Journal of Mathematical Chemistry Vol. 36, No. 2, June 2004 (© 2004)

A novel preprocessing algorithm for three-way HPLC data

Bi Xian, Tonghua Li*, Chen Kai, Tongcheng Cao and Jianrong Huang

Department of Chemistry, Tongji University, Shanghai 200092, China
E-mail: lith@tongji.edu.cn

Yunpeng Qi

College of Pharmacology, Second Military Medical University, Shanghai 200433, China

Received 12 January 2004; revised 8 March 2004

Orthogonal signal correction (OSC) was a data preprocessing algorithm. It ensured that the filtered information was irrelevant to concentration data while using it to filter the noise from the original data. This paper extended the OSC application range from two-way data to three-way data. Two drug data sets, Enoxacin, Norfloxacin, Ciprofloxacin and Betamethasone, cortisone acetate, prednisone acetate, showed that the application of the OSC algorithm to three-way HPLC data was feasible and needed further research.

KEY WORDS: orthogonal signal correction, OSC, three-way, PLS, HPLC

1. Introduction

In spectra data analysis, data are usually preprocessed by chemometrics algorithm to correct the undesirable variation and to improve the quality of multivariate calibration. However, there is no assurance that any of signal-processing techniques will remove only irrelevant information from the measured data [1]. For this reason, Wold et al. [2] developed orthogonal signal correction (OSC) to remove systematic variation from the measured matrix that is unrelated, or called orthogonal, to the concentration matrix. Since then, several groups [3–10] have published various OSC algorithms. In the same journal issue of Wold's algorithm, OSC was applied to calibration transfer by Sjöblom et al. [13]. Approximately one year later, direct orthogonalization (DO) was presented by Andersson [4]. Wise and his colleagues [5] developed another OSC algorithm in his website. Recently, Fearn [6] introduced a new algorithm for OSC and pointed out at the same time that the two algorithms used in reference [2, 3], both originating from Wold were similar but not identical. Westerhuis et al. [8]

* Corresponding author.

0259-9791/04/0600-0129/0

also presented his direct OSC in 2001. Orthogonal projection to latent structures (OPLS) was recently developed by Trygg and Wold [9,10].

Meanwhile, with the development of modern analysis instrument, more and more three-way data appeared. Most two-way data are generated by the multidetection instruments which are more and more popular in laboratories, such as HPLC-DAD, GC-MS, etc. Normally, when several two-way data are folded, one has a three-way data. Chemists also presented many data analysis methods to the analysis of three-way data. For example, Bro [11] expanded the PLS algorithm on three-way data which can be directly used to three-way data quantitative analysis.

Up to now, the OSC algorithm was applied to the two-way data. This paper extended the OSC application range from two-way data to three-way data based on the thought of Bro's three-way PLS and a novel technology of projection. Two drug data sets were used to verify the three-way OSC and the result showed that the three-way OSC was feasible and needed to further optimize its effect.

2. Theory and algorithm

2.1. Notation

Three-way matrices are denoted by capital bold italic characters ($\mathbf{X}$, $\mathbf{Y}$), two-way matrices are denoted by capital bold characters ($\mathbf{X}$, $\mathbf{Y}$), column vectors by small bold characters ($\mathbf{p}$,$\mathbf{t}$) and row vectors by transpose vectors ($\mathbf{p}^{\mathrm{T}}$). Indices are designated by small nonbold characters (j), index limits by capital nonbold characters (K) and constants by small italic characters (r).

2.2. Orthogonal signal correction (OSC)

Orthogonal signal correction is a data treatment technique developed by Wold et al. [2], whose goal is to correct the X data matrix removing the information that is orthogonal to the concentration matrix Y.

