Mutual factor analysis for quantitative analysis by temperature dependent near infrared spectra

Xueguang Shaoa,b,c,d,e, Xiaoyu Cui^b, Xiaoming Yu^f, Wensheng Cai^b

⁎ Corresponding author at: Xinjiang Laboratory of Native Medicinal and Edible Plant Resources Chemistry, College of Chemistry and Environmental Science, Kashgar University, Kashgar 844006, China.
E-mail address: xshao@nankai.edu.cn (X. Shao).

ARTICLE INFO

Keywords:
Near infrared spectroscopy
Temperature
Mutual factor analysis
Standardized signal
Glucose measurement
Serum

ABSTRACT

Temperature dependent near infrared (NIR) spectroscopy has been developed for analyzing multi–component mixtures and understanding the molecular interactions in solutions. In this work, a chemometric method named as mutual factor analysis (MFA) was proposed for the analysis of temperature dependent NIR spectra. The method extracts the common spectral feature contained in the spectra of different temperature or different concentration. The relative quantity of the extracted spectral feature is proportional to the temperature or concentration. From the spectra of water–glucose mixtures, both the spectral variations induced by temperature and concentration are obtained and the variations are correlated with the inducements, respectively, in a very good linearity. Serum samples were used for validation of the method. An acceptable calibration model with a good correlation coefficient (R² = 0.8639) was obtained for glucose measurement. The relative deviations of the measured concentrations from the calibration model are in the range of –18.7–8.52%, which are in a reasonable level for clinical uses. More importantly, the calculations are based on the spectral information of water that has interactions with the analyte. This provides a new way for quantitative analyses of bio–systems.

1. Introduction

Temperature dependent near infrared (NIR) spectroscopy has been developed as a new technique to analyze the aqueous–based biological or environmental samples. The structural information in a solution system can be obtained from the spectra measured at different temperatures [1,2]. A two–state structural model for water was reported by analyzing the NIR spectra measured at 6–80 °C. The spectral features of hydrogen–bonded and non–hydrogen–bonded water species were found changing distinctly with temperature [3,4]. Furthermore, the effects of solutes, such as salts [5,6], alcohols [7–9], glucose [10,11], proteins [12–14] on the NIR spectra of water were observed. In these studies, the hydration process of salts, the formation of water–ethanol heterocluster, the bioprotective role of glucose and the transition of protein conformation were discussed. Therefore, the behavior of water can serve as a mirror to reflect the interactions in aqueous systems using temperature dependent NIR spectroscopy.

Temperature effect of NIR spectra has been a problem for quantitative analysis. In earlier works, efforts were made to correct the effect on the calibration models. For example, the temperature compensation was reported to build the model using a combined calibration set including the spectra measured at different temperature [15]. Simulation study was also conducted to reduce the prediction error by incorporating the varied factors such as scattering effect and temperature in the calibration model [16]. A tensor–based calibration method using parallel factor (PARAFAC) analysis was reported for quantitative analysis of the NIR spectra of water–ethanol–propanol and water–ethanol–glycerine mixtures measured at different temperature [17]. In our previous works, the quantitative spectra–temperature relationship (QSTR) models were established for the mixtures of water, methanol, ethanol, and ethylenediamine and quantitative analysis of the composition of the mixtures was achieved based on the difference between the QSTR model of the different mixtures [18,19]. Multilevel simultaneous component analysis (MSCA) was also used to study the QSTR and
quantitative models for the ternary mixtures of water—ethanol—isopropanol and real serum samples [11,20]. In a recent work, a new QSTR model was established [21]. In the method, a parameter named as temperature—induced spectral variation coefficient (TSVC), which describes the overall spectral variation induced by temperature, was defined and a QSTR model was built with TSVC and normalized squared temperature. Afterwards, the quantitative calibration curve was obtained by the slopes of the QSTR models.

