ISSN 1009-3095 Journal of Zhejiang University(SCIENCE) V.1,No.1,P.15-19,Jan.-Mar.,2000
http://www.zju.edu.cn; http://www.chinainfo.gov.cn/periodical

A NEW COMPUTING MULTIVARIATE SPECTRAL ANALYSIS METHOD BASED ON WAVELET TRANSFORM*

CHENG Yi-yu(程翼宇), CHEN Min-jun(陈闽军)

(*Dept. of Chemical Engineering, Yuquan Campus of Zhejiang University, Hangzhou, 310027, China*)

Received Dec.8, 1998; revision accepted May 20,1999

Abstract: This paper proposes a new algorithm for multivariate calibration named Principal Component Regression Based on Wavelet (PCRW) which combines wavelet decomposition technique with the factor analysis method for establishing a duplicate denoising mechanism. A practical example in spectral analysis of a typical multicomponent pharmaceutical system was used to verify the effectiveness of the algorithm. It was shown that PCRW produced fewer prediction errors than those obtained by using PCR.

Key words: spectral analysis, wavelet transform, multivariate calibration, chemometrics, PCRW
Document code: A **CLC number**: O652.9, O657

INTRODUCTION

Spectroscopic methods are increasingly being employed for quantitative application in chemistry, biology and medicine. While advances in instrumentation bring increased resolution and sensitivity, the multivariate calibration models as statistical tools have also been shown to improve analytical precision, accuracy, reliability and applicability of spectral analysis vis-a-vis the traditional univariate methods of data analysis (Henk et al., 1992).

Of available methods for analyzing multivariate spectral data, factor-based methods such as principal component regression (PCR) and partial least-squares (PLS) regression had been well-studied (Geladi et al, 1986; Halland et al., 1988). In these factor-based methods, the measurement matrix is decomposed into a set of orthogonal variables, with each of them being a linear combination of the measured spectral data. Some informative variables corresponding to the larger eigenvalues are used to regress the concentration value while the other smaller ones are discarded as noise. Because noise is distributed at random in each variable, some noises still existing in the retained variables after the above data processing procedure handicap the performance of the factor-based methods. Therefore a duplicate denoising mechanism is expected to improve the final calibration results. Unfortunately, it is known that noise cannot be removed further by a duplicate processing of fac-tor analysis(Pan et al., 1992).

Wavelet analysis, a new mathematics branch developed in recent years, is a perfect combination of harmonic analysis, function analysis, Fourier analysis and numerical analysis (Lu et al., 1996). As a new and more efficient method that can provide the information in local time and frequency scales, it had been applied to some areas of analytical chemistry, such as detecting peaks in flow injection analysis (Bos et al., 1992), compressing IR spectra data(Bos et al., 1994) and denoising in electro-analytical chemistry(Bao et al., 1997).

This paper presents a new algorithm for multivariate calibration: Principal Component Regression Based on Wavelet (PCRW). The data processing procedure of the presented algorithm, differing from that of the conventional regression methods only in spectral domain, has been extended to the wavelet-based space. The algorithm combines wavelet decomposition technique with the factor analysis method to establish an effective duplicate denoising mechanism. Use of an experimental dataset obtained from UV/Vis spectra to test its performance showed that compared to PCR, PCRW generates fewer average prediction errors, and so, has better performance.

* Project (39870940) supported by NSFC.

THEORY AND ALGORITHM

The multiresolution decomposition of a signal on an orthogonormal wavelet basis gives an intermediate representation between Fourier and spatial representations(Mallat , 1989). The wavelet transform of $f(t)$ is defined as follows

$$W_f(b,a) = |a|^{-1/2}\int_{-\infty}^{+\infty} h\left(\frac{t-b}{a}\right)f(t)\,dt,$$

$$a \neq 0 \qquad (1)$$

Here h is a mother wavelet. The discrete wavelet transform is defined as follows:

$$C_f(m,n) = \int_{-\infty}^{+\infty} h_{m,n}(t)f(t)\,dt \qquad (2)$$

Since $h_{m,n}(t)$ has no explicit expression, the computation of the discrete wavelet transform is difficult. For solving the problem, Mallat presented a pyramidal algorithm, based on which the multiresoltuion decomposition of a signal can be expressed as follows:

$$A_1^d f = D_{2^{-1}}f + D_{2^{-2}}f + \cdots + D_{2^{-J}}f + A_{2^{-J}}^d f \qquad (3)$$