The algorithm used in this type of correction is similar to the NIPALS algorithm, commonly used in PLS, in each step of both algorithms, the weight vector (w) is modified, imposing the condition that $t = X \cdot w$ is orthogonal to the Y matrix, and where t is the corresponding score vector. In PLS, the condition would be calculated to maximize the covariance among X and Y, but in OSC just the opposite is attempted, to minimize this covariance, making t as close as possible to the orthogonality with Y. The result of this way makes the scores and the loadings contain the information not related to the concentration. Each internal component of OSC removes a part of the X matrix variance.

As removed more variance, the number of internal components to use in the OSC model will be higher. These internal components are similar to factors

in a PLS calibration. Once the information not correlated with the concentration has been done, it will be removed from the spectral data, subtracting from the X matrix, the product of the scores orthogonal to the concentration (T) and the loadings matrices (P'):

$$\mathbf{X}_{\text{osc}} = \mathbf{X} - \sum_{i=1}^{n} \mathbf{T}_i \mathbf{P}'_i,$$

where n is OSC components' number.

2.3. The three-way data model and three-way PLS algorithm [12,13]

The three-way data model was $\mathbf{X} = \mathbf{Y} \otimes \mathbf{C} \otimes \mathbf{S} + \mathbf{E}$. X was the spectral matrix whose size was $L*M*N$, Y was the response matrix whose size was $L * K$. L was the number of rows of X and Y. M and N were the hits of spectrum and chromatogram. K was the number of compounds. $\otimes$ was the outer product of vectors or matrix, i.e. Kronecker product. The Tucker model was adapted to decompose the ordinary N-way data. The Tucker model includes Tucker2 model and Tucker3 model. Tucker2 model spreads the N-way data to two-way data, mainly spreads on the sample dimension. Tucker3 model regards the three-way data as the outer product between three matrixes and a core matrix and does not differentiate the chemistry meanings difference in every direction of three-way data.

The three-way PLS algorithm used the Tucker2 model and the core thought was familiar with the ordinary PLS.

The brief steps were as follows:

1. Spread $\boldsymbol{X}$ matrix by tucker2 model to $\mathbf{X}$ (L $\times$ MN) and PCs $= 0$.
2. Generate a random vector $\mathbf{u}$ as the original loading vector of $\mathbf{Y}$

$$\mathbf{u} = r \text{ and } (K, 1), \quad \text{PCs} = \text{PCs} + 1.$$

3. Calculate the weight vector $\mathbf{w}$

$$\mathbf{w} = \mathbf{X}'\mathbf{u}$$

(a) w is a vector whose size is $NM \times 1$, reshape it to a $N * M$ matrix

$$\mathbf{wW} = \text{reshape } (\mathbf{w})$$

(b) Singular value decompose the $\mathbf{wW}$

$$[\mathbf{w1}, \mathbf{d}, \mathbf{w2}] = \text{svd } (\mathbf{wW})$$

(c) recalculate the weight vector $\mathbf{w}$ and save $\mathbf{w}_1\mathbf{w}_2$ to the vector $\mathbf{www}$

$$\mathbf{w} = \mathbf{w1} \otimes \mathbf{w2}, \quad \mathbf{www} = [\mathbf{w1}; \mathbf{w2}]$$

4. Calculate the score of **X**

$$\mathbf{t} = \mathbf{Xw}$$

5. Calculate the loading of **Y**: **q**

$$\mathbf{q} = \mathbf{Y}'\mathbf{t}, \mathbf{q} = \mathbf{q}/||\mathbf{q}||$$

6. Calculate the new value of **u**

$$\mathbf{u} = \mathbf{Yq}$$

7. Check the astringency of **u**, if constringency go to step 8 otherwise go to steps 3–6
8. Save following result for latter steps.

$$\mathbf{T} = [\mathbf{T}; \mathbf{t}]; \mathbf{W} = [\mathbf{W}; \mathbf{www}]; \mathbf{U} = [\mathbf{U}; \mathbf{u}]; \mathbf{Q} = [\mathbf{Q}; \mathbf{q}]$$

9. Calculate the relationship between **T** and **U**

$$\mathbf{B}(1 : \mathrm{PCs}, \mathrm{PCs}) = (\mathbf{T} * \mathbf{T}')^{-1} * \mathbf{U}(:, \mathrm{PCs})$$

