ANALYTICAL SCIENCES JULY 2007, VOL. 23

Multivariate Estimation between Mid and Near-Infrared Spectra of Hexafluoroisopropanol–Water Mixtures

Liping ZHANG,*1,*2 Isao NODA,*3 Boguslawa CZARNIK-MATUSEWICZ,*4 and Yuqing WU*1†

*1 *Key Lab for Supramolecular Structure and Material of Ministry of Education, Jilin University, Changchun 130012, P. R. China*
*2 *Department of Foundation, Jilin Grain College, Changchun 130026, P. R. China*
*3 *The Procter & Gamble Company, 8611 Beckett Road, West Chester, OH 45069, USA*
*4 *Faculty of Chemistry, University of Wroclaw, F. Joliot-Curie 14, 50-383 Wroclaw, Poland*

Multivariate regression based on partial least squares (PLS2) was applied to estimating one spectral dataset from another set having an intrinsic relationship with each other. An estimation was successfully carried out between mid-infrared (IR) spectra in the range of 2980 – 3800 cm^{-1} and that of near-infrared (NIR) spectra in the range of 6000 – 7500 cm^{-1} for hexafluoroisopropanol (HFIP)–water mixtures. The result demonstrates that, after building a suitable regression model, not only NIR spectra, but also well-resolved IR spectra of HFIP–water mixture can be estimated properly in this way. The use of IR and NIR spectroscopy together with PLS2 regression will not only alleviate laborious and costly measurements, but also open a way to provide easier assignments of generally weak and highly overlapped NIR spectral bands.

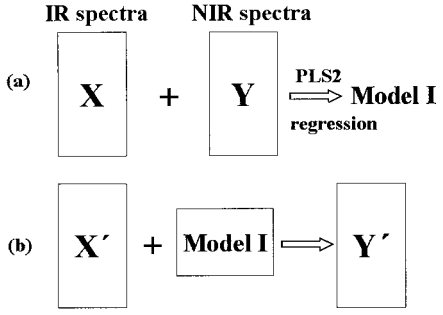
(Received March 26, 2007; Accepted April 16, 2007; Published July 10, 2007)

Introduction

Multivariate regression is a powerful chemometric method often used to predict (or estimate) the chemical components or the quality of samples based on the spectroscopic data.[1] Excellent applications of multivariate regression analysis can be found in a textbook for the quality prediction of various complicated practical systems, such as jam or peas, based on their spectral measurement.[1] Other examples of multivariate regression methods include the determination of the contents of free fatty acids and moisture in fish oils,[2] quantitative evaluation of the crystallinity of lactose in whey permeate powder,[3] and also the determination of energetic value of fruit and milk-based beverages.[4] The primary purpose of carrying out a multivariate regression analysis has been to replace laborious and costly measurements, such as sensory evaluation, by much less expensive and more straightforward instrumental measurements. In this paper, we describe another intriguing application of multivariate regression for the inter-conversion of spectral data.

By using multivariate regression methods, such as partial least squares (PLS2), it is also possible to obtain a model designed to estimate the corresponding intensity values of spectra used in the regression range. Starting with a known spectral data matrix, **X**, and another known spectral data matrix, **Y**, we can develop a multivariate regression model (as illustrated in Scheme 1(a)). Then the model can be applied to a new spectral data matrix, **X′**, to estimate the desired unknown spectra, **Y′**, corresponding to the type of data stored in **Y**, under the same or sometimes different physical conditions, such as temperature, pressure, or acidic value (as illustrated in Scheme 1(b)). In other words, by using multivariate regression methods, it is also possible to predict one spectrum from another. For example, Lew *et al.*[5] predicted NIR spectra from IR spectra, and Miller[6] attempted to predict Raman spectra from NIR spectra, although the results were still in a too rudimentary stage to be practical.