Where $A_1^d f$ is the original signal. $D_{2^j}f$ represents the detail of the signal at resolution $2^{-j}(-J \leq j \leq -1)$, which gives the difference between approximations of the signal at resolutions 2^{-j} and 2^{-j+1}. $A_{2^{-j}}^d f$ represents the approximation of the signal at resolution 2^{-j} (Xie et al., 1997). From Eq. 3, we have

$$\begin{cases} A_{2^j}^d f = \sum_{k=-\infty}^{+\infty} h(2n-k)A_{2^{j+1}}^d f \\ D_{2^j}f = \sum_{-\infty}^{+\infty} g(2n-k)A_{2^{j+1}}^d f \end{cases} \quad -J \leq j \leq -1 \qquad (4)$$

Here, the coefficients of h and g are determined by the selected wavelet function.

Usually, a signal from analytical instruments is contaminated by "noise": random perturbation of the sought signal. The complete separation of the information component and the noise of a signal is not easy in the signal domain. In other words, factor-based methods are limited in their effectiveness in cleanly filtering out noise. If a signal is transformed into the wavelet-based space, the information component and noise of the signal would have different distribution in the wavelet-based space.

From the mathematical point of view, a true signal variable is differentiable but a noise variable is not. So, with the resolution of wavelet transform becoming lower, the modulus maxima of the noise variable decrease, while the modulus maxima of the true signal variable increase (Mallat et al., 1992). Here, the modulus maximum is a measure of the signal energy. As a consequence, at the lower resolution, $A_{2^{-j}}^d f$ represents the information component of signals, and $D_{2^{-j}}f$ mainly represents noise. It can be used for separating the information component and the noise of a signal.

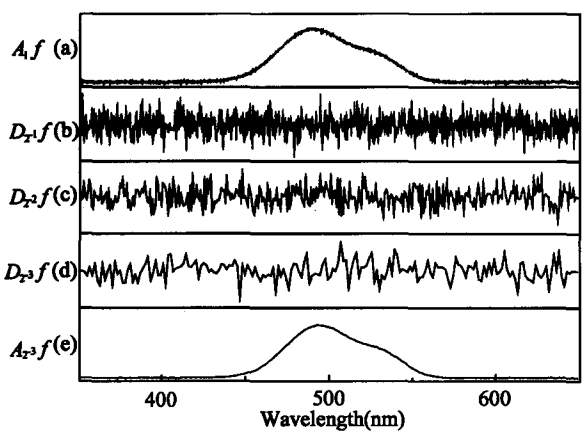

Fig. 1 Wavelet multiresolution decomposition of a typical simulated signal

(a) a typical simulated signal; (b) ~ (d) the detail of the signal at resolution 2^{-1}, 2^{-2} and 2^{-3}; (e) The approximation of the simulated signal at resolution 2^{-3}.

The above characteristic of the true signal and noise in the wavelet-based space is shown in Fig. 1. Assuming two overlapped Gaussian peaks as a true signal, the simulated signal shown in Fig. 1(a) is generated by the two Gaussian peaks and a white noise as follows:

$$a(i) = \exp[-(i-490)^2/800)] + 0.4\exp[-(i-530)^2/400)] + 0.01\mathrm{rand}[n(i)] \qquad (5)$$

Here, rand $[n(i)]$ is a normally distributed random numbers generator in Matlab, i is the wavelength, ranging from 350 nm to 650 nm at 0.2 nm intervals. It is seen in Fig. 1(b ~ d) that with the resolution of wavelet transform becoming lower and lower, the noise is separated

out step by step. Fig. 1(e) shows the approximation of the simulated signal at resolution 2^{-3} is almost the same as the true signal. The calculated results showed that its signal-to-noise ratio is a factor of 9 larger than that of the simulated signal. Thus, the wavelet multiresolution decomposition can be used for denoising effectively.

The PCRW algorithm uses the above denoising principle to improve the PCR method. Combining wavelet transform with factor analysis, PCRW develops a new duplicate denoising mechanism which is more effective than the denoising mechanism of PCR. Measured spectral data with noise are first filtered by using a specially designed wavelet decomposition algorithm, and then are processed with the PCR algorithm. The computing procedure of the PCRW algorithm is as follows:

CALIBRATION PROCEDURE

Step 1: Have the measurement matrix $\boldsymbol{X}(n$ samples by m wavelengths) and the concentration matrix $\boldsymbol{C}(n$ samples by k components) normalized.