10. Calculate the **X, Y** information in all components

$$\mathbf{Y}_{\mathbf{pred}} = \mathbf{T} * \mathbf{B}(1 : \mathrm{PCs}, 1 : \mathrm{PCs}) * \mathbf{Q}';$$
$$\mathbf{X}_{\mathbf{model}} = \mathbf{X}_{\mathbf{model}} + \mathbf{T}(:, \mathrm{PCs}) * \mathbf{w}'$$

11. Calculate the residua matrix of **X** and **Y**

$$\mathbf{Y} = \mathbf{Y0} - \mathbf{Y}_{\mathrm{pred}}$$
$$\mathbf{X} = \mathbf{X0} - \mathbf{X}_{\mathrm{model}}$$

3. Three-way orthogonal signal correction

3.1. The algorithm introduction and the improvement of projection

Our three-way OSC algorithm was based on the Tucker3 model. After decomposed three-way matrix $\boldsymbol{X}$, the biggest score vector t of the response matrix **X** was calculated. In the following steps the OSC was introduced and repeated until constringency. In the next step the loading vector **p** was calculated. Finally the revised **X** was achieved by $\mathbf{X} = \mathbf{X} - \mathbf{tp}'$. As for the unknown sample, used Tucker3 model to decompose the spectral matrix $\boldsymbol{X}$ and generated the score matrix **T**. The revised matrix **Xu** also was obtained by the score vector and loading vector. Repeated the above steps for certain times which were defined as the compounds number of three-way OSC. Same as OSC, the filtered information generally means the variety of noise.

In steps 4–5 of following algorithm, **T** replaced $\boldsymbol{X}$ to calculate the weight vector **w** and scoring vector **t**. This change is our improvement of three-way OSC. The reasons included two aspects. First, **T** was a two-way matrix and $\boldsymbol{X}$ was a three-way matrix, so using **T** could make the algorithm simpler. Moreover, $\mathbf{TT}^{+}$ and $\boldsymbol{XX}^{+}$ were equal mathematically. Therefore, it solved well the definition of generalized inverse of high-way matrix $\boldsymbol{X}$ when OSC was applied to high-way data. It is necessary to calculate weight vector **w** [14].

3.2. The detailed algorithm

The detailed algorithm was as follows:

1. Use the Tucker3 model to decompose three-way matrix **X**

$$[\mathbf{Factors}, \mathbf{G}, \mathbf{Explm}, \mathbf{Xm}] = tucker(\boldsymbol{X}, \mathrm{PCs})$$
$$[\mathbf{T}, \mathbf{WJ}, \mathbf{WK}] = factlet2(\mathbf{Factors}, \mathrm{PCs})$$

2. **t** is the biggest score vector of the response matrix **X**

$$\mathbf{t} = \mathbf{T}(:, 1)$$

3. Calculate the projection of **t** on the subspace of the property matrix **Y**

$$\mathbf{t} = (\mathbf{I} - \mathbf{YY}^{+})\mathbf{t}$$

4. Calculate the weight vector **w**

$$\mathbf{w} = \mathbf{T}^{+}\mathbf{t}$$

5. Calculate the new score vector

$$\mathbf{t} = \mathbf{Tw}$$

6. Repeat steps 3–5 until constringency
7. Calculate the loading vector **p**, **p** describe the biggest compound information in **X** which is irrelevant to **Y**

$$\mathbf{p} = \mathbf{X}'\mathbf{t}/(\mathbf{t}'\mathbf{t})$$

8. Calculate the revised **X**

$$\mathbf{X} = \mathbf{X} - \mathbf{tp}'$$

Process to the unknown sample:

