Available online at www.sciencedirect.com
SCIENCE DIRECT
Talanta 69 (2006) 676–680
Talanta
www.elsevier.com/locate/talanta

A new regression method based on independent component analysis

Xueguang Shao [a,b,*], Wei Wang [a], Zhenyu Hou [a], Wensheng Cai [a,b]

[a] *Department of Chemistry, University of Science and Technology of China, Hefei, Anhui 230026, PR China*
[b] *Department of Chemistry, Nankai University, Tianjin 300071, PR China*

Received 15 June 2005; received in revised form 30 October 2005; accepted 30 October 2005
Available online 1 December 2005

Abstract

Based on independent component analysis (ICA), a new regression method, independent component regression (ICR), was developed to build the model of NIR spectra and the routine components of plant samples. It is found that ICR and principal component regression (PCR) are completely equivalent when they are applied in quantitative prediction. However, independent components (ICs) can give more chemical explanation than principal components (PCs) because independence is a high-order statistic that is a much stronger condition than orthogonality. Three ICs are obtained by ICA from the NIR spectra of plant samples; it is found that they are strongly correlated to the NIR spectra of water, hydrocarbons and organonitrogen compounds, respectively. Therefore, ICA may be a promising tool to retrieve both quantitative and qualitative information from complex chemical data sets.

Keywords: Independent component analysis (ICA); Principal component analysis (PCA); Independent component regression (ICR); Principal component regression (PCR); Near-infrared spectroscopy (NIR)

1. Introduction

Near-infrared spectroscopy (NIR) has been proved to be a very useful tool for both direct and indirect analysis, especially for the complex plant samples, such as petroleum, tobacco, agricultural and food products, etc. [1–4]. The efficient application of quantitative NIR method is strongly dependent on chemometrical methods, including partial least square (PLS) and principal component regression (PCR), since the NIR absorption bands are generally rather weak and mutually influenced, especially in complex plant samples. More and more chemometrical tools are being applied to improve the qualitative analysis and quantitative prediction ability of the NIR technique for interesting chemical components in complex [5–11].

In recent years, independent component analysis (ICA) [12–16] was greatly developed as a potential statistical technique for blind source separation (BSS). Through making full use of the high-order statistical characteristics of the source, i.e., the fourth-order central moment, ICA can effectively resolve the independent components (ICs) from the measured mixed signals without any additional information about the source signals. It had been widely applied in the signal processing fields, such as biomedical signals [17,18], image processing [19,20] and financial analysis [21]. Its applications in processing analytical chemistry signals, including IR [22,23], NIR [24], photoacoustic spectroscopy (PAS) [25], electron paramagnetic resonance (EPR) [26] and GC/MS [27], were also investigated by some researchers in their recent works.

Since ICA is commonly considered to be a further development of principal component analysis (PCA) [13,14], a similar regression method based on ICA, independent component regression (ICR), is proposed by Chen and Wang [24]. In their work, the NIR spectra of water, starch and protein mixtures were investigated. It was found that the qualitative spectral information related to the pure component and quantitative prediction can be obtained. In this work, ICR method is introduced and applied in the quantitative prediction of routine components in tobacco samples from NIR spectra. Furthermore, the similarity and difference between ICR and PCR are discussed in detail. Although the application of ICA is generally thought to limited by its demand of the normal distribution of the source signals, it was found that ICR and PCR are completely equivalent in quantitative prediction because both principal components (PCs) and independent components are latent factors of the NIR spectra.

* Corresponding author. Tel.: +86 22 23503430; fax: +86 22 23502458.
E-mail address: xshao@nankai.edu.cn (X. Shao).

0039-9140/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2005.10.039

However, ICs can give more chemical explanation than PCs because independence is a high-order statistic that is a much stronger condition than orthogonality.

2. Theory and algorithm

2.1. Noise-free ICA model

If noise term is omitted, the basic model of ICA can be written as the following expression:

$$\mathbf{X} = \mathbf{AS} \tag{1}$$

where **X** denotes the recorded NIR spectra matrix and **S** and **A** represent the independent components and the coefficient matrix, i.e., the mixing matrix of the ICs, respectively.

