Asian Journal of Chemistry; Vol. 23, No. 7 (2011), 3214-3216

ASIAN JOURNAL OF CHEMISTRY

www.asianjournalofchemistry.co.in

Delaunay Triangulation Local Method for Analysis of Near-Infrared Spectra of Plant Samples

Y.K. LI[1,*] and J. JING[2]

[1]College of Environment Science and Engineering, North China Electric Power University, Baoding, P.R. China
[2]Department of Electrical Engineering, North China Electric Power University, Baoding, P.R. China

*Corresponding author: Tel: +86 312 7522277; E-mail: lyk800@tom.com

(*Received*: 26 November 2010; *Accepted*: 21 March 2011) AJC-9765

The proposed delaunay triangulation local method is applied as a local calibration method in the analysis of near infrared data. The method forms mesh of simplexes in principal component space and then carries out calibration subsets selection based on the situation of unknown sample in the mesh. It obtains higher prediction accuracy than the local method, in which calibration subsets are the closest points chosen through immediate Euclidean distances between samples in principal component space. On comparing with global PLS method, delaunay triangulation local method have better results with few numbers of principal components. Furthermore, it is no need to establish the calibration model and it is fast and simple. As a result, the proposed delaunay triangulation method can be used as a valid local method for distinguishing similarity of complex samples in near infrared spectral data analysis.

Key Words: Delaunay triangulation, Local method, Principal component analysis, Near-infrared spectroscopy, Quantitative analysis.

INTRODUCTION

Local methods were applied in chemometrics, in which variant sample-dependent models are constructed by the sample subset selected from large calibration samples set for each predicted sample[1,2]. The calibration subset selection depends on location of the unknown sample in the calibration domain, that is to say, selection of calibration sample subset by definite "similarity criteria" between calibration samples and unknown sample. In local methods, Variant "distance" rules is usually used as similarity criteria, including Euclidean distance[3,4], Mahalanobis distance[5], Manhattan distance and Minkowski distance in spectra space or principal component (PC) space.

Topological methods including k-nearest neighbors (kNN) method[6], multi-dimensional simplex interpolation (MSI)[7], the law of mixtures (LM)[8] method and delaunay triangulation (DT)[9], *etc*., is another type of local method, which acquires calibration samples whose property values, *e.g*., concentrations are close to the predicted sample by topological technology and predict sample directly not in need of constructing calibration models. Along with developments of computer technology and computer graphics, delaunay triangulation method has been one of the most popular methods of full-automatic "high grade" meshes generation. It is an important pre-treatment method to numerical analysis or Graph theory. The delaunay triangulation has been widely used in many fields such as statistics, solid state physics, computational geometry, *etc*.[10-12] and a few papers have been reported in quantitative analysis of near infrared spectroscopy for the present[9,13,14].

Selection of calibration sample subset is the core of local method. In this work, delaunay triangulation local method builds a lattice in principal component space by principal component analysis[15,16]. The prediction of unknown object is calculated with the calibration samples constructing the surrounding or enclosing simplex for each point (object). In calculation procedure, Euclidean distances between them were evaluated as weighted values. Simultaneously, the closest points immediately chosen through Euclidean distances in principal component space between samples were investigated, which got worse results than delaunay triangulation method. At last, compared with global PLS[17] method, prediction results are improved significantly in analysis of complex plant samples by near infrared.

Theory and algorithm

Principal component analysis (PCA): When large multivariate datasets are analyzed, it is often desirable to reduce their dimensionality. Principal component analysis is one technique for doing this. Its goal is to extract the important information from the datasets, to represent it as a set of new orthogonal variables called principal components (PCs). The mathe-matical formula of principal component analysis can be expressed as follows:

$$X_{n\times p} = T_{n\times f} P_{f\times p} \quad (1)$$

where $X_{n\times p}$ is the matrix of response variables, $T_{n\times f}$ and $P_{f\times p}$ are the score and loading vectors, respectively. n is the number of samples, p is the number of variables and f is the number of principal components.

Delaunay triangulation (DT) local method: Delaunay triangulation is one of the most popular methods for generation of unstructured meshes. It originates from the study of structures in computational geometry and can generate "high grade" meshes quickly. Mesh partitioning is a key step in the pre-treatment of finite element analysis. Delaunay triangulation has two important restrictions, namely, maximum and minimum angle criterion and circle criterion, which ensure that the simplex surrounding the point is "good". The detail steps to construct delaunay triangulation mesh can refer to literature[14].