9. Use Tucker3 model to decompose the spectral matrix $\boldsymbol{X}$ and generate the score matrix $\mathbf{T}$
10. Calculate the score vector $\mathbf{t}$

$$\mathbf{t} = \mathbf{Tw}$$

11. Calculate the loading vector $\mathbf{p}$

$$\mathbf{p} = \mathbf{X}'\mathbf{t}/(\mathbf{t}'\mathbf{t})$$

12. Calculate the revised matrix $\mathbf{Xu}$

$$\mathbf{Xu} = \mathbf{Xu} - \mathbf{tp}'.$$

4. Experimental

4.1. Apparatus

The HPLC instrument is Waters 966 HPLC-DAD with Waters 515 HPLC pump. The chromatographic column is Hypersil C18 (5 μm, 250 mm $\times$ 4.6 mm). Mobile phase is 65% alcohol, flow rate 1.0 ml/min, wavelength 210–340 nm, sensitivity 1.00 AUF, the column temperature room temperatures and sample size 20 μl.

4.2. Software

The signal-processing programs were implemented in MATLAB v.6.5 (Mathworks) and run on a PC with a Pentium IV 1.6 GHz processor and 256 MB memory.

4.3. Data sets

4.3.1. Drug data set I

The data set I was a three compound sample: Enoxacin, Norfloxacin, and Ciprofloxacin. There were eight samples and each sample generated a two-way data by HPLC-DAD. Superposed eight samples and achieved a three-way data set. The real size of the sample is $294 * 167 * 8$. The spectrum sampling point is 167 and that of chromatogram is 294.

4.3.2. Drug data set II

The data set II was a three compound sample: Betamethasone, cortisone acetate, prednisone acetate. There are five samples and each sample generated a two-way data by HPLC-DAD. Superposed five samples and achieved a three-way

data set. The real size of sample is $400 * 83 * 5$. The spectrum sampling point is 83, and that of chromatogram is 400.

4.4. Data processing

4.4.1. Chromatogram peak alignment

Because every sample was measured respectively, the sampling point of the chromatogram was not equal among these samples. It was necessary to align the chromatogram peak. The authors used Lagrange Interpolating for three points to align the chromatogram peak. Please refer to correlative books or references to see the detail algorithm.

4.4.2. Data selection

The size of the first sample data was $36 * 80$. These data were chosen from nearly 50,000 data by wiping off those data that were approximately equal to zero. The reasons were that: first, improve the calculation speed, second, in this algorithm, X needs to singular value decompose, wiping off some data can avoid the instability.

4.4.3. Cross-validation

To verify the quality of the model, we performed full cross-validation of the constructed models. The cross-validation method was the traditional "leave one out."

5. Results and discussion

We used the RT and RS as the evaluation index. RT is the percentage and RS is standard deviation. Their calculation equations were:

$$\mathrm{RT}_i = \frac{y_{i,\mathrm{pre}}}{y_{i,\mathrm{obs}}} \times 100$$

$$\mathrm{RS}^2 = \sum_{i=1}^{K} \frac{\left(y_{i,\mathrm{obs}} - y_{i,\mathrm{pre}}\right)^2}{K}.$$

5.1. The result of data set

The concentration of Enoxacin, Norfloxacin and Ciprofloxacin (unit: μg/ml) was listed in table 1.

From the table 1, the RT of three-way OSC-PLS was better than that of three-way PLS except the Ciprofloxacin. As for the RS, the three-way OSC-PLS result was all better than that of three-way PLS.

Table 1
The result of drug data set I.

	Enoxacin			Norfloxacin			Ciprofloxacin		
	Experiment value	Three-way PLS	Three-way OSC	Experiment value	Three-way PLS	Three-way OSC	Experiment value	Three-way PLS	Three-way OSC
1	30	29.55	29.56	30	30.16	30.21	30	29.83	29.68
2	30	29.91	29.98	30	30.03	29.99	30	29.96	29.97
3	30	30.30	30.21	30	29.94	30.01	30	30.13	30.13
4	10	10.09	10.20	30	30.05	29.96	20	20.06	20.07
5	10	10.11	10.10	30	30.02	30.07	20	20.07	20.07
6	10	9.82	9.75	30	29.93	29.94	20	19.89	19.87
7	40	40.97	40.41	16	15.43	15.56	30	30.28	30.17
8	40	39.16	39.42	16	16.80	16.69	30	29.87	29.83
RT		99.98	99.99		100.23	99.89		100.04	99.91
RS		0.50	0.33		0.35	0.21		0.16	0.14

Table 2
The result of drug data set II.