The goal of ICA is to find a proper linear representation of non-Gaussian vectors so that the estimated vectors are as independent as possible, and the mixed signals can be denoted by the linear combinations of these independent components, ICs. The basic noise-free ICA model is very similar to that of PCA, in which mixed signals are denoted by the linear combinations of some orthogonal principal components, PCs. The main difference between ICA and PCA is that they have different criterion in linear representation. The former is finding independent ICs and the latter is finding orthogonal PCs. In probability theory, independence is a high-order statistic and it is a much stronger condition than orthogonality. Therefore, ICA is considered to be more powerful in analyzing multivariate data sets because the high-order statistic of ICs can reflect the intrinsical properties of mixed signals better.

2.2. Independent component regression

Since the mixed signals, **X**, can be described by its coefficient matrix, **A**, the multiple linear regression equation between **A** and chemical components concentration matrix, **C**, can be written as following like PCR:

$$\mathbf{C} = \mathbf{AB} \tag{2}$$

The whole ICR algorithm is very similar to that of PCR. The only difference is to use ICs and their coefficient matrix obtained by ICA in the regression operation, instead of PCs and score matrix obtained by PCA.

The quantitative prediction results are evaluated by three criteria in this work, which are correlation coefficient (R), root mean square error of prediction (RMSEP) and mean relative error (MRE), according to the following formulae:

$$\mathrm{RMSEP} = \left(\frac{\sum_{i=1}^{m} (\hat{y}_i - y_i)^2}{m} \right)^{1/2} \tag{3}$$

$$\mathrm{MRE}\,(\%) = \frac{1}{m} \sum_{i=1}^{m} \frac{|y_i - \hat{y}_i|}{y_i} \times 100 \tag{4}$$

The details of the ICR calculation can be summarized as the following five steps:

(1) Input the preprocessed NIR data set, including the training set (**X**), prediction set ($\mathbf{X}_\mathrm{p}$) and the concentration of the training samples (**C**).
(2) Calculate the ICs (**S**) and the coefficient matrix (**A**) of the training set using the ICA program. The JADE algorithm downloaded from http://www.tsi.enst.fr/~cardoso/RR_ECG/jadeR.m was used. Furthermore, in this step, the number of ICs should be estimated. Generally, the suitable number of ICs can be estimated by the relative root of sum of square differences (RRSSQ) [27] of the original data and the reconstructed data from the ICs. However, in this study, the number of ICs is assumed to be the same as the number of PCs, which was estimated by PCA.
(3) Calculate the regression coefficients **B** in Eq. (2) with $\mathbf{A}^+\mathbf{C}$, where $\mathbf{A}^+$ is the pseudo inverse of **A**.
(4) Calculate the coefficient matrix ($\mathbf{A}_\mathrm{p}$) of the prediction set with $\mathbf{X}_\mathrm{p}\mathbf{S}$, and then predict the concentration of prediction set ($\mathbf{C}_\mathrm{p}$) with $\mathbf{A}_\mathrm{p}\mathbf{B}$.
(5) Finally, calculate the RMSEP and MRE using Eqs. (3) and (4).

3. Experimental and data preprocessing

NIR spectra of 46 tobacco lamina samples were measured on a Bruker Vector 22/N FT–NIR System. Each NIR spectrum was recorded in the wave number range 4000–9000 cm^{-1} with the digitization interval 1 cm^{-1}. The concentration of the four chemical components, including total sugar (TS), total nitrogen (TN), nicotine and water, was measured on an AutoAnalyzer III instrument (Bran and Luebbe, Germany) following the procedures of the standard methods. TS means the total amount of sugar, mainly composed of glucose and levulose, and a small amount of sucrose, maltose and other glucide compounds. TN means the total amount of organonitrogen compounds, mainly composed of proteins and small amount of amino acids and organic amines.

