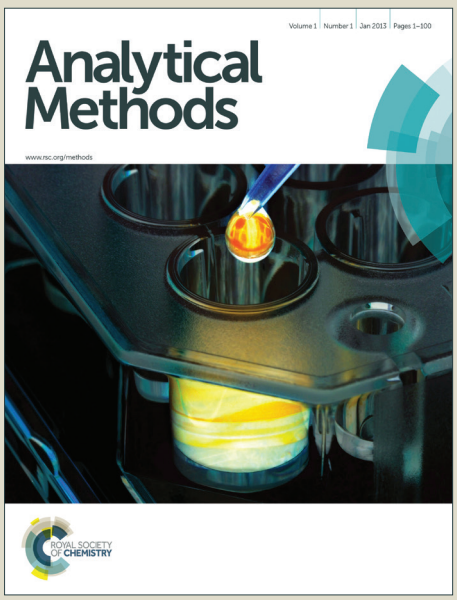
Analytical Methods

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/methods

Analytical Methods

RSCPublishing

ARTICLE

Coding method for study of intrinsic mechanism of spectral analysis

Cite this: DOI: 10.1039/x0xx00000x

Mei Zhou, [a] Qingli Li,*[a] Gang Li[b] and Ling Lin[b]

Received 00th January 2012,
Accepted 00th January 2012

DOI: 10.1039/x0xx00000x

www.rsc.org/

In view of many factors that influence the in-situ real-time spectral measurements, various correction and modeling methods have been applied on spectral analysis. However, the intrinsic mechanism of these methods in actual applications was not usually demonstrated, so chance correlations could be unavoidable. This paper presents a new method coding the absorbance spectrum of each component in a multi-component compound to quantize the size relation between different absorbance of various components. To describe the implementation process, spectra of the compound with three components were developed and partial least-squares regression was used to construct calibration models. The results verify the feasibility of the coding method used for studying the intrinsic mechanism of spectral analysis. The new method provides a way to analyze the influence from different components of the object qualitatively and quantitatively, and it can help us to grasp the action mechanism of these correction and modeling methods. Based on this, we can choose and design a suitable method to enhance the accuracy and reliability of spectral analysis.

Introduction

Spectroscopy technology due to its advanced features of non-invasive, high-speed, high-precision, multi-information and easy-to-operate, has been widely applied in the fields, such as food, agriculture, environment, and medicine.[1-3] It is usually implemented in laboratory analysis and has gradually been extended to the in-situ real-time monitoring and analysis.[4] However, for the in-situ real-time measurement, the spectrums of different components are overlapping so that the selectivity of the measured spectra is degraded, and also the spectra could vary with measurement conditions and different physical properties of the object (*e.g.*, object surface and particle size). To reduce the influence of factors mentioned above and achieve more accurate results of quantitative spectral analysis, the reported methods mainly fall into three categories: spectral correction methods, wavelength (or wavelength region) selection methods,[5,6] and linear and non-linear modeling methods.[7,8] Spectral correction methods include multiplicative signal correction (MSC),[9] standard normal variate (SNV),[10] optical path-length estimation and correction (OPLEC)[11] and so on. Wavelength selection methods include stepwise regression, genetic algorithm (GA),[12] interval partial least-squares (iPLS)[13] and so on. And linear/non-linear modeling methods include partial least-squares (PLS),[14,15] principal components regression (PCR), support vector machine (SVM)[16] and so on. Most of the researches used finite experimental data sets to verify the feasibility of these methods, and they didn't analyze the intrinsic mechanism of the used method in actual spectral processing. Therefore, chance correlations are unavoidable,[17,18] which limits their extention to other applications simply.

To enhance the accuracy of spectral analysis, this paper proposes a novel coding method which generates a simulated spectrum of a multi-component compound through coding the absorbance spectrum of each component. The coding method can quantize the size relation between absorbance of multi components in the object. Utilizing our proposed method, it is easy to analyze the influence from each component qualitatively and quantitatively, and also we can clearly understand the action mechanism of traditional correction and modeling methods. The principle and the implementation of the new method are described as follows. The spectra of a compound with three components were developed, and the PLS regression was used to construct calibration models to verify the feasibility.

