Journal of Zhejiang University SCIENCE
ISSN 1009-3095
http://www.zju.edu.cn/jzus
E-mail: jzus@zju.edu.cn

Background correction in near-infrared spectra of plant extracts by orthogonal signal correction*

QU Hai-bin (瞿海斌), OU Dan-lin (欧丹林), CHENG Yi-yu (程翼宇)†‡

(*Pharmaceutical Informatics Institute, Zhejiang University, Hangzhou 310027, China*)

†E-mail: chengyy@zju.edu.cn

Received Jan. 17, 2005; revision accepted Apr. 30, 2005

Abstract: In near-infrared (NIR) analysis of plant extracts, excessive background often exists in near-infrared spectra. The detection of active constituents is difficult because of excessive background, and correction of this problem remains difficult. In this work, the orthogonal signal correction (OSC) method was used to correct excessive background. The method was also compared with several classical background correction methods, such as offset correction, multiplicative scatter correction (MSC), standard normal variate (SNV) transformation, de-trending (DT), first derivative, second derivative and wavelet methods. A simulated dataset and a real NIR spectral dataset were used to test the efficiency of different background correction methods. The results showed that OSC is the only effective method for correcting excessive background.

Key words: Background correction, Plant extracts, Orthogonal signal correction, Near-infrared spectroscopy
doi:10.1631/jzus.2005.B0838 **Document code:** A **CLC number:** TQ461

INTRODUCTION

When using near-infrared (NIR) spectroscopy to analyze plant extracts, excessive background often exists within the NIR spectra. It is mainly caused by the strong absorbance of the solvent, such as water and ethanol. In addition, the absorbance of the active constituents in plant extracts is often weak and disguised by the excessive background. It is necessary to correct the excessive background, but how to achieve this goal is still troublesome.

The classical method to subtract background is to fit the background as a line or a polynomial curve. Offset correction is an old method applied to correct flat background (Candolfi *et al*., 1999). Multiplicative scatter correction (MSC) (Geladi *et al*., 1985) and standard normal variate (SNV) transformation (Barnes *et al*., 1989) are proposed as methods for correcting multiplicative scatter effect in NIR spectra, but they can also be used to correct sloping background. De-trending (DT) (Karstang and Kvalheim, 1991) is a method to correct curvilinear background by modelling the background as a function of wavelengths with a second-degree polynomial. But in many cases, the background is not as "ideal" as a line or a polynomial curve, so the above methods are inefficient.

Another very common method for background correction is mathematical derivative of the spectrum (Tahboub and Pardue, 1985) usually first derivative and second derivative. A flat background can be removed by first derivative, and a sloping background can be removed by second derivative. In fact, derivative can be seen as removing low-frequency components and amplifying high-frequency components of the spectrum. By selecting proper order of derivative, the background can be removed as low-frequency components, but meanwhile the noise is amplified. However, wavelet method is a more efficient method than derivative method. The spectrum can be decomposed into different frequency components by

‡ Corresponding author
* Project supported by the Zhejiang Province Key Technologies R & D Program (No. 021103549) and the National Key Technologies R & D Program (No. 2001BA701A45), China

wavelet functions, and the background can be removed by setting the very-low-frequency components to zero values (Alsberg *et al*., 1997), whereas the noise is not amplified and noise can be removed as very high frequency components (Barclay and Bonner, 1997). Background correction based on wavelet had been reported. Ma and Zhang (2003) applied wavelet transform to background correction in inductively coupled plasma atomic emission spectrometry. Tan and Brown (2002) developed a criterion based on Shannon entropy to judge and remove the low-frequency background reasonably and automatically. Mittermayr *et al*.(2001) removed the highly varying background successfully by wavelet vanishing moments. But when excessive background existed, the frequency contributions of analytical signal and background can hardly be discriminated, and then background cannot be separated from spectrum by derivative or wavelet method.

Some methods based on multivariate calibration had been proposed for background correction. Schechter (1995) proposed a method for correcting nonlinear fluctuating background by using two reference spectra with known concentrations. But it can only be used in monovariate analytical systems. Sun (1997) proposed a PCA (Principal Component Analysis)-like method to remove the excessive background from a set of NIR spectra of human blood by assuming some of the principal components represented the background variations. A specially designed experiment was used to identify principal components whether they represented the excessive background or not, although the principal components representing background often contain the information of interest components and could not be removed. In practice, the complexity and the need of some prior knowledge restrict the application of these methods.