Chemometric methods are generally needed to extract the information hidden in NIR spectra. Principal component analysis (PCA) was applied to study the spectral features of water OH stretching at 75 °C and their variation with time for understanding the progress of amyloidogenic nucleation [22]. Evolving factor analysis (EFA) was used to analyze the temperature—dependent NIR spectra of bovine serum albumin (BSA) in aqueous solutions [23]. From the variation of the eigenvalues obtained in forward and backward EFA analysis, the hydration and secondary structural change of BSA with temperature was discussed. Two—dimensional (2D) correlation analysis was employed to investigate the order of spectral changes with temperature for comprehending the interaction between ethylene glycol and water [24]. Studies on hydrogen bonding using multivariate curve resolution—alternating least squares (MCR—ALS) [25] and independent component analysis (ICA) [26] were also reported. For the spectra of multi—samples measured at different temperature, however, the data is in a form of three or even higher dimensional array. In such cases, high order chemometric algorithms, such as multiway principal component analysis (NPCA) [27], parallel factor analysis (PARAFAC) [28] and alternating trilinear decomposition (ATLD) [29], are needed. The performance of the three algorithms in analyzing the high dimensional temperature dependent NIR spectra was investigated in our recent work [9]. It was indicated that both the spectral variations induced by temperature and concentration can be obtained. The former can be used for structural analysis and the latter can be used for building a calibration model for quantitative analysis. However, due to the complexity of the spectral changes, the result is still dependent on the algorithm and even the number of latent variables used in the calculation. It is also difficult to obtain the quantitative information directly from the scores, and the multiple linear regression is needed.

In this work, particular attention is paid on developing a new algorithm for analyzing high dimensional temperature dependent NIR spectral data. Combining the spectra measured at different temperature of different samples, a combined spectral matrix can be obtained. If PCA is performed on the combined data matrix, the difference between the samples or temperature can be reflected by the loadings. Taking the measurement of glucose in aqueous solutions, the method was tested, and the feasibility was proved by the obtained calibration models of both temperature and concentration. Determination of glucose content in real serum samples was also carried out.

2. Materials and spectral measurement

2.1. Reagents and sample preparation

Three groups of samples were prepared in this study. The first two groups, A and B, are composed of aqueous glucose solutions in different concentrations. Group A contains five samples with the concentration of 0, 20.00, 40.00, 60.00 and 80.00 g L\(^{-1}\), and group B includes six samples with the concentration of 0, 0.90, 1.20, 1.80, 2.40 and 3.00 g L\(^{-1}\). The former was used to validate the method, and the latter was designed to simulate the glucose contents in human serum. Glucose is of analytical grade and purchased from Concord Technology Co., Ltd. Tianjin, China. Double distilled water was used for preparation of the solutions. The third group, C, contains 22 human serum samples were collected from different patients by the clinical laboratory of People’s Hospital of Gaomi (Gaomi, China), and the glucose contents were measured by biochemical analysis in the laboratory. The serum samples were frozen at −20 °C to preserve their chemical composition and integrity until spectral measurement.

2.2. Temperature control and spectral measurement

The temperature in the experiment was controlled by a model 2216e temperature controller (Bruker Optics Inc., Ettlingen, Germany). The precision of the equipment for temperature control is ± 0.1 °C. For group A, the temperature changed from 30° to 70°C with a step of 5 °C. For group B and C, the temperature changed from 30° to 60°C with a step of 5 °C. The spectra were also measured at temperature 36, 37, 38 and 39 °C to test the performance of the method in a small temperature range. The temperature range may be more suitable for bio-sample systems to avoid the structural changes of proteins. The spectrum at each temperature was measured 30 min later when the temperature was changed.

All NIR spectra were measured from 4000 to 12,000 cm\(^{-1}\) by a Vertex 70 spectrometer (Bruker Optics Inc, Ettlingen, Germany), using 1 mm cuvette. A tungsten—halogen light source and InGaAs detector were used. The spectra are digitalized with ca. 2 cm\(^{-1}\) interval in Fourier transform. To increase signal to noise ratio, both air reference and the spectra were measured with scan number 64.

As examples, Fig. 1 shows the measured NIR spectra under three temperatures (30, 45 and 60 °C) of pure water, glucose solution with the concentration of 3.00 (solution I) and 80.00 g L\(^{-1}\) (solution II) and a serum sample. Only the spectra in the wavenumber range of 6000 cm\(^{-1}\) and 8000 cm\(^{-1}\) were plotted in the figure, because the peak around 5100 cm\(^{-1}\) is over the range of the spectrometer and absorption after 8000 cm\(^{-1}\) is relatively low. Clearly, only a broad peak around 6900 cm\(^{-1}\) can be found in the spectra, which includes the absorptions of water (in different structures), glucose, and even the proteins for the serum samples.