Methods

Coding Method

Suppose that the object to be measured is *S* that is made up of three components: *a*, *b*, *c*. For one of the components such as *a*, there are five kinds of size relations between absorbance *A* of another component such as *b* and itself at a given wavelength. They are shown as follows: $A_a=A_b$, $A_a>A_b$, $A_a>>A_b$, $A_a<A_b$, $A_a<<A_b$. The relations between any other two components are similar, and they are not listed here individually. To represent the size relations among absorbance of three components, we firstly quantize the absorbance into three numerical values: 1, 2, 3. Therefore, two same values stand for the relation "=", two adjacent values stand for ">" or "<", and two non-adjacent values stand for ">>" or "<<". We use three numerical values to express the absorbance of three components in the object *S*, and

there are 3×3×3 combinations in total, which are shown in Table 1. The process that we use the setting values to quantize the absorbance of multi components in the object is called the coding method. One combination is called a kind of coding.

Table 1 All the coding for three components *a*, *b*, *c* based on the setting numerical values 1, 2, 3.

a	*b*	*c*	*a*	*b*	*c*	*a*	*b*	*c*
1	1	1	2	1	1	3	1	1
1	1	2	2	1	2	3	1	2
1	1	3	2	1	3	3	1	3
1	2	1	2	2	1	3	2	1
1	2	2	2	2	2	3	2	2
1	2	3	2	2	3	3	2	3
1	3	1	2	3	1	3	3	1
1	3	2	2	3	2	3	3	2
1	3	3	2	3	3	3	3	3

Design of Coding Absorbance Curve

The relations among absorbance of different components in a compound are complex in the ultraviolet and visible or near infrared spectral range. They can be usually described by most of the coding in Table 1. To perform a complete analysis of different relations, 27 coding in Table 1 were all utilized to construct absorbance curves of three components. The absorbance curve covers 270 wavelengths with a wavelength resolution of 1nm. The whole wavelength region is divided into 27 segments with a step size of 10 wavelengths. The same segment of three components corresponds to one kind of the coding. The constructed absorbance curves which are also called coding absorbance curves are shown in Fig. 1. Moreover, considering the practical implication of coding absorbance curves, we multiplied coding absorbance curves of three components *a*, *b*, *c* by 0.3 respectively. The calculated results were all defined as the absorbance curves with the concentration of 1mol/L and the optical path length of 1cm.

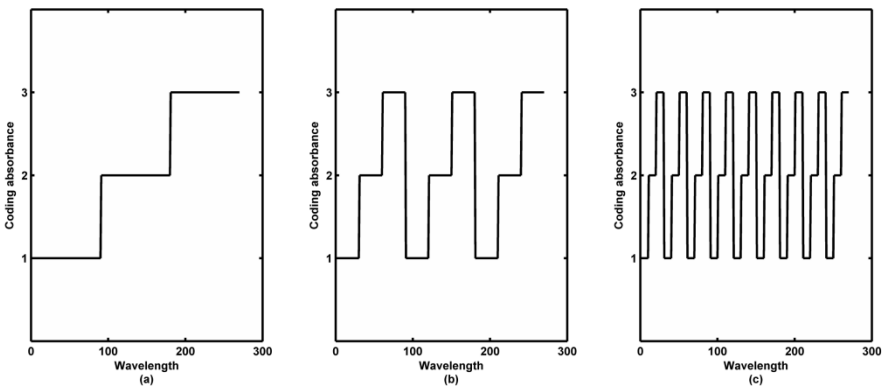


Fig. 1 Coding absorbance curves of three components, (a) component *a*, (b) component *b*, (c) component *c*.

Construction of Simulated Spectra

The setting concentration ranges of three components are 1-2mol/L (component *a*), 0.1-0.5mol/L (component *b*), and 0.001-0.01mmol/L (component *c*) respectively. In the concentration range of each component, 300 random concentration values were generated, so no correlation existed between the concentrations of any two components. According to additive effect of absorbance in Eq. (1), 300 pure spectra of the compound with different concentrations of three components were calculated by multiplying random concentration values of three components and their absorbance curves.

$$A_{total} = A_a + A_b + A_C \quad (1)$$

However, the measured spectra in practical situations are unavoidably disturbed by random noise and system noise. Therefore, we added white Gaussian noise, whose amplitude was the average absorbance value of the component *c* with the concentration of 0.0001mol/L at all the wavelengths, to the simulated pure spectra. Any four absorbance spectra from 300 simulated ones after adding noise are shown in Fig. 2. We can clearly see that the spectra of the compound have the features of absorbance curves of components *a* and *b*, which is due to higher concentrations of the two components.