In practice, one often encounters a situation, such that the spectral measurement of **Y** is often expensive, difficult, labour intensive, time consuming, dangerous, or sometimes experimentally impossible compared with a measurement of **X**. The concept we are introducing here is quite simple, such that the matrix **Y**, which in traditional multivariate regression problem is typically composed of concentrations or chemical composition, flavor, viscosity, or some other quality, is now comprised of another set of spectral data that we are interested in. Based on the general rules of PLS2, it is noted that there must be a linear relationship between the two spectral matrices,



Scheme 1 (a) Regression, to make a model from known spectral matrix **X** and **Y**; (b) prediction, based on another known spectral matrix **X′** to predict the desired spectral matrix **Y′** by using the constructed regression model.

† To whom correspondence should be addressed.
E-mail: yqwu@jlu.edu.cn

X and **Y**, either in the original form or after a reasonable mathematical manipulation. For example, those of mid-infrared (IR) and near-infrared (NIR) spectra clearly satisfy such a requirement. Both types of spectra obey the Lambert–Beer rule with respect to the composition of the same sample. Therefore, good linearity should exist between the IR and NIR spectra. It should, in principle, be possible to predict the NIR spectra (or IR spectra) from a measured set of IR (or NIR) spectra by using an appropriately constructed regression model. Also, a good estimation can be achieved only after a good model based on the general rules of PLS2 has been built. More importantly, such an estimation scheme between the IR and NIR spectra may also open a possible way to obtain an easier assignment of generally weak and highly overlapped NIR spectral bands.

In the present work, such a scheme was successfully applied to a prediction between the mid-IR spectra in the range of 2980 – 3800 cm^{-1} and that of the NIR spectra in the range of 6000 – 7500 cm^{-1}, measured at different concentrations of hexafluoroisopropanol (HFIP)–water mixtures. This mixture system is of particular interest because of the following two reasons: one is that HFIP interacts with water in a specific way, forming micelles for a 30% (v/v) aqueous solution. Interestingly, the conformational changes in protein or peptides, from a β-sheet to an α-helical structure, are said to occur in a HFIP solution of such a volume fraction. The induced formation of α-helix in peptides and proteins in such aqueous solutions has long been studied in the biophysics and biochemistry fields.[7-9] The second reason for interest in this system is that, while HFIP may exist in two different conformations, there is no agreement about which conformation is the dominating one.

The evidence for microheterogeneity in HFIP–water mixtures comes primarily from small-angle X-ray scattering (SAXS),[10] wide-angle X-ray scattering (WAXS), NMR and mass spectrometry (MS) studies.[11] In the study of Kuprin *et al*.,[10] a sharp turn at 30% HFIP and another inflecting point at about 70% HFIP were found in the average scattering intensity dependent plot. Yoshida *et al*.[11] discovered that the structural transition of clusters occurs at a mole fraction of HFIP (χ_{HFIP}) between 0.06 – 0.1 (corresponding to 30 – 35% HFIP by volume). That is, the microheterogeneity reaches a maximum value at around that concentration. Moreover, they suggested that the inflection point in the ^{17}O-δ_{H_2O} values indicates another structural transition at χ_{HFIP} ~0.7 (corresponding to ~90% HFIP by volume). Meanwhile, by using PCA, we also explored the concentration-dependent interaction kinetics between HFIP and water; the results demonstrate a significant turn around 70% concentration (the result is not shown here). Although we have not fully determined that the structural variations occur at exactly 70%, it must be highly probable that the microheterogeneity of HFIP–water in the moderate concentration range is different from that both in the water-poor and water-rich range. Therefore, we focus our efforts on the moderate concentrational range, 35 – 65% of HFIP, which is relatively stable in microscopic components.

Experimental

Sample preparation

1,1,1,3,3,3-Hexafluoro-2-propanol was purchased from Aldrich. Appropriate volumes of alcohol and distilled water were mixed by shaking for half a minute at room temperature several hours before the measurement. Alcohol concentrations reported here are volume(%) = $volume_{alcohol}/(volume_{alcohol} + volume_{water})\cdot 100$, where volumes were measured before mixing.

Instrumentations

ATR/IR spectra of HFIP–water solution were measured at a 0.5-cm^{-1} resolution with a Nicolet Magna 760 FTIR spectrometer equipped with a DTGS detector. NIR spectra of the HFIP–water solutions were measured with an 1-cm^{-1} resolution by using the same spectrometer. To yield high a signal-to-noise ratio, 512 interferograms were coadded both for the IR and NIR measurements.