3. Algorithm and calculation

3.1. Continuous wavelet transform

Continuous wavelet transform (CWT) has been proved to be an efficient tool for processing analytical signals [30–32], and is generally used for enhancing the spectral resolution and removing the background and noise [8–11,13,14,19–21]. In this work, CWT was employed for resolution enhancement of the spectra. Symmlet filter with a vanishing moment 6 (‘Sym6’) was adopted and the scale parameter was set to 50. CWT with ‘Sym6′ filter is an approximate equivalence of the sixth derivative but smoothing can be achieved simultaneously [32].

![Fig. 1. Measured spectra of water (blue), glucose solution of 3.0 g L\(^{-1}\) (solution I, magenta), glucose solution of 80.0 g L\(^{-1}\) (solution II, green) and human serum (red) measured at 30(solid line), 45 (dot line) and 60 (dash line) °C, respectively.](image-url)
3.2. Mutual factor analysis

PCA has been the most commonly used technique to extract information and reduce dimensionality for a 2-dimensional or bilinear data matrix. PCA model of a data matrix, \( X \), can be expressed by:

\[
X = CS^T + E = TP^T + E
\]

(1)

where superscript ‘\( T \)’ denotes the mathematical operation of transposition; \( C \) and \( S \) are the matrices of concentration and spectra of the components in the samples; \( T \) and \( P^T \) represent the scores and the loadings of the principal components (PCs); and \( E \) contains the residuals that are not fitted in the model. For the data matrix of a sample measured at different temperature, the spectral features will be decomposed into the orthogonal loadings that account for as much variance as possible of the data matrix, and the corresponding scores are the weight of the loadings that can be further explained by the contribution of the spectral features to the data matrix.

If \( n \) samples were measured at the same temperature schedule, a combined spectral matrix can be formed as,

\[
X_{\text{comb}} = [X_1, X_2, \ldots, X_n]
\]

(2)

If PCA was performed on \( X_{\text{comb}} \), the following model can be obtained,

\[
X_{\text{comb}} = [X_1, X_2, \ldots, X_n] = T[P^T_1, P^T_2, \ldots, P^T_n] + E
\]

(3)

where \( X_i = TP^T_i \). Because the same \( T \) is used, and the differences between the sample is reflected in \( P^T_i \). Using the relationship, the quantity of the spectral pattern of \( X_i \) contained in \( X_j \) \((i \neq j)\) can be calculated. Using this idea, spectral space transformation (SST) was proposed for eliminating the spectral differences induced by the changes in instruments or measurement conditions [33,34], and a generalized standard addition method was developed for analyzing the specific analyte in real samples with complex matrices [35].

In this study, the spectral information of a reference sample represented by the model is defined as the standardized signal and denoted by \( SS \), i.e.,

\[
SS = TP^T_{\text{ref}} = X_{\text{ref}}(P^T_{\text{ref}})(P^T_{\text{ref}})^{-1}
\]

(4)

where the superscript ‘\( + \)’ denotes pseudoinverse of the matrix. Thus, the relative quantity of the standardized signal contained in each \( X_i \) can be obtained by,

\[
z_i = \text{trace}(X_iSS^*)
\]

(5)

It should be noted that, theoretically, any one of the \( n \) samples can be used as the reference, but generally the blank sample or the sample with lowest or highest concentration is used. Furthermore, because the spectra \( (X_i) \) of the samples with different concentration of the analyte are used in the combined spectral matrix, the standardized signal \( (SS) \) dominantly accounts for the information of the analyte, and thus \( z_i \) mainly reflects the spectral variation of the analyte.

Because the essence of the algorithm is to extract and compare the factor mutually contained in the spectral data of different samples, the algorithm is named as mutual factor analysis and abbreviated as MFA. There are two ways to construct the combined spectral matrix. One way is to take the spectra of a sample measured at different temperature as an \( X_i \), and the other way is to construct the \( X_i \) by the spectra of different samples measured at a temperature. In the former case, the parameter \( z_i \) (relative intensity) is an equivalence of the relative concentration of the samples, and the effect of temperature is contained in \( SS \). In the latter case, however, the parameter \( z_i \) shows the effect of temperature, and the information of concentration is included in \( SS \). In this work, the two cases are denoted as MFAc and MFAt, respectively.