Modeling Analysis

Partial least-squares regression was used to build calibration models to verify the feasibility of the coding method for spectral analysis. The component *c* which has the minimum concentration was chosen as the component to be measured. The simulated spectra were sorted according to numerical order of the concentration values of the component *c*, with consideration of the effect of the concentration distribution. For every three continuous spectra, the middle one was chosen for the prediction set. And 200 spectra were for the calibration set and 100 for the prediction. Then the calibration models were built with the number of loading vectors changing from 1 to 6. The number of loading vectors which provides the minimum the root-mean-square of prediction (RMSEP) was lastly determined. And loading vectors with the optimal number were extracted to compare with absorbance curves of three components respectively. Moreover, the PLS regression coefficients of the model by using the optimal number of loading vectors were compared with the absorbance curves of

three components.

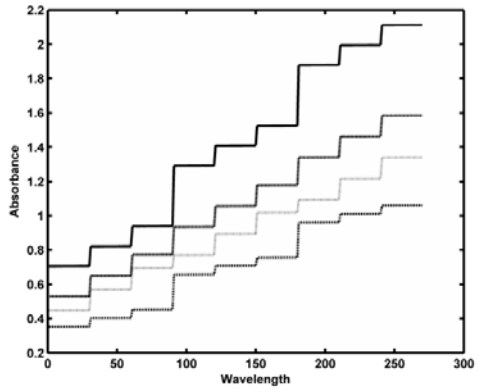


Fig. 2 Any four absorbance spectra after adding noise.

Results and Discussion

For the PLS calibration models based on the simulated spectra, the changing relationship between RMSEP values and the number of loading vectors is shown in Fig. 3. When the number of loading vectors is equal to 3, the RMSEP reaches its minimum value. However, when the number of loading vectors gets greater than 3, the RMSEP value gradually increases. When the number of loading vectors is equal to 1 or 2, the calibration models have hardly any prediction performance for the component *c*. The results show that the calibration model achieves the best performance when the number of loading vectors is equal to the number of components in the object *S*, and the third loading vector determines the performance of the calibration model.

Three loading vectors used at last in the PLS calibration model are shown in Fig. 4. Compared with the changing trend of absorbance curves of three components, the first loading vector has positive correlation with both absorbance curves of components *a* and *b*, the second loading vector has a positive correlation with the component *a* and a negative correlation with the component *b*; and the third loading vector has a positive correlation with the component *c* and negative correlation with both components *a* and *b*. Therefore, the third loading vector includes the information of the component *c* and determines the model performance, which is consistent with the above result (the changing relationship between the number of loading vectors and RMSEP values). On the other hand, the first loading vector is dependent to some degree on relative intensities of spectral bands, and the information in the first loading vector is less useful if the component of interest only has relatively small spectral features.

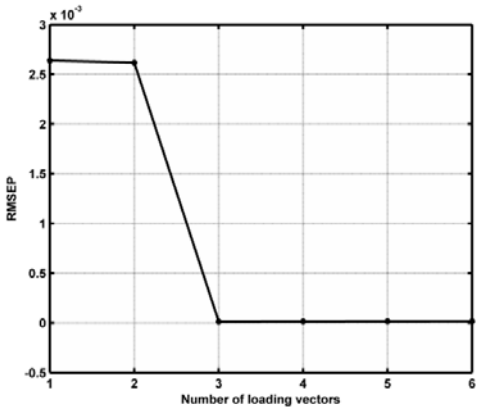


Fig. 3 The changing relationship between RMSEP values and the number of loading vectors for the PLS calibration models based on simulated spectra.

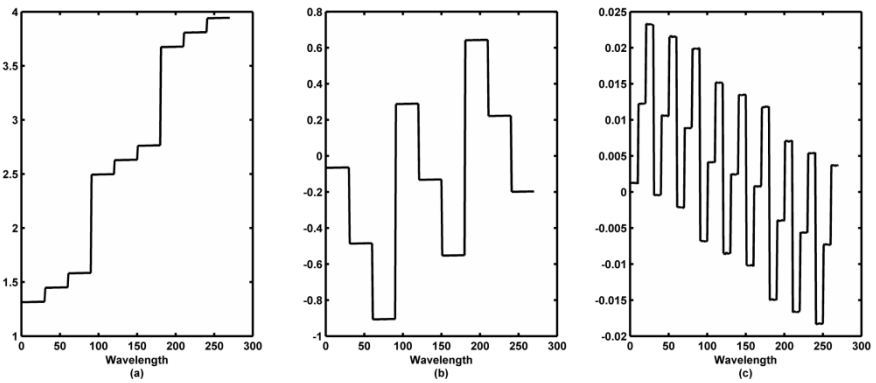


Fig. 4 Three loading vectors used in the PLS calibration model, (a) the first loading vector, (b) the second loading vector, (c) the third loading vector.