To reduce the influences of noise and background, a continuous wavelet transform (CWT) operation with scale factor 100 was adopted. Then, the standardization operation was applied to make the whole data set to be zero-mean and unit-variance. All the calculations were implemented with our previously developed programs [28–30] on a Pentium IV (3.0 GHz) PC with 512 M memory. Fig. 1 demonstrates the measured NIR spectra and the spectra after preprocessing of 20 randomly selected samples. The whole spectral data set was randomly divided into three parts, i.e., the training set with 26 spectra, the validation set with 10 spectra and the prediction set with 10 spectra.

4. Results and discussion

4.1. Chemical explanation of ICs and PCs

The differences between ICA and PCA have been described in many aspects since ICA was proposed [13,14,31]. In this work, the comparison between PCs and ICs is discussed when they are used to explain the latent factors in NIR spectra of complex plant samples.

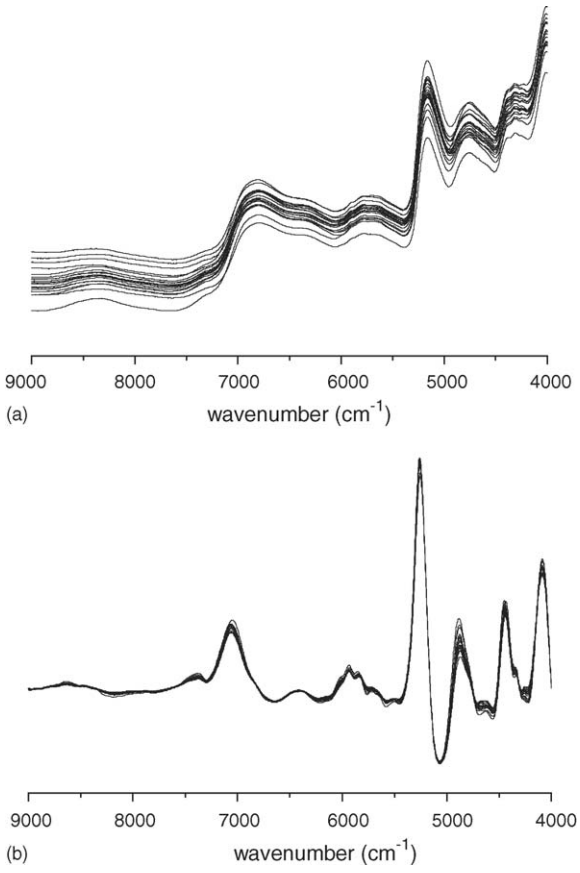

Fig. 1. The measured NIR spectra (a) and the spectra after preprocessing with CWT (b) of 20 randomly selected samples.

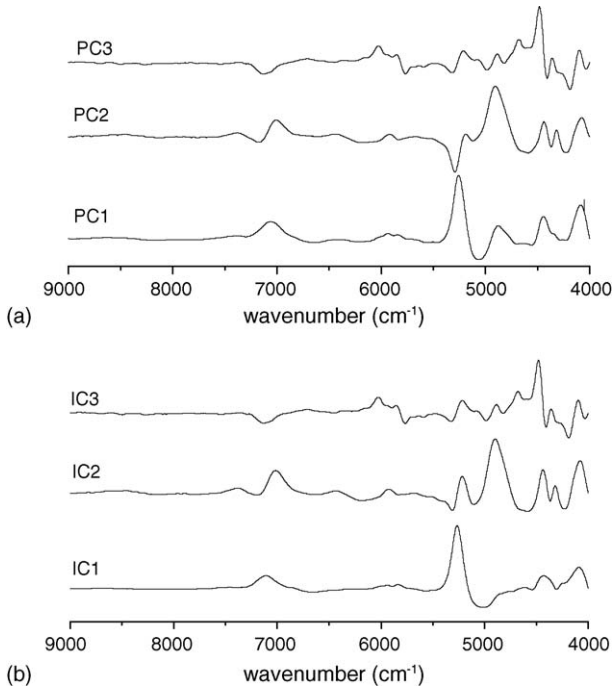

Fig. 2. The three PCs (a) and ICs (b) obtained from the preprocessed NIR signals by PCA and ICA, respectively.