4. Results and discussion

4.1. Spectral analysis and resolution enhancement

Fig. 1 shows the broad band around 6900 cm\(^{-1}\), which is mainly composed of the first overtone of the stretching vibrations (\( v_1 + v_2 \)) of OH in water [1]. The spectral information related to the absorption of NH and CH groups is also hidden in the broad band [14,23,36], but it is very small due to the weak absorption and low concentration. Comparing the intensity of the spectra for the four samples, it can be found that the absorbance decreases in an order of water, solution I, solution II and serum. This is apparently caused by the decrease of the water content in the samples with the increase of the solute content. On the other hand, the peak shifts to higher wavenumber with the rise of temperature. This indicates that water molecules undergo changes in structure and the relative content of the structures, particularly the changes in hydrogen bonding [1,10]. Generally, in the mixture-model theory water structure is described as an equilibrium mixture of different species, mostly two or five. In the two species model, water is composed of hydrogen–bonded (HB) and non–hydrogen–bonded (NHB) structures [3,4]. In the five species model, however, water is considered as the structural composition of \( S_0, S_1, S_2, S_3 \) and \( S_4 \) referring to the water molecule with none, one, two, three and four hydrogen bonds, respectively [1,9–11]. Even for the quantitative analysis, resolution enhancement is also needed to enlarge the difference between the spectra and to emphasize the effect of temperature.

To obtain the high resolution spectra, CWT with the wavelet filter of ‘Sym6’ was applied. The transform is an approximate equivalence of the sixth derivative [32]. Fig. 2 shows the transformed spectra corresponding to those in Fig. 1. Taking the advantage of the higher order derivative, narrow peaks with an almost zero baseline were obtained. Compared with the results by the first or second derivatives used in our previous works [8–11,14–18–20], the resolution of the transformed spectra is further improved.

Comparing the spectra of the samples, the profiles are still similar to each other, indicating that the main information of the spectra is from the water. The difference for the intensity, however, can be seen, particularly for the high concentration sample and the serum sample. Only

Fig. 2. Transformed spectra of water (blue), glucose solution of 3.0 g L\(^{-1}\) (solution I, magenta), glucose solution of 80.0 g L\(^{-1}\) (solution II, green) and human serum (red) measured at 30 (solid line), 45 (dot line) and 60 (dash line) °C, respectively.
A tiny difference can be seen between the spectra of water and solution I due to low concentration. More importantly, the effect of temperature is highlighted in the transformed spectra. The spectral change induced by temperature is even larger than that induced by the concentration. Furthermore, as shown in the inset of Fig. 2, the spectral information of the solutes can be found in the spectral range of 6050–6400 cm\(^{-1}\). The spectral intensity of water is near zero, a little difference can be seen between the spectra of water and solution I, and obvious difference can be found for the spectra of solution II and the serum sample. These spectral information should stem from the absorption of NH and CH groups \([14,23,36]\). Therefore, high order derivative using CWT technique may provide more information for analyzing the variations induced by temperature and concentration, and even for the quantitative analysis of the temperature dependent NIR spectra.

4.2. Quantitative analysis by temperature–induced variation

To quantitatively analyze the temperature effect on the NIR spectra, MFAt was employed. Firstly, the spectra of different samples measured at a temperature were constructed as \(X_i\) for the combined spectral matrix. Performing PCA on the matrix and taking the spectra measured at 30 °C as the reference, the standardized signal (SS) can be obtained using Eq. (4). Then, the relative quantity \(z_i\) of the standardized signal contained in each \(X_i\) was extracted according to Eq. (5). Because \(z_i\) is a reflection of the spectral variation induced by temperature, the relationship between the spectra and temperature should be obtained, similarly as the QSTR in our previous works \([11,18–20]\). The spectra of samples in group A was firstly used to test the feasibility of the method. Fig. 3(a) shows the result. Clearly, the relationship is a linear model with a very high correlation coefficient (R\(^2\)), which can be used for predicting temperature by the NIR spectra.

For further explanation of the standardized signal, SS is plotted in Fig. 3(b). Theoretically, the information in SS is extracted to measure the temperature–induced variation. Thus, it should be independent to the temperature. The intensity of the signal should be only related to the concentration of the samples. To validate the assumption, the intensity change at 7096 cm\(^{-1}\) with the concentration was plotted in the embedded graph. Inspiringly, a linear model with R\(^2\) being 0.9995 was obtained, indicating that the model can be used as a calibration equation for quantitative analysis. Such model was known as quantitative spectra–concentration relationship (QSCR) in our previous works \([11,18–20]\).