Both IR and NIR spectra were collected at room temperature (25˚C) for HFIP–water solution in different ratios. The spectra of pure water and pure HFIP were measured under the identical conditions for a comparison.

Data analysis

The spectra obtained by the ATR and NIR techniques for a series of HFIP–water solutions were employed to build a regression model by using PLS2. Before we carried out the calculation of PLS2, the raw spectra were subjected to a series of pretreatment to minimize undesirable effects. First, an ATR correction was performed to each individual IR spectrum.[12] Then, subtraction of the contribution of vapor water from each spectrum was performed.

After a pretreatment, the appropriately formatted IR and NIR data were exported to Unscrambler 6.0 for the PLS2 regression, validation and estimation performance.

Results and Discussion

IR and NIR spectra of HFIP–water mixtures

Figure 1 shows both the IR and NIR spectra of the HFIP/water mixtures at full concentration range. In the IR range (Fig. 1(a)) one can see three bands at 3628, 3591, and 3422 cm^{-1} for neat HFIP. The first two arise from the O–H stretching vibrations in monomers for *gauche*- and *trans*-conformers of HFIP, respectively.[13,14] Because upon dilution of HFIP by water, intensity of the band at 3591 cm^{-1}, assigned to the *trans*-conformer, decreases more rapidly than that at 3628 cm^{-1} due to the *gauche* conformer, probably the most stable one in the polar medium. At the same time, upon water dilution, three new bands with maxima at 3230, 3400, and 3672 cm^{-1} were developed. According to the two-state model of water structure,[15,16] the bands at 3230 and 3400 cm^{-1} could be assigned to the more and less structured water, respectively. The band around 3672 cm^{-1}, where the intensity noticeably increases in the range from 70 to 95%, then decreases from 65 to 35%, and finally disappears upon further dilution, could be attributed to the free OH groups of water molecules that are confined in the hydrophobic interior formed by the CF_3 groups. The assignment was confirmed by the sum-frequency generation (SFG) spectra of the air/water interface, where the peak at 3680 cm^{-1} has been attributed to OH stretches associated with dangling OH bonds.[17]

In the NIR spectra (Fig. 1(b)), two narrow bands observed at 7039 and 7109 cm^{-1} for neat HFIP are attributed, by analogy to the IR range, to the first overtone of the OH stretching vibration in monomers for *gauche*- and *trans*-conformers of HFIP, respectively. With dilution, in place of the former two bands, two others at 7158 and 6868 cm^{-1} arise. Because the intensity of that at a lower frequency increases with the water content, it should be attributed to the first overtone of the stretching vibrations of the hydrogen-bonded OH groups. The dilution-induced changes of the band at 7158 cm^{-1} have the same character as those of the band at 3672 cm^{-1}; therefore, it has been assigned to the first overtone of the free OH group trapped in the hydrophobic interior.

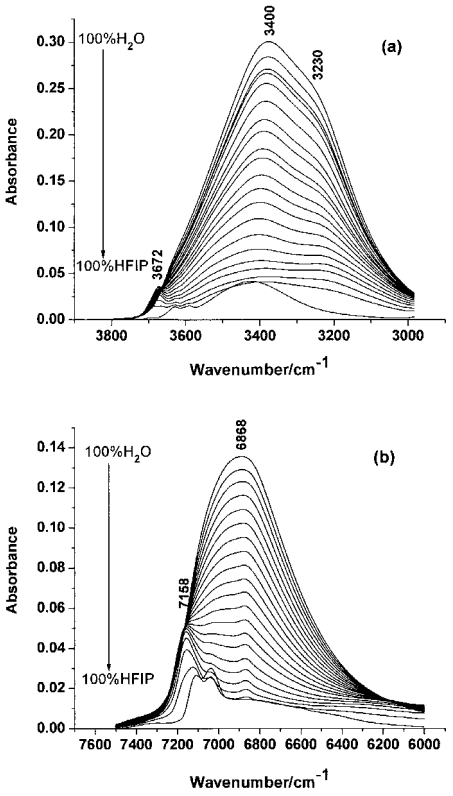


Fig. 1 (a) IR spectra in 2980 – 3800 cm^{-1}, and (b) NIR spectra in 6000 – 7500 cm^{-1} of HFIP–water mixtures in the full concentration range.