The whole 46 NIR spectra obtained above are tested to compare the ICs and PCs. At first, the number of components is set to 3 because three PCs can contain 99.9% information of the original data in PCA. Fig. 2 shows three PCs and ICs obtained from the NIR data set, respectively. It can be seen in Fig. 2 that PCs, especially the first PC, are still complex and similar to the preprocessed signals in Fig. 1(b). Such results can give little help to explain the latent factors in NIR spectra. However, three ICs can distinguish the six main peaks in Fig. 1(b) clearly. This will offer more information and make the variables explanation easier. Such a conclusion can also be further validated with the correlation between the ICs and the concentrations of the components of the samples. The first three columns of Table 1 give the correlation coefficients, R, between the mixing coefficients $\mathbf{A}$ of ICs and the concentrations of total sugar, total nitrogen, nicotine and water. It can be found that the mixing coefficients of an IC, which describe the contribution of this IC to the spectra of these samples, have a strong correlation with the concentrations of a specific chemical component. Concretely, the concentration of TS is shown to have an obvious positive correlation with the mixing coefficients of IC2, which means that IC2 contains more information of hydrocarbons. In similarity, water is more correlated with IC1 and organonitrogen compounds, including nicotine and TN, are more correlated with IC3.

As a comparison, the columns 4–6 of Table 1 show the correlation coefficients between the coefficients of PCs and these concentrations. It is obvious that, much different from the first IC, the first PC, which is the most important one containing 99.5% information of the data set, has no correlation with any concentration of the four components. The different property

Table 1
The correlation coefficients between concentrations and coefficients of ICs

Number of ICs or PCs	No. of IC or PC	TS	TN	Nicotine	Water
3	IC1	−0.68	0.21	0.14	0.78
	IC2	0.74	−0.29	−0.23	−0.72
	IC3	−0.56	0.82	0.86	−0.52
	PC1	−0.06	0.03	−0.06	0.37
	PC2	0.73	−0.28	−0.21	−0.73
	PC3	−0.56	0.82	0.85	−0.53
10	IC1	−0.66	0.27	0.17	0.80
	IC2	0.66	−0.16	−0.13	−0.76
	IC3	−0.52	0.79	0.92	−0.45
	IC4	−0.04	0.51	0.45	−0.76
	IC5	0.20	−0.65	−0.63	0.66
	IC6	−0.90	0.70	−0.55	0.28
	IC7	0.38	−0.22	−0.24	−0.65
	IC8	0.63	0.29	−0.19	−0.12
	PC9	−0.24	0.27	0.25	−0.28
	PC10	0.10	−0.23	−0.38	−0.31
	PC1	−0.06	0.03	−0.06	0.37
	PC2	0.73	−0.28	−0.21	−0.73
	PC3	−0.56	0.82	0.85	−0.53
	PC4	−0.13	−0.14	−0.27	−0.10
	PC5	0.04	−0.08	−0.07	−0.18
	PC6	0.16	−0.03	0.17	0.11
	PC7	0.04	−0.26	0.19	0.09
	PC8	−0.03	−0.27	−0.09	0.06
	PC9	−0.16	0.02	0.07	0.12
	PC10	0.14	0.08	0.04	0.10

of IC and PC is clearly indicated by the comparison. However, the second and third PCs have correlation coefficients with the corresponding IC. This should be a coincidence that in this case the IC and PC are equivalent to best represent remained spectral information after the first IC and PC. In other words, the remaining data fit in well with both the mathematical conditions of ICA and PCA. After all, the PCA is adopted as the first step in ICA calculation.