Fig. 3. Results obtained by MFAt for the spectra of the samples in group A. (a) Relationship between the relative quantity \((z)\) and temperature. (b) Standardized signal (SS) and (inset) the relationship between the intensity at 7096 cm\(^{-1}\) and glucose concentration.

To make the method more practical, the spectra of the samples in group B with low glucose concentration were analyzed by MFAt with the same calculation. Fig. 4 shows the results. It can be seen from Fig. 4(a) that a perfect QSTR model with R\(^2\) = 0.9998 was obtained, even for the temperature from 36° to 39°C with 1 °C increment. In Fig. 4(b), an acceptable calibration model with R\(^2\) = 0.9059 can still be established by the intensity change with the concentration, although the spectra in SS are almost coincident with each other due to the low concentration.

Fig. 4. Results obtained by MFAt for the spectra of the samples in group B. (a) Relationship between the relative quantity \((z)\) and temperature. (b) Standardized signal (SS) and (inset) the relationship between the intensity at 7096 cm\(^{-1}\) and glucose concentration.
4.3. Quantitative analysis by concentration-induced variation

In order to do the quantitative analysis based on the spectral variation induced by the analyte in the solution, MFAc was applied. In the calculation of MFAc, the spectra of a sample measured at different temperatures are taken as an $X_i$ to form the combined data matrix. Then, PCA was performed, and SS was calculated using the spectra of water as the reference. In this case, the relative quantity ($z_i$) of the standardized signal is a reflection of the concentration of the samples, and the effect of temperature on the spectra is contained in SS. Fig. 5 demonstrates the results for the spectra of group A samples with high glucose concentration. Clearly, Fig. 5(a) shows a very good linearity between parameter $z_i$ and the glucose concentration. Fig. 5(b) shows the extracted standardized signal and the relationship between the intensity and temperature. Obviously, the spectral variation induced by temperature is contained in SS. A very good linearity between the intensity and temperature ($R^2 = 0.9994$) was obtained for the QSTR model.

The spectra of the samples in group B with low glucose concentration were also studied using MFAc. The results are shown in Fig. 6. Compared with that in Fig. 4(b), a slightly better result of the calibration model is obtained in Fig. 6(a). The result suggests that the method may be a feasible way for blood glucose detection. In Fig. 6(b), the QSTR is still as perfect as those Figs. 3, 4, and 5. These results sufficiently prove that the NIR spectra of water are very sensitive to temperature, and the sensitivity can be used to analyze the changes in aqueous solutions. Water may be a promising probe for analyzing the aqueous systems like bio-fluids by temperature dependent NIR spectroscopy.

4.4. Glucose detection for human serum samples

To further investigate the feasibility of the method for real samples, the spectra of the human serum samples in group C were calculated by MFA and MFAc, respectively. Just like discussed above, the two ways of calculation produce similar result of the QSTR and the calibration model, but the calibration model obtained by MFAc is slightly better. The feasibility of the method for real sample analysis is proved by the results.

Fig. 7 shows the results obtained by MFAc for the spectra of the human serum samples. Clearly, the calibration model ($R^2 = 0.8639$) shown in Fig. 7(a) is acceptable for the detection of the glucose content in serum samples, although the model is slightly worse than that for the prepared solutions due to the complexity of biological samples. The relative deviations of each sample from the fitted line (model) are listed in Table 1. The relative deviations are between $-18.7\%$ and $8.52\%$, which is similar with the results obtained in our previous work [11]. Fig. 7(b) shows the extracted SS calculated using the spectra of the sample with lowest glucose concentration as the reference. Although the spectral information of CH and NH can be seen in the spectral range $6000–6500\text{cm}^{-1}$, the main profile is similar with that in Figs. 3–6 for the solution samples. It further confirms that the results are mainly obtained from the spectral information of water. Furthermore, the QSTR model, as shown in the inset, is as good as that for the solution samples.
Therefore, both the calibration model for quantitative analysis and the QST model can be obtained. With three groups of the prepared aqueous glucose solutions and human serum samples, the feasibility of the method for glucose detection was proved. Furthermore, this study also proves that changes in NIR spectra can be quantitatively analyzed taking advantage of the temperature effect. Water may be a promising probe for analyzing the changes in aqueous systems.

Acknowledgements

This study is supported by National Natural Science Foundation of China (Nos. 21475068 and 21775076).

References

[23] Y. Chen, Y. Ozaki, M.A. Czarnecki, Molecular structure and hydrogen bonding in