Step 2: Give the coefficients h and g corresponding to the given filter, and the cutoff resolution 2^{-s}.

Step 3: Use Eq. 4 to transform all the raw vectors $\boldsymbol{x}^0$ of the normalized matrix $\boldsymbol{X}$ into the wavelet-based space with the given filter and cutoff resolution.

Step 4: Retain all the approximations $\boldsymbol{x}^s(1 \times p)$ to construct a matrix $\boldsymbol{Y}(n \times p)$.

Step 5: Decompose the matrix $\boldsymbol{Y}$ by using following SVD algorithm

$$\boldsymbol{Y}_{n\times p} = \boldsymbol{U}_{n\times n}\boldsymbol{S}_{n\times p}\boldsymbol{V}^t_{p\times p}$$

Here, $\boldsymbol{S}$ is a diagonal matrix, and the eigenvalues of $\boldsymbol{Y}$ are distributed in descending order along its diagonal.

Step 6: Determine the number of optimum latent factors (q) by using some criteria such as REV, IND, IE, etc.

Step 7: Truncate the matrix of $\boldsymbol{U}, \boldsymbol{S}, \boldsymbol{V}$ and decompose the matrix $\boldsymbol{Y}$ into the product of a scores matrix $\tilde{\boldsymbol{T}}(n \times q)$ and a loadings matrix $\tilde{\boldsymbol{V}}$ $(p \times q)$ plus a residual matrix $\boldsymbol{E}(n \times p)$. Here and elsewhere, the tilde ("~") is used to distinguish the truncated matrix.

$$\tilde{\boldsymbol{T}}_{n\times q} = \tilde{\boldsymbol{U}}_{n\times q}\tilde{\boldsymbol{S}}_{q\times q}$$
$$\boldsymbol{Y}_{n\times p} = \tilde{\boldsymbol{T}}_{n\times q}\tilde{\boldsymbol{V}}^t_{p\times q} + \boldsymbol{E}_{n\times p}$$

Step 8: Find the regression coefficient matrix $\boldsymbol{B}$.

$$\boldsymbol{B} = (\tilde{\boldsymbol{T}}^t \times \tilde{\boldsymbol{T}})^{-1} \times \tilde{\boldsymbol{T}}^t \times \boldsymbol{C}$$

PREDICTION PROCEDURE

Step 1: Have the prediction measurement matrix $\boldsymbol{X}_{un}$ normalized as in step 1 above.

Step 2: Transform the normalized matrix $\boldsymbol{X}_{un}$ according to the above procedure from step 2 to step 4 and obtain a matrix $\boldsymbol{Y}_{un}$.

Step 3: construct the normalized matrix $\boldsymbol{C}_{un}$ below

$$\boldsymbol{C}_{un} = \boldsymbol{Y}_{un} \times \tilde{\boldsymbol{V}} \times \boldsymbol{B}$$

Step 4: Recover $\boldsymbol{C}_{un}$ by the inverse calculation of normalization to obtain the predicted concentration matrix.

EXPERIMENTAL SECTION

1. Apparatus and reagent

Chloramphenicolum, Dexamethasoni Acetas, Aethylparabenum, ethanol and sodium hydroxide were all analytical reagent grade; UV/Vis – 7530G spectrophotometer (HP Shanghai Analytical Instrument Limited Company, China); All the computations here were performed on a pentium/133 computer using Matlab 5.1 (the MathWorks, Natick, MA) with our own written programs.

2. Experiment procedure

The experimental dataset was obtained through a carefully designed experiment involving three-component mixtures of Chloramphenicolum, Dexamethasoni Acetas and Aethylparabenum. Stock solutions of the pure components were prepared at concentration of 12.00, 7.50 and 5.00 mg/L for them respectively. Fourteen samples were prepared by combining the various stock solutions and diluting them to 25 ml with 0.1 mol/L sodium hydroxide. Calibration and prediction data sets consisted of 8 and 6 spectra of the samples, respectively. Final concentration ranges were 0.74 – 2.00 mg/L for Chloramphen-

icolum, 0.92 – 1.50 mg/L for Dexamethasoni Acetas and 0.37 – 1.00 mg/L for Aethylparabenum. To minimize the effect of instrument drift, a reference spectrum was run prior to calibration of each new sample. All the spectral data were recorded over the range of 210 – 310 nm at 1 nm intervals on a UV/Vis – 7530G spectrophotometer.