The regression coefficients of the PLS calibration model with three loading vectors are shown in Fig. 5. The changing trend of regression coefficients is positively correlated with the absorbance curve of the component *c*, but negatively correlated with components *a* and *b*. Further, we calculated the average regression coefficient in every wavelength segment (the difference between regression coefficients in each wavelength segment was due to the noise), and then sorted 27 average regression coefficients. The number of wavelength segments from short wavelength to long wavelength is 1-27. The number of wavelength segments, the average regression coefficients and the corresponding coding are shown in order in Table 2. According to the sequence and the number of wavelength segments in Table 2, we can clearly understand the relations between regression coefficients and absorbance of multi components. We come to conclusions as follows: (1) when the

absorbance of the component of interest is far greater than the other two components (“>>”), the regression coefficient can reach a positive maximum; (2) when it is equal to the other two (the boxed coding in Table 2) (“=”), the greater the absorbance value is, the greater the regression coefficient is; (3) when it is less than the other two (“<”) or is less than the other one and equal to another one (“<”, “=”), the regression coefficient is negative; (4) when it is far less than the other two (“<<”), the coefficient is a negative minimum. In a word, if the absorbance of the component to be measured is greatly different from other components, the corresponding regression coefficient is absolute maximum.

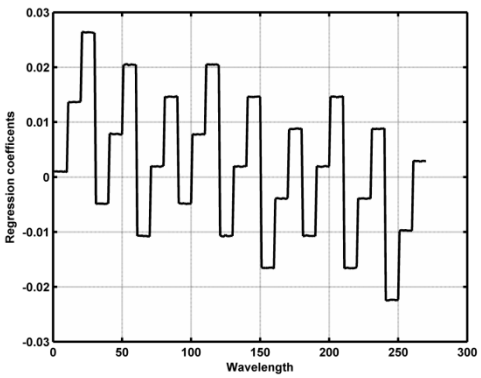


Fig. 5 Regression coefficients of the PLS calibration model with three loading vectors.

There are many reports about wavelength selection based on the regression coefficient vector of PLS.[19,20] The wavelengths with relatively large regression coefficients were selected to build a calibration model. The practical experimental results demonstrated that the method can be used to improve the model performance. Base on existing conclusions and above analysis results, we can infer that the spectra in the wavelength range where the absolute values of regression coefficients are relatively larger (namely the absorbance of the component of interest is greatly different from the others) have more effective information. The analysis result also shows that the coding method can be used to select wavelength variables and further design better wavelength selection methods.

Meanwhile, the analysis process clearly accounts for the action mechanism of PLS regression, including loading vectors and selection of loading vectors. We could use the coding method to study other modeling methods. We believe the new method can help other researchers study the principle of various modeling methods and provide the reference for choosing a better one.

It is worth noting that we didn’t consider nonlinear effects when constructing the coding spectra. In practical situations, there are various nonlinear effects, such as scattering influence from the object itself, temperature changes, and interaction between different components. We can highlight influence of these nonlinear factors in the coding spectra for further study, for example, by adding the square of the coding. Moreover, the discrete steps which don’t conform to the smoothness of the spectra in visible and near-infrared range are the biggest drawback in the coding absorbance curve. They will weaken the relationship between adjacent wavelengths. Therefore, it is very difficult for the coding method to evaluate the performance of some spectral preprocessing methods such as smoothing methods and the first and second order derivatives. Although most of modeling methods such as PLS also neglect the relationship between adjacent wavelengths and take spectra as a matrix, there is no denying that the changing trend of adjacent wavelengths includes a lot of information. Maybe we can change the code expression to simulate the adjacent relationship in the future research.

Table 2 The number of wavelength segments, the average regression coefficients and the corresponding coding.

Sequence	Number of wavelength segment	Average regression coefficient	coding (*a*, *b*, *c*)
1	3	0.0263	(1, 1, 3)
2	6, 12	0.0205	(1, 2, 3) (2, 1, 3)
3	9, 15, 21	0.0146	(1, 3, 3) (2, 2, 3) (3, 1, 3)
4	2	0.0137	(1, 1, 2)
5	18, 24	0.0088	(2, 3, 3) (3, 2, 3)
6	5, 11	0.0078	(1, 2, 2) (2, 1, 2)
7	27	0.0029	(3, 3, 3)
8	8, 14, 20	0.0019	(1, 3, 2) (2, 2, 2) (3, 1, 2)
9	1	0.0010	(1, 1, 1)
10	17, 23	-0.0039	(2, 3, 2) (3, 2, 2)
11	4, 10	-0.0048	(1, 2, 1) (2, 1, 1)
12	26	-0.0098	(3, 3, 2)
13	7, 13, 19	-0.0107	(1, 3, 1) (2, 2, 1) (3, 1, 1)
14	16, 22	-0.0166	(2, 3, 1) (3, 2, 1)
15	25	-0.0224	(3, 3, 1)