	Betamethasone			Cortisone acetate			Prednisone acetate		
	Experiment value	Three-way PLS	Three-way OSC	Experiment value	Three-way PLS	Three-way OSC	Experiment value	Three-way PLS	Three-way OSC
1	0.1	0.100	0.099	0.1	0.099	0.100	0.1	0.101	0.100
2	0.1	0.101	0.102	0.1	0.096	0.102	0.1	0.099	0.098
3	0.1	0.099	0.099	0.1	0.102	0.098	0.1	0.099	0.101
4	0.1	0.100	0.099	0.1	0.100	0.098	0.1	0.098	0.102
5	0.1	0.097	0.095	0.1	0.098	0.096	0.1	0.099	0.097
RT		99.40	98.80		99.00	98.80		99.60	99.20
RS		0.15	0.25		0.23	0.24		0.13	0.19

The concentration of betamethasone, cortisone acetate, prednisone acetate (unit: μg/ml) was listed in table 2. This set was used to verify the stability of two algorithms, so the concentration of all samples was same.

From table 2, the RT of three-way OSC was slightly worse than that of three-way PLS. The RS of three-way OSC was approximatively as much as that of three-way PLS. From the repeat experiment, the stability of three-way PLS was slightly better than three-way OSC. It was concluded that the three-way OSC algorithm was feasible on quantitative analysis and needed further research.

6. Conclusions

The algorithm realized the application of OSC from two-way to three-way. Two examples showed that the application of three-way OSC algorithm to three-way data was feasible, especially in the quantitative prediction of compound concentration. We believed that three-way OSC algorithm could also improve the qualitative and quantitative analysis ability of other calibration methods.

Acknowledgments

The authors are grateful to the National Natural Science Foundation of China for financial support. (No. 20275026) The authors also thank Dr. Qi for providing the experimental data.

References

[1] Baibing Li, A. Julian Morris and Elaine B. Martin, J. Chemometr. 16 (2002) 556.
[2] Svante Wold, Henrik Antti, Fredrik Lindgren and JerKer Ohman, Chemometr. Intell. Lab. Syst. 44 (1998) 175.
[3] J. Sjöblom, O. Svensson, M. Josefson, H. Kullberg and S. Wold, Chemometr. Intell. Lab. Syst. 44 (1998) 229.
[4] C.A. Andersson, Chemometr. Intell. Lab. Syst. 47 (1999) 51.
[5] B.M. Wise and N.B. Gallagher, http://www.eigenvector.com/MATLAB/OSC.html
[6] T. Fearn, Chemometr. Intell. Lab. Syst. 50 (2000) 47.
[7] J.A. Fernańdez Pierna, D.L. Massart, O.E. de Noord and Ph. Ricoux, Chemometr. Intell. Lab. Syst. 55 (2001) 101.
[8] J.A. Westerhuis, S. de Jong and A.K. Smilde, Chemometr. Intell. Lab. Syst. 56 (2001) 13.
[9] S. Wold, J. Trygg, A. Berglund and H. Antii, Chemometr. Intell. Lab. Syst. 58 (2001) 131.
[10] J. Trygg and S. Wold, J. Chemometr. 16 (2002) 119.
[11] Rasmus Bro, J. Chemometr. 10 (1996) 47.
[12] http://www.models.kvl.dk/courses/tucker.
[13] http://newton.foodsci.kvl.dk/foodtech.
[14] Huang Jian-Rong, Li Tong-Hua and Chen Kai, J. Chin. Univ. 24 (2003)1009.