Unfortunately, it seems that none of the classical methods are suitable for correcting excessive background. A new popular method called orthogonal signal correction (OSC) showing potential for correcting excessive background was introduced by Wold *et al*.(1998) for removing synthetical noise including background, scatter effect, etc. It works by removing the parts linearly unrelated (orthogonal) to the response of the calibration model. As a preprocessing method, OSC is widely and successfully applied to spectral analysis of NIR (Sjoblom *et al*., 1998), NMR (Brindle *et al*., 2003), and DRIFTS (Peussa *et al*., 2000).

In this paper, OSC and other common background correction methods were used to correct the excessive background. A simulated dataset and a real NIR spectral dataset were analyzed by different background correction methods, whose results showed that the OSC method is the only effective method for correcting excessive background.

THEORY

The basic idea of OSC is very simple and can be understood easily. When pretreating the NIR spectra in multivariate analysis, we simply want to remove from the spectral matrix ($\boldsymbol{X}$) only the part that is unrelated to the response matrix ($\boldsymbol{Y}$), namely the part that is orthogonal (or as close as possible orthogonal) to the response matrix ($\boldsymbol{Y}$). To achieve such orthogonality, the matrix $\boldsymbol{X}$ can be decomposed into score matrix $\boldsymbol{T}$, which is required to be orthogonal to $\boldsymbol{Y}$, and loading matrix $\boldsymbol{P}$.

To obtain the OSC score matrix $\boldsymbol{T}$, different algorithms have been developed. (Wold *et al*., 1998, Sjoblom *et al*., 1998, Fearn, 2000), direct orthogonalization (DO) (Andersson, 1999), direct OSC (DOSC) (Westerhuis *et al*., 2001), orthogonal projection to latent structures (OPLS) (Trygg and Wold, 2002). A previous study of ours showed that the results obtained by different OSC algorithms were similar, and that the best were those of Sjoblom *et al*.(1998). So the Sjoblom's OSC algorithm is used, which is described in the following text.

For distinguishing ability, the final score vector, loading vector and weight vector of OSC are represented by $\boldsymbol{t}_\perp$, $\boldsymbol{p}_\perp$ and $\boldsymbol{w}_\perp$.

Before the calculations, $\boldsymbol{X}$ and $\boldsymbol{Y}$ are centered and scaled as usual.

In the first step, the first principal component of $\boldsymbol{X}$ is calculated as the initial score vector, $\boldsymbol{t}$. This ensures that $\boldsymbol{t}$ is an optimal linear summary of $\boldsymbol{X}$.

Then orthogonalize $\boldsymbol{t}$ to $\boldsymbol{Y}$

$$\boldsymbol{t}_{\text{new}}=(1-\boldsymbol{Y}(\boldsymbol{Y}^{\text{T}}\boldsymbol{t})^{-1}\boldsymbol{Y}^{\text{T}})\boldsymbol{t} \quad (1)$$

Next, calculate a new weight vector $\boldsymbol{w}$ by

$$\boldsymbol{w}=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{t}_{\mathrm{new}}/(\boldsymbol{t}_{\mathrm{new}}^{\mathrm{T}}\boldsymbol{t}_{\mathrm{new}}) \tag{2}$$

And scale $\boldsymbol{w}$ to unit length

$$\boldsymbol{w}=\boldsymbol{w}/\|\boldsymbol{w}\| \tag{3}$$

Now calculate the new score vector $\boldsymbol{t}$

$$\boldsymbol{t}=\boldsymbol{X}\boldsymbol{w} \tag{4}$$

Repeat Eqs.(1)~(4) until $\boldsymbol{t}$ becomes stable. The final $\boldsymbol{t}$ is now a good describer of the part of $\boldsymbol{X}$ that is orthogonal to $\boldsymbol{Y}$.

The next step is to build a PLS (partial least square) model with $\boldsymbol{X}$ calibrated against $\boldsymbol{t}$. Proper PLS-components are selected to ensure that $\boldsymbol{t}$ is described well by the model. Calculate the prediction vector $\hat{\boldsymbol{b}}$ of the PLS model as the weight vector $\boldsymbol{w}_{\perp}$ of OSC,

$$\boldsymbol{w}_{\perp}=\hat{\boldsymbol{b}}=\boldsymbol{W}(\boldsymbol{P}^{\mathrm{T}}\boldsymbol{W})^{-1}\boldsymbol{q} \tag{5}$$

where $\boldsymbol{W}$ is the weight matrix of the PLS model above, $\boldsymbol{P}$ is the loading matrix, $\boldsymbol{q}$ is the regression coefficient vector for the inner relation between $\boldsymbol{t}$ and the score matrix $\boldsymbol{T}$.