To further investigate the difference of ICs and PCs, the correlation of the coefficients of ICs and PCs with the concentrations by using different number of latent variables was calculated. As an example, the results obtained with 10 latent variables are listed in the bottom part of Table 1. By comparison of the correlation coefficients of PCs, it can be found that the first three PCs are completely identical, and from the fourth PC, there is almost no correlation between the coefficients and concentration. It is not difficult to understand such results because PCA finds PCs one by one to represent as much as the information of the processing data set. However, by comparison in the same way, it can be found that different ICs are obtained when different number of ICs is used in the calculation, although there is some similarity for the first three ICs. Furthermore, up to eighth IC there is correlation (positive or negative) between each IC and the concentration of some component. This sufficiently demonstrates the difference of ICs and PCs that PC tries to represent as much as the information but IC tries to get the information as independent as possible.

Theoretically, the rotation immutability of ICA is one possible reason for its ability in variable explanation. Since the goal of ICA is to find statistically independent latent variables, the ICs obtained by ICA can be determined only if the source signals are same, no matter what the mixing coefficient matrix is. In contrast, PCs obtained by PCA are different if the rotation matrix is different. Taking the signals of complex samples as an example, if the components in a mixture are determined, the ICs should be completely identical. However, the PCs will be different if the concentration matrix is different. In this sense, ICs are more appropriate to describe the intrinsical properties of a certain system than PCs. Therefore, ICA may play a more important role in explaining latent factors in NIR spectra than PCA.

4.2. Quantitative prediction by ICR and PCR

To evaluate the quantitative prediction ability of ICR and PCR, the NIR data sets mentioned above are investigated. Table 2 shows the predicted results of the two methods. From the results obtained using three ICs or PCs, it can be seen that ICR and PCR give completely identical results. This phenomenon should be explained by the fact that both the PCs and ICs are latent factors of the NIR spectra, and the PCs and ICs are transformable each other. In fact, in ICA algorithms, PCA is commonly used as a preprocessor to reduce the number of components and simplify the problem. According to the analysis in Section 2, the ICs estimated by ICA are mutually statistically independent, and consequently orthogonal. Independent ICs should be obtained by linear transformation from orthogonal PCs. Therefore, ICs and PCs should contain the same information about the mixed signals. Consequently, ICA and PCA will give the same prediction results when they are used as regression method.

Table 2
The quantitative prediction results of ICR and PCR

Number of ICs or PCs	Component	*R*		RMSEP		RME (%)	
		ICR	PCR	ICR	PCR	ICR	PCR
3	TS	0.949	0.949	1.937	1.937	7.23	7.23
	TN	0.944	0.944	0.111	0.111	3.97	3.97
	Nicotine	0.930	0.930	0.336	0.336	12.9	12.9
	Water	0.973	0.973	0.327	0.327	2.20	2.20
10	TS	0.950	0.950	1.549	1.549	6.46	6.46
	TN	0.948	0.948	0.119	0.119	4.68	4.68
	Nicotine	0.927	0.927	0.272	0.272	9.50	9.50
	Water	0.934	0.934	0.364	0.364	2.33	2.33

To further investigate the identity between ICs and PCs in quantitative prediction, results obtained by using different number of latent variables (ICs or PCs) are calculated. It was found that the results by ICs and PCs are always identical no matter how many latent variables are used. As an example, the bottom part of Table 2 shows the results using 10 latent variables. On the other hand, by comparison of the results with 3 and 10 latent variables, it can be found that there is no significant change in all the three items (*R*, RMSEP and RME). This also indicates that three latent variables are enough to represent the information of NIR spectra.

5. Conclusion

Independent component regression was developed to build the model of NIR spectra and the routine components of plant samples. Compared with PCR, quantitative prediction ability is found to be completely equivalent because both the PCs and ICs are latent factors of the NIR spectra. However, ICA is more powerful in qualitatively describing the intrinsical properties of the NIR spectra because ICs are obtained under a high-order statistic, i.e., independence, that is a much stronger condition than orthogonality. With NIR spectra of plant samples, it was proved that ICs obtained by ICA are strongly correlated to the specific component of the sample. Therefore, ICA may be a promising tool to retrieve both quantitative and qualitative information from complex chemical data sets.