RESULATS AND DISCUSSIOIN

1. Determination of the wavelet filter and cutoff resolution

For the wavelet multiresolution decomposition, its denoising effectiveness mainly depends on the wavelet function and cutoff resolution. In general, the rule for selecting the wavelet function is that it must have not only support property but also smoothing property. Here, filter 5 in the Daubechies wavelet family was found to be the best by experiments. On the other hand, an overly low cutoff resolution may cause partial loss of the information component. After comparing the prediction results with the calibration concentration data set, the cutoff resolution was determined at resolution 2^{-3}.

2. Determination of the optimum number of PCs

Here, the criteria IE, IND, REV, ER and VPVRS(Pan et al., 1992) were used to determine the optimum number of principal components (PCs). The results of the factor analysis in the wavelet-based space listed in Table 1 showed that, the optimum number of PCs was four based on the calculation of the criteria IE, REV, ER and VPVRS. In fact, there were only three components in the experimental samples. Therefore, the 4th principal component should not be considered as a real component. Further investigation showed that there was no good linear addition relationship between the components. The interaction between some components in these mixtures yields the 4th principal component.

Table 1 Results of the factor analysis in wavelet-based space

No.	EV[a] ($\times 10^2$)	IE ($\times 10^3$)	IND ($\times 10^3$)	REV ($\times 10^3$)	ER	VPVRS
1	136124	427.73	0.854	2181	214.6	297.9
2	634.13	50.384	0.554	11.24	3.536	3.730
3	179.34	33.762	0.347	3.544	3.124	2.131
4	57.409	20.282	0.035	1.276	296.9	499.3
5	0.1933	1.8630	0.031	0.005	2.456	2.231
6	0.0787	1.4530	0.030	0.002	2.878	6.490
7	0.0274	1.1810	0.036	0.001	1.407	
8	0.0194			0.001		

[a] eigenvalue

3. Comparison between results obtained by PCRW and PCR

The prediction errors of different algorithms can be evaluated by the value of the root-mean square error of prediction (RMSEP)(Xie et al., 1997) usually used as a criterion to compare the performance of multivariate calibration algorithms. The smaller the value of RMSEP is, the better the algorithm. The RMSEP given by the following expression

$$\mathrm{RMSEP} = \sqrt{\sum_{i=1}^{N_{\mathrm{pred}}} (c_i^{\mathrm{pred}} - c_i^{\mathrm{true}})^2 / N_{\mathrm{pred}}} \qquad (6)$$

Here, c_i^{pred} and c_i^{true} represent the predicted and actual concentrations of an analyte in the predicted sample i respectively, and N_{pred} is the number of predicted samples. After the RMSEPs of the three analytes are calculated respectively, the total RMSEP can be obtained as follows

$$\mathrm{RMSEP}_{\mathrm{total}} = \sqrt{(\mathrm{RMSEP}_{\mathrm{Ch}}^2 + \mathrm{RMSEP}_{\mathrm{De}}^2 + \mathrm{RMSEP}_{\mathrm{Ae}}^2)/3} \qquad (7)$$

The values of RMSEP obtained by using PCR and PCRW are listed in Table 2. showing that

Table 2 Comparison of the RMSEP values calculated by PCR and PCRW

Component	RMSEP (mg/L)		
	PCR	PCRW	PCR-PCRW reduction(%)[a]
Choloramphenicolum	0.014	0.012	14
Dexamethasoni Acetas	0.054	0.022	59
Aethylparabenum	0.040	0.019	53
Total	0.040	0.018	47

[a] The column shows the percent reduction in switching from PCR to PCRW

theprediction errors of PCRW are much lower than those of PCR. This means the performance of PCRW is much better.

Furthermore, the predicted concentrations of the three components are given in Fig. 2. showing that PCRW substantially improves the predictions of all three components in the samples.