Then the score vector $\boldsymbol{t}_{\perp}$ of OSC is calculated by

$$\boldsymbol{t}_{\perp}=\boldsymbol{X}\boldsymbol{w}_{\perp} \tag{6}$$

And the loading vector $\boldsymbol{p}_{\perp}$ of OSC is calculated by

$$\boldsymbol{p}_{\perp}=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{t}_{\perp}/(\boldsymbol{t}_{\perp}^{\mathrm{T}}\boldsymbol{t}_{\perp}) \tag{7}$$

The corrected spectra can be obtained by subtracting the orthogonal part from $\boldsymbol{X}$,

$$\boldsymbol{X}_{\mathrm{OSC}}=\boldsymbol{X}-\boldsymbol{t}_{\perp}\boldsymbol{p}_{\perp}^{\mathrm{T}} \tag{8}$$

Several OSC-components can be removed by repeating the steps above.

To correct the new spectra, a score matrix is first calculated.

$$\boldsymbol{T}_{\mathrm{new}}=\boldsymbol{X}_{\mathrm{new}}\boldsymbol{W}_{\perp} \tag{9}$$

And then the OSC-component is removed using loading matrix $\boldsymbol{P}_{\perp}$

$$\boldsymbol{X}_{\mathrm{OSC}}=\boldsymbol{X}_{\mathrm{new}}-\boldsymbol{T}_{\mathrm{new}}\boldsymbol{P}_{\perp}^{\mathrm{T}} \tag{10}$$

EXPERIMENTAL DETAILS

Simulated dataset

The simulated data mimicked common NIR spectral data with excessive background. A simulated spectrum was obtained by adding a background (a broader Gaussian peak), an analytical signal (a narrower Gaussian peak), and normal distributed noise. The Gaussian peaks were generated by

$$x_i=A\exp[-\frac{1}{2}(x_i-x_0)^2/\sigma^2] \tag{11}$$

where A is the height at the centre, x_0 is the position of the centre and σ is the standard deviation of the Gaussian peak, and $\sqrt{2}\sigma$ is the peak width (Brereton, 2003).

A simulated dataset with 40 samples and 400 variables were used in this paper. The parameters x_0, σ of background and analytical signal are shown in Table 1. The peak heights A are proportional to concentrations for analytical signal, whereas the peak heights for background are random. The simulated spectra are shown in Fig.1.

Table 1 The parameters of the simulated dataset

Parameters	Background	Analytical signal
x_0	180	250
σ	100	10

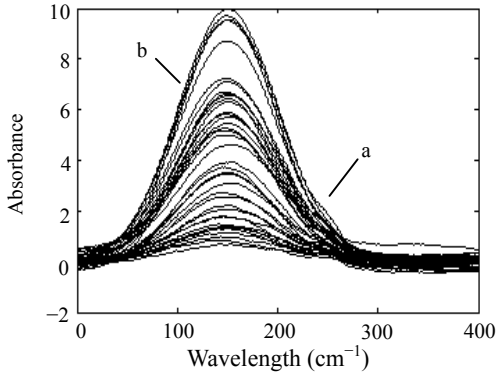

Fig.1 The simulated NIR spectra
a: The simulated analytical signal, b: The simulated background

Real dataset

NIR spectral data on process analysis of purification of natural herb (Yang *et al*., 2003) was used to test background correction methods. The eluting process of a natural herb, *Coptis chinensis*, was monitored by NIR spectroscopy simultaneously. The spectra were measured in the range 11000 to 4000 cm^{-1} with a resolution of 8 cm^{-1} (1816 points). Forty samples were obtained in eluting process (Fig.2). The contents of an active constituent in *Coptis chinensis*, berberine were determined by HPLC.

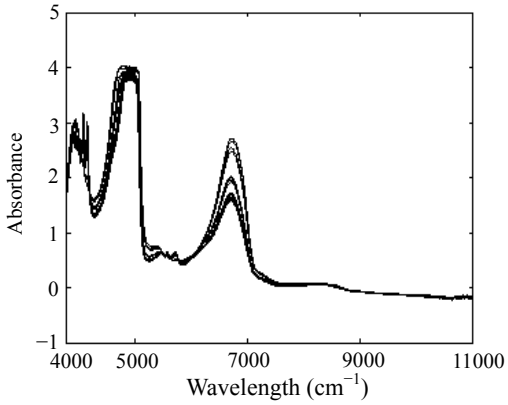

Fig.2 The NIR spectra of process analysis of purification of *Coptis chinensis*

Computation

All computation was performed in Matlab Version 6.5. All wavelet transforms were achieved using the Mathworks wavelet toolbox for Matlab.

