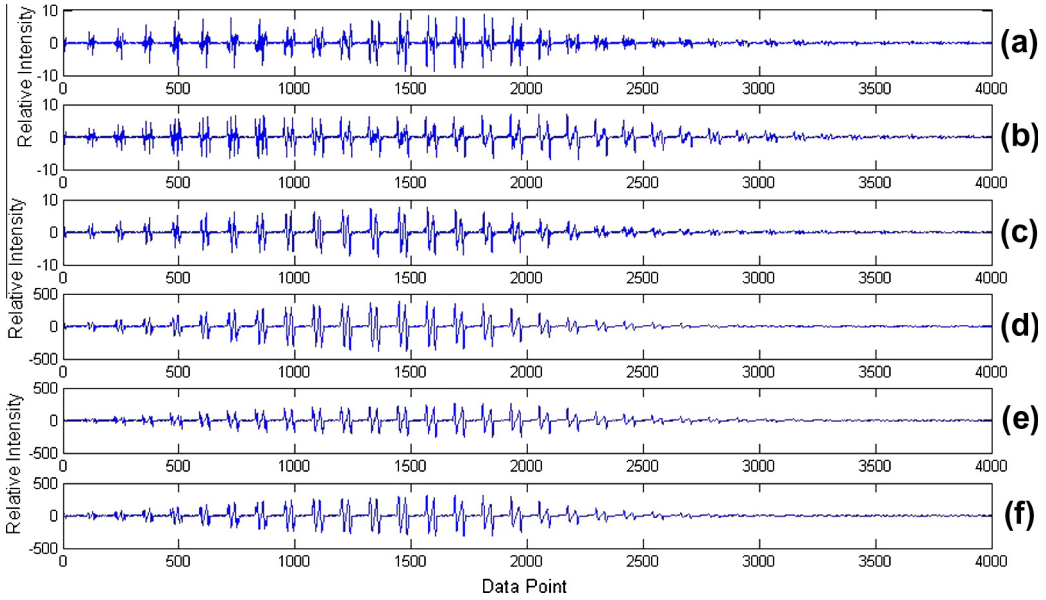


Fig. 5. Norm and calculated differential emission spectra of the three phenolic compounds (a–c) are the norm emission differential spectra. (d–f) are the calculated emission differential spectra.

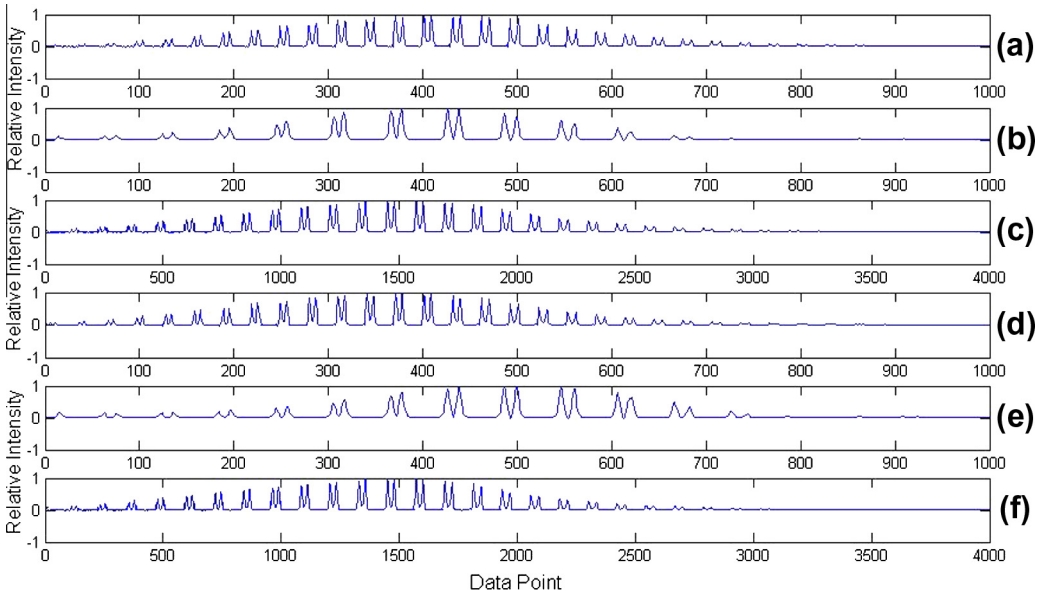


Fig. 6. Norm emission and calculated excitation frequency domain spectra of the three phenolic compounds. (a–c) are the norm emission frequency domain spectra. (d–f) are the calculated emission frequency domain spectra.

The p of the three phenolic compounds obtained after wavelet analysis are the smallest among three methods shown in Table 1. The decline in p enables more efficient separation.

Experiment results and discussion

Five mixtures of different concentrations of phenol, *m*-cresol, and thymol are presented in Table 2.

Taking the spectra of no. 1 as an example, each row of the three-dimensional fluorescence spectrum connected with the next one forms a large emission spectrum, and each column connected forms a large excitation spectrum. After executing differential and wavelet analyses on the emission spectra, three calculated emission spectra can be extracted from the mixed spectra through ICA, as shown in Figs. 5 and 6. The component number is 3 in this experiment.