Acknowledgements

This work is supported by the outstanding youth fund (No. 20325517) from the National Natural Science Foundation of China (NNSFC), and the Teaching and Research Award Program for Outstanding Young Teachers (TRAPOYT) in Higher Educations of MOE, PR China.

References

[1] A.A. Christy, S. Kasemsumran, Y.P. Du, Y. Ozaki, Anal. Sci. 20 (2004) 935.

[2] A.M. Bruno-Soares, I. Murray, R.M. Paterson, J.M.F. Abreu, Anim. Feed Sci. Technol. 75 (1998) 15.
[3] J.T. Diffee, Pract. Spectrosc. 13 (1992) 433.
[4] M. Blanco, S. Maspoch, I. Villarroya, X. Peralta, J.M. Gonzalez, J. Torres, Appl. Spectrosc. 55 (2001) 834.
[5] O. Svensson, T. Kourti, J.F. MacGregor, J. Chemom. 16 (2002) 176.
[6] T.V. Karstang, O.M. Kvalheim, Anal. Chem. 63 (1991) 767.
[7] R.P. Paradkar, R.R. Williams, Appl. Spectrosc. 51 (1997) 92.
[8] P. Chalus, Y. Roggo, S. Walter, M. Ulmschneider, Talanta 66 (2005) 1294.
[9] X.J. Cui, Z.Y. Zhang, W.L. Ren, S.D. Liu, P.D. Harrington, Talanta 64 (2004) 943.
[10] D. Chen, X.G. Shao, B. Hu, Q.D. Su, Anal. Chim. Acta 511 (2004) 37.
[11] X.G. Shao, F. Wang, D. Chen, Q.D. Su, Anal. Bioanal. Chem. 378 (2004) 1382.
[12] A. Hyvarinen, J. Karhunen, E. Oja, Independent Component Analysis, John Wiley & Sons, 2001.
[13] C. Jutten, J. Herault, Signal Process. 24 (1991) 1.
[14] P. Comon, Signal Process. 36 (1994) 287.
[15] A. Hyvärinen, E. Oja, Neural Networks 13 (2000) 411.
[16] L. De Lathauwer, B. De Moor, J. Vanderwalle, J. Chemom. 14 (2000) 123.
[17] R.N. Vigário, Electroencephalogr. Clin. Neurophysiol. 103 (1997) 395.
[18] J.V. Stone, Trends Cogn. Sci. 6 (2002) 59.
[19] S. Makeig, M. Westerfield, T.P. Jung, S. Enghoff, J. Townsend, E. Courchesne, T.J. Sejnowski, Science 295 (2002) 690.
[20] A. Hyvärinen, Neural Comput. 11 (1999) 1739.
[21] A.D. Back, A.S. Weigend, Int. J. Neural Syst. 8 (1997) 473.
[22] E. Visser, T.W. Lee, Chemom. Intell. Lab. Syst. 70 (2004) 147.
[23] X. Bi, T.H. Li, L. Wu, Chem. J. Chin. Univ. 25 (2004) 1023.
[24] J. Chen, X.Z. Wang, J. Chem. Inf. Comput. Sci. 41 (2001) 992.
[25] A. Pichler, M.G. Sowa, J. Mol. Spectrosc. 229 (2005) 231.
[26] J.Y. Ren, C.Q. Chang, P.C.W. Fung, J.G. Shen, F.H.Y. Chan, J. Magn. Reson. 166 (2004) 82.
[27] X.G. Shao, G.Q. Wang, S.F. Wang, Q.D. Su, Anal. Chem. 76 (2004) 5143.
[28] X.G. Shao, Y.D. Zhuang, Anal. Sci. 20 (2004) 451.
[29] C.X. Ma, X.G. Shao, J. Chem. Inf. Comput. Sci. 44 (2004) 907.
[30] X.G. Shao, A.K.M. Leung, F.T. Chau, Acc. Chem. Res. 36 (2003) 276.
[31] M.S. Bartlett, J.R. Movellan, T.J. Sejnowski, IEEE Trans. Neural Networks 13 (2002) 1450.