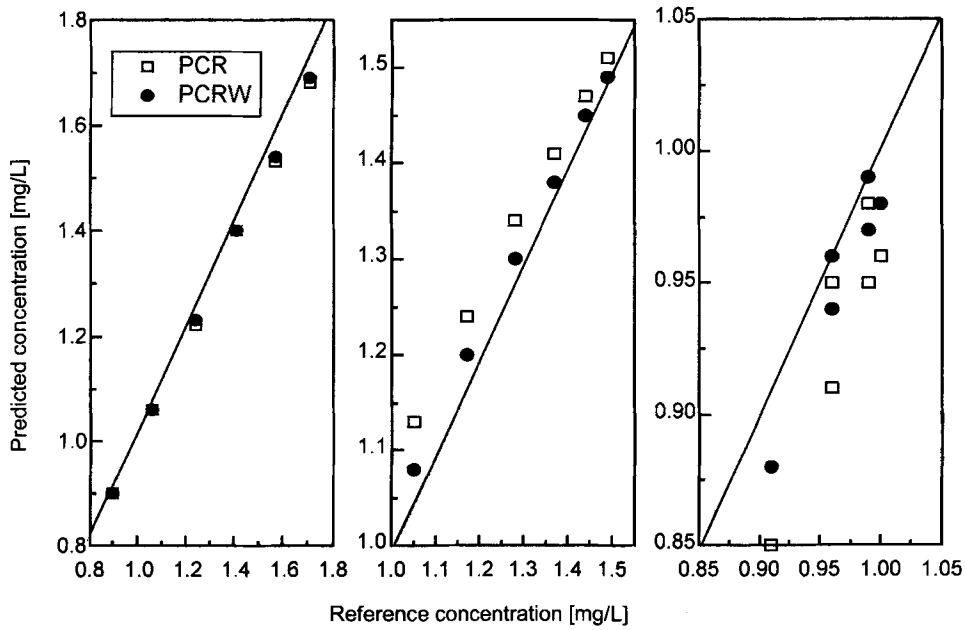

Fig. 2 Prediction concentration plot with PCR and PCRW
(a) Chloramphenicolum; (b) Dexamethasoni Acetas; (c) Aethylparabenum
(The diagonal represents the perfect prediction)

CONCLUSION

The new algorithm for multivariate calibration named Principal Component Regression Based on Wavelet effectively reduces prediction errors. The algorithm was applied to the practical spectral analysis of a typical multicomponent pharmaceutical system. The results showed that PCRW is a better computing multivariate spectral analysis method than the factor-based methods such as PCR.

References

Bao Lunjun, Mo Jinyuan and Jang Zuying, 1997. The application in processing analytical chemistry signal of a cardinal spline approach to wavelets. *Anal. Chem.* **69**: 3053

Bos, M. and Hoogendam, E., 1992. Wavelet transform for the evaluation of peak intensities in flow-injection analysis. *Anal. Chim.* Acta, **267**:73.

Bos, M. and Vrielink, J. A. M., 1994. The wavelet transform for pro-processing IR spectra in the identification of mono and di-Substituted benzenes. *Chemom. & Intell. Lab. Syst.*, **23**:115.

Geladi, P., Kowalski, B. R., 1986. Partial lest-squares regression: A tutorial. *Anal. Chim. Acta*, **185**:1.

Halland, D. M. and Thomas, E. V., 1988. Partial least-squares methods for spectral analysis. *Anal. Chem.*, **60**:1193.

Henk, L. C. and Thomas, L., 1992. Computer-Enhanced analytical spectroscopy, Volume III. Plenum Press, New York and London, p.1 - 2.

Lu Xiaoquan, Mo Jingyuan, 1996. Wavelet analysis as a new method in analytical chemometrics. *Chinese J. of Anal. Chem.*, **24**: 1100(in Chinese with English abstract).

Mallat, S., 1989. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. Pattern Anal. Machine Intell.*, **11**: 674.

Mallat, S. and Hwang, W. L., 1992. Singularities detection and processing with wavelets. *IEEE Trans. on Inform. Theory*, **38**:617.

Pan Zhongxiao, Si Shengzhu, Nie Shengzhe et al., 1992. Factor analysis in chemistry. Chinese Science and Technology Unversity Press, Hefei, p.184(in Chinese).

Yulong Xie, Kalivas, J. H., 1997. Loal prediction models by principal component regression. *Anal. Chim. Acta.*, **348**:29.