The p values obtained by proposed method are larger than those obtained by two other methods, as shown in Fig. 7. The difference in emission spectra is smaller than the difference in excitation spectra, but the p values of the emission wavelet spectra and emission differential spectra (0.9786 and 0.9647, respectively) are larger than those of their excitation counterparts respectively (0.9188 and 0.9128, respectively). Although superiority in frequency domain p is less significant than that in the differential domain, superiority in the excitation spectra is significant and increases of p in phenol and *m*-cresol are 1.4% and 2.1%, respectively. In the excitation spectra of phenol, the p is low when analyzed by the original method, partly because phenol and *m*-cresol show similar waveform peak point positions and waveform broadening

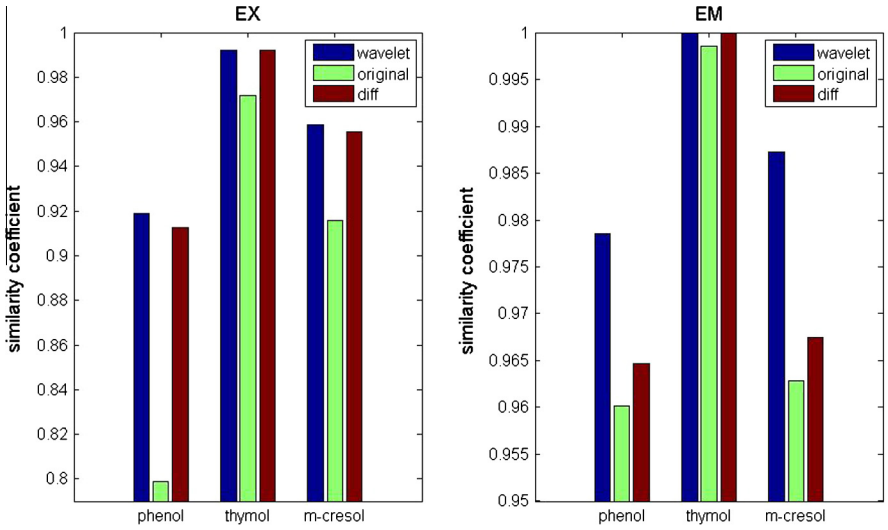


Fig. 7. Similarity coefficients of the three phenolic compounds after transformation.

Table 3
Correct classification rates of the three methods.

Correct classification		Excitation spectra			Emission spectra		
		Phenol (%)	Thymol (%)	*m*-Cresol (%)	Phenol (%)	Thymol (%)	*m*-Cresol (%)
Rate	Original	84	82.5	83.4	87.4	79.5	82.6
	Diff	93.7	89.7	92.1	92.3	94	89.4
	Wavelet	98.2	96.4	99.9	98.8	98.6	99.8

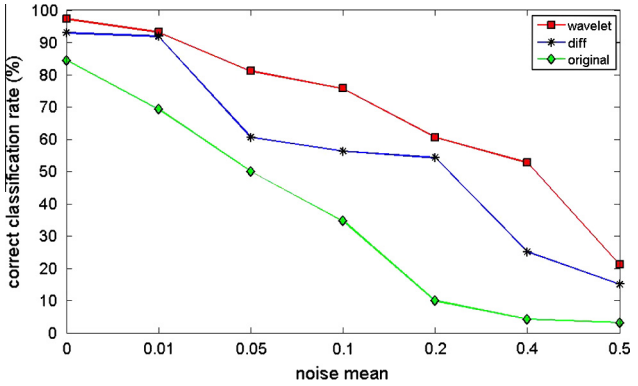


Fig. 8. Correct classification rate of the three methods at different noise mean.

as shown in Fig. 2. The correct classification rates which are the numbers of correctly discriminated in the percentage of the total number are listed in Table 3. The results indicate that the correct classification rates are higher than 0.98 for phenol and *m*-cresol using the proposed method; by contrast, a slight decrease is observed for thymol.

To improve the correct recognition rates in multi-component fluorescence spectra, the similarity coefficients of the emission and excitation spectra were combined. Only the two coefficients simultaneously exceeded the threshold; thus, the spectra are recognized as the specific matter.

Noise immunity

To investigate the anti-noise properties of the three methods, noise of different averages was added to the spectral data. The correct classification rates of the three methods at different noise averages are shown in Fig. 8.

While ICA can overcome the adverse effects of noise to a certain extent, dividing the calculated spectra into different classes is difficult when the noise increases to a specific mean. For example, when the noise mean is 0.1, the correct classification rate drops to 34.7%. Given that differential transformation is sensitive to noise, the roughness penalty approach should be used. The accuracy of differentiation and denoising increases but remains lower than that obtained from wavelet analysis, likely because wavelet analysis simultaneously performs time-domain and frequency-domain transformations, whereas differential transformation is performed separately from denoising, which leads to loss of some important information. The two other organics show consistent trends. The method used in this paper has features of robustness and anti-attacking. As the original spectra are difficult to recover correctly from heavy noise, signals must contain as little noise as possible.

