High dimensional chemometric algorithms for analyzing temperature dependent near infrared spectra

Xiaoyu Cui, Wensheng Cai, Xueguang Shao*

Research Center for Analytical Sciences, College of Chemistry, Nankai University, Tianjin, China, 300071

**E-mail: xshao@nankai.edu.cn*

Summary: High dimensional data analysis has gained widespread attention with the rapid development of analytical instruments and experimental techniques. Benefiting from the second-order advantage, high order chemometric algorithms have shown a great ability to match the nature of data and extract the latent components from the data. In this study, multiway principal component analysis (NPCA), parallel factor analysis (PARAFAC) and alternating trilinear decomposition (ATLD) were employed, respectively, to extract the information from temperature dependent near infrared (NIR) spectra of alcohol aqueous solutions. With the knowledge of aquaphotomics, the structural variations of water and ethanol induced by temperature and concentration in the solutions were analyzed by the three algorithms.

Introduction: High dimensional data has been produced with not only the hyphenated analytical techniques, but also the different kinds of measuring methods. Temperature dependent NIR spectra, as an example, were generated by NIR spectroscopy under a regular change of temperature. These spectra can be employed to obtain the structural and quantitative information in solutions with the knowledge of aquaphotomics [1]. Ozaki et al. [2] reported a two-state structural model for water with the application of PCA. Tsenkova et al. [3] studied the effects of perturbations including temperature and salts on the NIR spectra of water in terms of hydrogen bonding using multivariate curve resolution-alternating least squares (MCR-ALS). In addition, six types of water species were analyzed by Gaussian fitting to investigate water structure induced by glucose of different concentration [4]. Studies on hydration shell of proteins using independent component analysis (ICA) [5] and two-dimensional correlation analysis [6] were also reported. On the other hand, a quantitative spectra-temperature relationship (QSTR) model was established by partial least squares (PLS) model [7] or the between-temperature model obtained by multilevel simultaneous component analysis (MSCA) [8,9]. Moreover, quantitative analysis can be achieved by the concentration-induced variation [7-9]. Furthermore, mutual factor analysis was proposed for quantitative analysis of real samples [10]. The disadvantage of these approaches is that more than one model is needed to explore the effect of temperature and concentration, and the models are not isolated. Therefore, high order chemometric algorithms may be more attractive for analyzing the high dimensional data.

Taking the advantage of the high order algorithms, the quantitative and structural information contained in the spectra was extracted and this information can be used for quantifying the content or understanding the function of the analytes. To compare the analytical performance of high order chemometric algorithms for structural and quantitative analysis, three methods, NPCA, PARAFAC and ATLD were used to decompose the temperature dependent NIR spectra data of water-ethanol mixtures [11].

Methods: Aqueous solutions containing 0, 10, 20, 30, 40, 50, 60, 70, 80, and 90 (%, v/v) ethanol were prepared and measured at different temperature (31 to 40 °C with a step of *ca.* 1 °C) for generating the three-way data. As an unfold decomposition method, NPCA is based on the ordinary two-way PCA. PARAFAC is a method considered as a generalization of the bilinear PCA for multi-way arrays. ATLD is a method for iterative trilinear decomposition of a three-way array based on the principle of ALS. The three-way NIR spectral data can be decomposed into loadings, temperature and concentration scores by the three methods, which explain the spectral features, and the variation with temperature and concentration of the corresponding features, respectively.

Results and Discussion: To compare the performance of NPCA, PARAFAC and ATLD in extraction of the spectral variations from the temperature dependent NIR spectra of water-ethanol mixtures, Figure 1 illustrates the loadings obtained by the three algorithms. Figure 1 (a1) shows the first loading obtained by NPCA, which is similar with the average spectrum of dataset with the maximum variance. The two peaks can be assigned as the overtone stretching vibrational mode of free OH group (S_0), and OH with two hydrogen bonds (S_2), respectively [1]. In the second and third loadings, three peaks are included, respectively, as shown in Figure 1 (a2) and (a3). The second loading contains the spectral features of CH_3, CH_2 and S_2, while the third loading includes the information of CH_3 and S_0 [10,11]. The spectral profiles obtained by PARAFAC are shown in Figure 1 (b1) to (b3). According to the assignments above, the first component describes the information of CH in ethanol molecules. In Figure 1 (b2), only the spectral information of S_0 is included. The third component illustrates the spectral features of both CH and S_0 similar with that in Figure 1 (a3). The profile may correspond to a component of water-ethanol heterocluster in the mixture. The three profiles captured by ATLD are displayed in Figure 1 (c1) to (c3). It is clear that the first two profiles are almost the same as those obtained by PARAFAC, but the third profile includes the spectral features of S_1 (water with one hydrogen bond) [1] and S_2. Therefore, spectral features obtained by NPCA explain the maximum variances, while the spectral profiles computed by PARAFAC and ATLD contain the spectral information of the components.