RESULTS AND DISCUSSION

For evaluating the performances of the different background correction methods, calibration models were established by partial least square (PLS) regression in the corrected spectra. The datasets were divided into two sets, one for building PLS model (calibration set) and the other for validating the model (validation set). The optimal number of PLS components was determined by Leave-one-out Cross-Validation. The room mean square errors of correction (RMSEC, calibration set) and prediction (RMSEP, validation set) and the multiple correlation coefficient r^2 of two sets were calculated.

Analysis of simulated dataset

The simulated data mimicked typical NIR spectral data with excessive background. The peak heights and widths of the background are obviously larger than those of the analytical signals (Fig.1).

Forty samples were randomly divided into two datasets, 20 in calibration set and the others in validation set. OSC and some classical background correction methods were selected for correction of excessive background. After being corrected by different background correction method, a PLS model of analytical signals was built. The results are shown in Table 2.

Without background correction, the RMSEP was 3.536 and the r^2 of the validation set was only 0.8862. The results showed that the prediction of PLS model should be improved. But after correcting by offset correction, MSC, SNV, DT and second derivative, poor results were obtained. It was clear that these methods were unsuitable for the correction of excessive background. The results of the correction by first derivative method and wavelet method were better than the original PLS, but still not acceptable.

Table 2 The results of the simulated dataset

Correction methods	Calibration set			Validation set	
	L.V.*	*RMSEC*	r^2	*RMSEP*	r^2
None	7	2.006	0.97430	3.536	0.88620
Offset correction	6	2.277	0.96690	4.447	0.81190
MSC	1	12.070	0.07027	10.200	0.03805
SNV	1	11.980	0.08472	10.330	0.05060
DT	4	2.331	0.96530	3.984	0.85500
First derivative	3	1.428	0.98700	3.172	0.91200
Second derivative	2	1.888	0.97720	10.080	0.11930
Wavelet	4	1.843	0.97830	2.947	0.91970
OSC	2	2.011	0.97420	2.533	0.94640

* L.V. is the latent variables of PLS models

Use of OSC for correction yielded perfect result. The prediction error and complexity of models were decreased distinctly. Compared to original PLS, RMSEP decreased from 3.536 to 2.533, r^2 increased from 0.8862 to 0.9464, and latent variables of PLS models decreased from 7 to 2. The results of simulated dataset showed that OSC was the only effective method for excessive background correction.

Analysis of the real dataset

In order to test the efficiency of OSC in practice, the method was tested by real NIR spectral data on *Coptis chinensis* extracts. Half of the samples were used in the calibration set and the other half in the validation set. Fig.2 shows that the absorbance peaks of the O-H bond at 6944 cm^{-1} and 5155 cm^{-1} are very large, these peaks are mainly attributed to water and alcohol. But the absorbance of berberine was so weak that no absorbance peaks can be found in the spectra. Building a PLS model of berberine would be a challenging work.

After correction by different background correction method, a PLS model of berberine was built. The results are shown in Table 3 showing that without data correction, poor result was obtained. Slightly better or worse results were obtained by the classical data correction methods such as offset correction, MSC, SNV, DT, wavelet, first and second derivative. It seems that these methods are poorly suited for correction of excessive background variations. The best results were obtained by OSC, as shown in the simulated datasets.

Table 3 The results of the real NIR dataset

Correction methods	Calibration set			Validation set	
	L.V.*	*RMSEC*	r^2	*RMSEP*	r^2
None	2	5.6830	0.6021	5.356	0.4787
Offset correction	4	2.1610	0.9424	4.954	0.6984
MSC	2	4.9940	0.6927	4.994	0.5282
SNV	2	4.9810	0.6942	4.999	0.5289
DT	2	5.6990	0.5998	5.378	0.4528
First derivative	2	2.3600	0.9314	4.447	0.7383
Second derivative	2	4.4660	0.7543	4.967	0.5255
Wavelet	7	0.5967	0.9956	3.612	0.8160
OSC	3	0.8184	0.9852	2.144	0.9487

*L.V. is the latent variables of PLS models

CONCLUSSION

This paper discussed the pretreatment of NIR spectra of plant extracts with emphasis on the problem of excessive background. As mentioned before, many classical background correction methods have been developed. But the classical methods of offset correction, MSC, SNV, DT, first and second derivative, and wavelet, seem unable to correct the excessive background. Thus a novel popular method, OSC, was used to correct the excessive background in this paper.