Conclusions

In this study, wavelet analysis combined with ICA was applied for component recognition of seriously overlapped, multi-component, three dimensional fluorescence spectra. Wavelet analysis can extract salient features of spectra, enabling more efficient differentiation of phenolic homologs. ICA was used to separate single component. The proposed method was proved to be successful in reducing error recognition rates and enriching the spectral library.

Experiments on laboratory samples were performed to determine the effectiveness of the proposed method. The proposed method can be applied in water pollution monitoring and cases that do not fit a three-linear model. Therefore, blind separation

of wavelet components is a potential alternative to current methods of performing three dimensional spectral analysis.

Acknowledgments

This work is supported by National Natural Science Foundation of China (61378041), National 863 Program of China (2013AA065502), Excellent Youth Foundation of Anhui Scientific Committee (1108085J19), Foundation of Director of Anhui Institute of Optics and Fine Mechanics (Y03AG31144), the Natural Science Foundation of Anhui Province (11040606M26) and Chinese Academy of Equipment Functional Development of Technological Innovation Project (yg2012071).

References

[1] C.A. Stedmon, S. Markager, R. Bro, Mar. Chem. 8 (2003) 239–254.
[2] L. Poryvkina, S. Babichenko, A. Leeben, in: Proc. EARSeL-SIG-Workshop LIDAR, vol. 6, 2000, pp. 224–232.
[3] R.A. Harshman, UCLA WPP. 16 (1970) 1–84.
[4] J.B. Kruskal, Proc. Symp. Appl. Math. 28 (1983) 75–704.
[5] C.W. Snyder, W.D. Walsh, P.R. Pamment, J. Appl. Psychol. 68 (1983) 572–583.
[6] A.C. Olivieri, G.M. Escandar, A. Muñoz de la Peña, TrAC 30 (2011) 07–617.
[7] M. Bahram, R. Bro, Anal. Chim. Acta 584 (2007) 397–402.
[8] R. Tauler, Chemom. Intell. Lab. Syst. 30 (1995) 133–146.
[9] M.M. Reis, S.P. Gurden, A.K. Smilde, M.M.C. Ferreira, Anal. Chim. Acta 422 (2000) 21–36.
[10] S. Wold, P. Geladi, K. Esbensen, J. O hman, J. Chemom. 1 (1987) 41–56.
[11] R. Bro, J. Chemom. 10 (1996) 47–61.
[12] S. Haykin, Neural Networks, NJ, USA, 1999.
[13] F. Marini, R. Bucci, A.L. Magri, A.D. Magri, Microchem. J. 88 (2008) 178–185.
[14] D.J.R. Bouveresse, H. Benabid, D.N. Rutledge, Anal. Chim. Acta 589 (2007) 216–224.
[15] P. Common, Signal Process. 36 (1994) 287–314.
[16] F. Ammari, C.B.Y. Cordella, N. Boughanmi, D.N. Rutledge, Chemom. Intell. Lab. Syst. 113 (2012) 32–42.
[17] F. Zhang, R. Su, X. Wang, et al., J. Exp. Biol. Ecol. 368 (2009) 37–43.
[18] F. Zhang, R. Su, J. He, et al., Spectrochim. Acta, Part A 75 (2010) 578–584.
[19] L. Wang, C. Ji, Cot. Sci. 2 (2006) 124–128.
[20] I. Daubechies, IEEE T. Inform. Theory 36 (1990) 961–1005.
[21] S. Surya, J.E. Powers, W.M. Grady, IEEE T. Power Deliver 4 (1996) 924–930.
[22] P. Abry, D. Veitch, IEEE T. Inform. Theory 1 (1998) 2–15.
[23] I. Daubechies, Ten Lectures on Wavelets, Philly, USA, 1992.
[24] E. Visser, T.W. Lee, Chemom. Intell. Lab. Syst. 70 (2004) 147–155.
[25] G.D. Brown, S. Yamada, T.J. Sejnowski, Trends Neurosci. 24 (2001) 54–63.
[26] C.M. Kim, H.M. Park, T. Kim, Y.K. Choi, S.Y. Lee, IEEE Trans. Neural Netw. 14 (2003) 1038–1046.
[27] M. Kano, S. Tanaka, S. Hasebe, I. Hashimoto, H. Ohno, AlChE J. 49 (2003) 969–976.
[28] J.M. Lee, C. Yoo, I.B. Lee, J. Chem. Eng. Jpn. 36 (2003) 563–577.
[29] C.K. Yoo, J.M. Lee, P.A. Vanrolleghem, I.B. Lee, Chemom. Intell. Lab. Syst. 70 (2004) 151–163.
[30] A.J. Bell, T.J. Sejnowski, Neural Comput. 7 (1995) 1129–1159.
[31] A. Hyvärinen, E. Oja, Neural Comput. 9 (1997) 1483–1492.
[32] K. Wongravee, T. Parnklang, P. Pienpinijtham, et al., Phys. Chem. Chem. Phys. 15 (2013) 4183–4189.