General procedure: According to each individual unknown sample situated in the delaunay triangulation mesh, its calibration subsets are chosen from the whole calibration data sets. Then calculate the content of unknown sample by using content of the selected subsets. The detailed procedures can be described as follows: (1) The near infrared spectral data are randomly divided into a calibration set, an assessing set and a prediction set. (2) At number of principal component n_f, scoring matrix T of calibration set and assessing set are obtained, respectively by using principal component analysis, then delaunay triangulation mesh is constructed by calibration set in f-dimensional space and assessing set are cast in this space. (3) In f-dimension space, if an unknown sample M falls within a simplex defined by f + 1 neighbor samples (M_1, M_2, …, M_{f+1}), its associated property can be calculated through the properties of f + 1 neighbours according to the following equations:

$$d_{M_1} = \sqrt{\sum_{i=1}^{f}(t_{iM} - t_{iM_1})^2} \quad (2)$$

$$d_{M_{f+1}} = \sqrt{\sum_{i=1}^{f}(t_{iM} - t_{iM_{f+1}})^2} \quad (3)$$

$$D = d_{M_1} + d_{M_2} + \ldots + d_{M_{f+1}} \quad (4)$$

$$\alpha_{M_1} = \frac{D - d_{M_1}}{fD} \quad (5)$$

$$\alpha_{M_{f+1}} = \frac{D - d_{M_{f+1}}}{fD} \quad (6)$$

where, $t_{iM}, \ldots, t_{iM_{f+1}}$ are the scores of the objects in the projected principal component-space and $d_{M_1}, \ldots, t_{iM_{f+1}}$ are the Euclidean distances between unknown object and its neighbours. D is the summation of f + 1 distance. $\alpha_{M_1}, \ldots, \alpha_{M_{f+1}}$ are the contributions of samples M_1, …, M_{f+1}, individually, which is inverse ratio to the distances between itself and the unknown object. Then the property of an unknown sample can calculated as follows:

$$y_M = \alpha_{M_1} y_{M_1} + \alpha_{M_2} y_{M_2} + \ldots + \alpha_{M_{f+1}} y_{M_{f+1}} \quad (7)$$

where $y_M, y_{M_1}, \ldots, y_{M_{f+1}}$ are the property of calibration subset samples M, M_1, …, M_{f+1}, respectively. By using above formulation, predict every sample in the assessing set one by one. (4) In f-dimension space, if an unknown sample M falls outlier of the whole DT mesh, f + 1 neighbor samples with the shortest distance are used. As well, according to eqn. 8 calculate its content. (5) At optimum number of principal components decided by assessing set, predict the samples in prediction set with the same procedures.

EXPERIMENTAL

Data set consists of 80 samples of corn measured on 3 different near infrared spectrometers. (Provided by http://software.eigenvector.com/Data/Corn/index.html). The wavelength range is 1100-2498 nm at 2 nm intervals (700 channels). Moisture, oil, protein and starch components of the corn samples are predicted.

Eighty samples (or spectra) were arbitrarily divided into 3 sets. 48 samples were used for calibration data set, 16 samples were used for validation data set and the remaining 16 samples were used as prediction data set. In optimization of procedure parameters, the root mean squared error of prediction (RMSEP) of assessing set is used as evaluation criterion.

$$RMSEP = \left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right]^{1/2} \quad (8)$$

where $\hat{y}_i$ is the prediction concentration of the ith sample, y_i is the true concentration of the ith sample, n is the number of prediction samples.

RESULTS AND DISCUSSION

Determination of number of principal components: To determine the optimal number of principal components (n_f), RMSEP of assessing set at different n_f were given in the Table-1. It is shown that RMSEP almost decreased with the increase of n_f, so n_f = 4 is chosen as the optimal number of principal components. Take 3 dimensions for example, the distributions of samples projected to principal component-space are shown in Fig. 1. The delaunay triangulation-meshes are constructed by training set and the points indicate the assessing set.

TABLE-1
PREDICTION RESULTS OF ACCESSING SET BY DELAUNAY TRIANGULATION AT DIFFERENT NUMBER OF PRINCIPAL COMPONENTS (n_f)

Components	Number of principal components	RMSEP
Moisture	2	0.39
	3	0.30
	4	0.21
Oil	2	0.21
	3	0.24
	4	0.16
Proteins	2	0.62
	3	0.46
	4	0.28
Starch	2	1.15
	3	0.77
	4	0.60

Prediction results of prediction set: At n_f = 4, prediction results of the prediction set by delaunay triangulation

TABLE-2
PREDICTION RESULTS OF PREDICTION SET BY DELAUNAY TRIANGULATION, LOCAL METHOD AND PLS METHODS

Components/methods		Delaunay triangulation	Local method	PLS
Moisture	RMSEP	0.19	0.37	0.42
	Recovery (%)	102.08-96.70	93.97-105.52	112.83-94.58
Oil	RMSEP	0.17	0.17	0.18
	Recovery (%)	107.95-91.38	90.42-105.00	92.56-113.04
Proteins	RMSEP	0.28	0.40	0.32
	Recovery (%)	106.92-93.93	108.50-90.51	110.55- 95.09
Starch	RMSEP	0.57	0.75	0.89
	Recovery (%)	102.01-99.06	102.79-98.11	103.83-98.20