A simulated datasets and a real NIR dataset were used for testing the efficiency of different background correction methods. The result of both simulated dataset and real dataset showed that the OSC is the only effective method for correction of excessive background.

References

Alsberg, B.K., Woodward, A.M., Kell, D.B., 1997. An introduction to wavelet transforms for chemometricians: A time-frequency approach. *Chemometrics and Intelligent Laboratory Systems*, **37**:215-239.

Andersson, C.A., 1999. Direct orthogonalization. *Chemometrics and Intelligent Laboratory Systems*, **47**:51-63.

Barclay, V.J., Bonner, R.F., 1997. Application of wavelet transforms to experimental spectra: Smoothing, denosing, and data set compression. *Analytical Chemistry*, **69**:78-90.

Barnes, R.J., Dhanoa, M.S., Lister, S.J., 1989. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy*, **43**:772-777.

Brereton, R.G., 2003. Chemometrics: Data Analysis for the Laboratory and Chemical Plant. John & Wiley, Chichester, England, p.123.

Brindle, J.T., Nicholson, J.K., Schofield, P.M., Grainger, D.J.,

Holmes, E., 2003. Application of chemometrics to H-1 NMR spectroscopic data to investigate a relationship between human serum metabolic profiles and hypertension. *Analyst*, **128**:32-36.

Candolfi, A., Maesschalck, R.D., Jouan-Rimbaud, D., Hailey, P.A., Massart, D.L., 1999. The influence of data pre-processing in the pattern recognition of excipients near-infrared spectra. *Journal of Pharmaceutical and Biomedical Analysis*, **21**:115-132.

Fearn, T., 2000. On orthogonal signal correction. *Chemometrics and Intelligent Laboratory Systems*, **50**:47-52.

Geladi, P., MacDougall, D., Martens, H., 1985. Linearization and scatter-correction for near-infrared reflectance spectra of meat. *Applied Spectroscopy*, **39**:491-500.

Karstang, T.V., Kvalheim, K., 1991. Multivariate prediction and background correction using local modelling and derivative spectroscopy. *Analytical Chemistry*, **63**:767-772.

Ma, X.G., Zhang, Z.X., 2003. Application of wavelet transform to background correction in inductively coupled plasma atomic emission spectrometry. *Analytica Chimica Acta*, **485**:233-239.

Mittermayr, C.R., Tan, H.W., Brown, S.D., 2001. Robust calibration with respect to background variation. *Applied Spectroscopy*, **55**:827-833.

Peussa, M., Harkonen, S., Puputti, J., Niinisto, L., 2000. Application of PLS multivariate calibration for the determination of the hydroxyl group content in calcined silica by DRIFTS. *Journal of Chemometrics*, **14**:501-512.

Schechter, I., 1995. Correction for nonlinear fluctuating background in monovariable analytical systems. *Analytical Chemistry*, **67**:2580-2585.

Sjoblom, J., Svensson, O., Josefson, M., Kullberg, H., Wold, S., 1998. An evaluation of orthogonal signal correction applied to calibration transfer of near infrared spectra. *Chemometrics and Intelligent Laboratory Systems*, **44**:229-244.

Sun, J.G., 1997. Statistical analysis of NIR data: Data pretreatment. *Journal of Chemometrics*, **11**:525-532.

Tahboub, Y.R., Pardue, H.L., 1985. Evaluation of multiwavelength first- and second-derivative spectra for the quantitation of mixtures of polynuclear aromatic hydrocarbons. *Analytical Chemistry*, **57**:38-41.

Tan, H.W., Brown, S.D., 2002. Wavelet analysis applied to removing non-constant, varying spectroscopic background in multivariate calibration. *Journal of Chemometrics*, **16**:228-240.

Trygg, J., Wold, S., 2002. Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics*, **16**:119-128.

Westerhuis, J.A., de Jong, S., Smilde, A.K., 2001. Direct orthogonal signal correction. *Chemometrics and Intelligent Laboratory Systems*, **56**:13-25.

Wold, S., Antti, H., Lindgren, F., Ohman, J., 1998. Orthogonal signal correction of near-infrared spectra. *Chemometrics and Intelligent Laboratory Systems*, **44**:175-185.

Yang, N.L., Cheng, Y.Y., Qu, H.B., 2003. A new method for process analysis of purification of natural herb using NIRS. *Acta Chimica Sinica*, **61**:742-747.