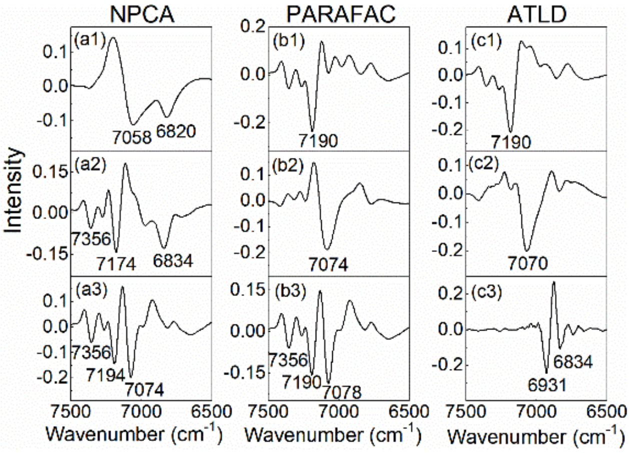

Figure 1. Loadings obtained by NPCA (a1–a3), PARAFAC (b1–b3) and ATLD (c1–c3).

According to the model of the three-way data array, two scores, along the dimensions of temperature and concentration can be obtained, which describes the variations of the spectra induced by temperature and concentration, respectively. The temperature scores obtained by NPCA, PARAFAC and ATLD show the relationship between temperature and the scores, respectively. The relationship can be defined as the QSTR model in our previous works [6-8], which can be used to predict temperature from the NIR spectra. The results indicate that all the three methods can extract the spectral variation induced by temperature from the dataset, and the variation of each loading changes almost linearly with temperature. The underlying causes of the QSTR model are the changes of the molecular structures with the temperature. The variation of the concentration scores obtained by NPCA, PARAFAC and ATLD reflect the spectral change induced by the content of ethanol in the mixture, and the underlying cause is the content variation of the component represented by the corresponding loading. Through analyzing the variations, the structural change can be revealed with the assistance of aquaphotomics and quantitative determination can be achieved. Therefore, high order chemometric algorithms may be the best way for analyzing the high dimensional temperature dependent NIR spectra.

References:

[1] R. Tsenkova, Introduction: Aquaphotomics: Dynamic spectroscopy of aqueous and biological systems describes peculiarities of water, J. Near Infrared Spectrosc. 17 (2009) 303–314.

[2] V.H. Segtnan, S. Sasic, T. Isaksson, Y. Ozaki, Studies on the structure of water using two–dimensional near–infrared correlation spectroscopy and principal component analysis, Anal. Chem. 73 (2001) 3153–3161.

[3] A.A. Gowen, J.M. Amigo, R. Tsenkova, Characterization of hydrogen bond perturbations in aqueous systems using aquaphotomics and multivariate curve resolution–alternating least squares, Anal. Chim. Acta 759 (2013) 8–20.

[4] X.Y. Cui, W.S. Cai, X.G. Shao, Glucose induced variation of water structure by temperature dependent near infrared spectra, RSC Adv. 6 (2016) 105729–105736.

[5] D. Cheng, W.S. Cai, X.G. Shao, Understanding the interaction between oligopeptide and water in aqueous solution using temperature-dependent near-infrared spectroscopy, Appl. Spectrosc. 0 (2018) 1–8.

[6] M. Li, X.Y. Cui, W.S. Cai, X.G. Shao, Understanding the function of water during the gelation of globular proteins by temperature dependent near infrared spectroscopy, Phys. Chem. Chem. Phys. 20 (2018) 20132–20140.

[7] X.G. Shao, J. Kang, W.S. Cai, Quantitative determination by temperature dependent near–infrared spectra, Talanta 82 (2010) 1017–1021.

[8] R.F. Shan, Y. Zhao, M.L. Fan, X.W. Liu, W.S. Cai, X.G. Shao, Multilevel analysis of temperature dependent near–infrared spectra, Talanta 131 (2015) 170–174.

[9] X.Y. Cui, X.W. Liu, X.M. Yu, W.S. Cai, X.G. Shao, Water can be a probe for sensing glucose in aqueous solutions by temperature dependent near infrared spectra, Anal. Chim. Acta 957 (2017) 47–54.

[10] X.G. Shao, X.Y. Cui, X.M. Yu, W.S. Cai, Mutual factor analysis for quantitative analysis by temperature dependent near infrared spectra, Talanta 183 (2018) 142–148.

[11] X.Y. Cui, J. Zhang, W.S. Cai, X.G. Shao, Chemometric algorithms for analyzing high dimensional temperature dependent near infrared spectra, Chemom. Intell. Lab. Syst. 170 (2017) 109−117.