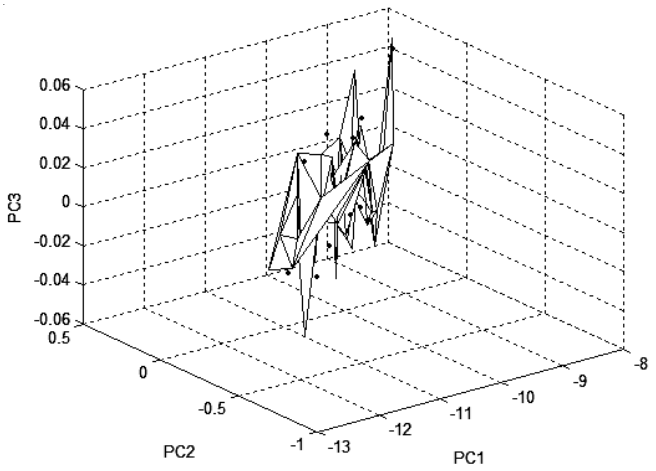

Fig. 1. Distributions of the samples projected to principal component-space of 3 dimensions

method were list in Table-2. At the same time, the prediction results by local method which selects n_{f+1} closest points (samples) of training sets to unknown object through immediately Euclidean distances between them in principal component space were also investigated. Global PLS method ($n_f = 13$) was also used to predict the same prediction set.

From Table-2, it can be concluded that the proposed delaunay triangulation method performs better than local method. It indicates that in principal component-space, concentrations of the calibration samples constructing the surrounding or enclosing simplex are closer than the closest calibration samples to each point. So it may be interpreted that delaunay triangulation mesh could establish more accurate intrinsic correlation between spectral data and properties data than "distance" in principal component-space, as a result, adjacency of spectra values between different samples according with adjacency of concentration values more exactly in delaunay triangulation method. Then compared with PLS method, delaunay triangulation get better results at small number of principal components (low dimensions), furthermore, it need not constructing a calibration model, which is more simple and quick, avoiding redundant interferences.

Conclusion

The proposed delaunay triangulation local method gets better prediction results with few principal components compared with local method and global PLS method in prediction of components of corn. It can be concluded that in principal component-space, delaunay triangulation mesh could reflect accurate intrinsic relationship in terms of spectra and concentrations, *etc*., between samples of different properties. Therefore, the proposed delaunay triangulation method may be an effective tool for quantitative analysis of complex samples in near infrared spectra.

ACKNOWLEDGEMENTS

This study is supported by the Fundamental Research Funds for the Central Universities (No. 09QL52).

REFERENCES

1. F. Despagne and D.L. Massart, *Anal. Chem*., **72**, 1657 (2000).
2. E.L. Dong, J.H. Song, S.O. Song and E.S. Yoon, *Ind. Eng. Chem. Res*., **44**, 2101 (2005).
3. B.J.C. Baxter. *Comput. Math. Appl*., **51**, 1163 (2006).
4. W.S. Cleveland and S.J. Devlin, *J. Am. Stat. Assoc*., **83**, 596 (1988).
5. X. Shi, W.S. Cai and X.G. Shao, *Chin. J. Anal. Chem*., **36**, 1093 (2008).
6. Y. Estrin,, S. Arndt, M. Heilmaier and Y. Brechet, *Acta Mater.*, **47**, 595 (1999).
7. R. Danielsson and G. Malmquist, *Chemom. Intell. Lab. Syst*., **14**, 115 (1992).
8. L. Jin, J.A.F. Pierna, F. Wahl, P. Dardenne and D.L. Massart, *Anal. Chim. Acta*, **476**, 73 (2003).
9. L. Jin, J.A.F. Pierna, Q. Xu, F. Wahl, O.E. de Noord, C.A. Saby and D.L. Massart, *Anal. Chim. Acta*, **488**, 1 (2003).
10. K.F. Mulchrone, *J. Struct. Geol*., **25**, 529 (2003).
11. K.F. Mulchrone, *J. Struct. Geol*., **25**, 689 (2003).
12. J. Lepinoux and Y. Estrin, *Acta Mater.*, **48**, 4337 (2000).
13. L. Jin, Q.S. Xu, J. Smeyers-Verbeke and D.L. Massart, *Chemom. Intell. Lab. Syst*., **80**, 87 (2006).
14. Y.K. Li, In 4th International Conference on Bioinformatics and Biomedical Engineering (ICBBE), IEEE Computer Society, US (2010).
15. H.L. Huang and A. Paolo, *J. Appl. Meteor*., **40**, 365 (2001).
16. H. Abdi and L.J. Williams, *Wiley Interdiscip. Rev.: Comput. Statist*., **2**, 433 (2010).
17. D.V. Haaland and E.V. Thomas, *Anal. Chem*., **60**, 1193 (1988).