Conclusions

The coding method used to analyze the internal mechanism of spectral analysis has been presented in this paper. Simulated spectra of the compound with three components were constructed based on

the coding absorbance curves of three components, and the calibration models were built using PLS regression. The results demonstrate that the new method is effective in qualitative and quantitative analysis of influence from absorbance of three components, and can clearly show the relationships among loading vectors, the number of loading vectors, regression coefficients of PLS and the spectra. The results verify the feasibility of the coding method. And we believe the coding method would also be a promising method in wavelength selection, analysis of other modeling methods and influence from non-linear factors, which are also the focus of our future research.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (Grant No. 61177011, 61377107), the Science and Technology Commission of Shanghai Municipality (Grant No. 14DZ2260800), and the Project supported by the State Key Development Program for Basic Research of China (Grant No. 2011CB932903).

Notes and references

[a] *Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University, Shanghai 200241, China. Fax: +8621-5434-5119; Tel: +8621-5434-5163; E-mail: qlli@cs.ecnu.edu.cn*

[b] *State Key Laboratory of Precision Measurement Technology and Instruments, Tianjin University, Tianjin 300072, China.*

1 Y. He, L. Tang, X. Wu, X. Hou and Y. Lee, *Appl. Spectrosc. Rev.*, 2007, **42**, 119.

2 W. J. Dong, Y. N. Ni and S. Kotot, *Appl. Spectrosc.*, 2014, **68**, 245.

3 D. Sorak, L. Herberholz, S. Lwascek, S. Altinpinar, F. Pfeifer and H. W. Siesler, *Appl. Spectrosc. Rev.*, 2012, **47**, 83.

4 S. Wang, L. Li, L. Zhong and Z. Chen, *Journal of Analytical Science*, 2011, **27**, 779.

5 M. Forina, S. Lanteri, M. C. C. Oliveros and C. P. Millan, *Anal. Bioanal. Chem.*, 2004, **380**, 397.

6 R. M. Balabin and S. V. Smirnov, *Anal. Chim. Acta*, 2011, **692**, 63.

7 R. M. Balabin, R. Z. Safieva and E. I. Lomakina, *Chemometr. Intell. Lab.*, 2007, **88**, 183.

8 X. M. Wu, Z. Q. Liu and H. Li, *Anal. Methods*, 2014, **6**, 4056.

9 P. Geladi, D. MacDougall and H. Martens, *Appl. Spectrosc.*, 1985, **39**, 491.

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

11 J. Jin, Z. Chen, L. Li, R. Steponavicius, S. N. Thennadil, J. Yang and R. Yu, *Anal. Chem.*, 2012, **84**, 320.

12 J. Koljonen, T. E. M. Nordling and J. T. Alander, *J. Near Infrared Spec.*, 2008, **16**, 189.

13 L. Norgaard, A. Saudland, J. Wagner, J. P. Nielsen, L. Munck and S. B. Engelsen, *Appl. Spectrosc.*, 2000, **54**, 413.

14 M. J. McShane, G. L. Cote and C. H. Spiegelman, *Appl. Spectrosc.*, 1998, **52**, 878.

15 D. M. Haaland and E. V. Thomas, *Anal. Chem.*, 1988, **60**, 1193.

16 I. Barman, C. Kong, N. C. Dingari, R. R. Dasari and M. S. Feld, *Anal. Chem.*, 2010, **82**, 9719.

17 N. C. Dingari, I. Barman, G. P. Singh, J. W. Kang and R. R. Dasari and M. S. Feld, *Anal. Bioanal. Chem.*, 2011, 400, 2871.

18 W. J. Zhang, R. Liu, W. Zhang, H. Jia and K. X. Xu, *Biomed. Opt. Express*, 2013, **4**, 789.

19 A. G. Frenich, D. Jouan-Rimbaud, D. L. Massart, S. Kuttatharmmakul, M. M. Galera and J. L. M. Vidal, *Analyst*, 1995, **120**, 2787.

20 M. Hong and Z. Wen, *Spectrosc. Spect. Anal.*, 2010, **30**, 2088.