From the above discussions we can conclude that the dependence of the band shift on the concentration of HFIP in the NIR range is closely correlated with that in the IR range, and this consistent data structure is very important in applying PLS to a prediction of the infrared spectra.

Regression model building

In chemometrics, multivariate calibration modelling always concerns two matrices, an **X** and a **Y**. The **Y** matrix may, for example, consist of dependent variables, and the **X** contains the independent variables.[1] The multivariate model for **X** and **Y** is simply a regression relationship between the empirical (**X**, **Y**) relations,

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}.$$

This **B** matrix is calculated from

$$\hat{\mathbf{B}} = \mathbf{X}^{+}\mathbf{Y},$$

where $\mathbf{X}^{+}$ is an appropriate pseudo inverse of **X**. In a typical multivariate calibration case, **X** is a $k \times n$ matrix of spectral data comprised of k spectra with n wavenumber points, and **Y** is a $k \times r$ matrix with k different samples and each with r chemical responses, such as concentrations. In our case, both **X** and **Y** are spectral data matrices, such that each row corresponds to a spectrum.

We start with a known **X** and a known **Y** measured at 35, 45, 55% HFIP. From these data sets, we develop a multivariate regression model. The model is then applied to a newly measured matrix **X′** at 40, 50, 60% HFIP to predict the desired new **Y′**. In order to estimate IR and NIR spectral bands for HFIP–water mixtures, we establish a model by using a multivariate regression, PLS2, based on the concentration-dependent IR (2983 – 3800 cm^{-1} with 1694 variables as matrix **X**) and NIR (6009 – 7500 cm^{-1} with 1548 variables, serve as matrix **Y**) spectra. Since PLS is sensitive to the number of components (latent variables) in the models, and the introduction of a large number of components may significantly damage the performance of the method, the choice of the number of principal components is very important. The magnitude of the optimal number of latent variables in PLS2 is determined by a cross-validation technique.[18]

Table 1 Results of root mean square error of prediction (RMSEP)

	NIR	IR
RMSEP[a]	3.8×10^{-6} – 2.0×10^{-3}	7.4×10^{-6} – 5.3×10^{-3}
RMSEP[b]	2.0×10^{-3}	5.0×10^{-3}

a. Values for all the ranges.
b. Values for two specific bands at 6868 and 3400 cm^{-1}.

Validation of the predictive ability of the regression model

Validation is necessary to confirm the predictive ability of the model built by PLS2, or in other words, to check out how well the model will perform on new data. One of the methods that are available to estimate the prediction error is cross validation, with which the same samples are used both for model estimation and testing. A few samples are left out from the calibration data set and the model is calibrated on the remaining data points. Then, the values for the left-out samples are predicted and the prediction residuals are computed. The process is repeated with another subset of the calibration set, and so on, until every object has been left out once; then, all prediction residuals are combined to compute the validation residual variance and RMSEP (root mean square error of prediction). The simplest and most efficient measure of the uncertainty on future predictions is the RMSEP. This value (one for each response) means the average uncertainty that can be expected when predicting **Y**-values for new samples, expressed in the same units as the **Y**-variable. The results of RMSEP computed by the Unscrambler software are given in Table 1. It appears that the computed prediction errors are very small.

Prediction of NIR and IR spectra by using the regression model

We applied the model to a new matrix **X** (IR spectra data measured at 40, 50, 60% HFIP, 2983 – 3800 cm^{-1}) to estimate a new matrix **Y′**, and finally we could obtain the predicted NIR (6009 – 7500 cm^{-1}) spectra of 40, 50, 60% HFIP by plotting the data of matrix **Y′**. A plot of the predicted spectra is shown in Fig. 2.

In a similar manner, once we treat the NIR spectral data as the **X**-matrix, while IR spectral data as the **Y**-matrix, corresponding IR spectra can be estimated based on the models built by regression analysis (results are shown in Fig. 3). That is, by using the multivariate PLS2 regression model, it is possible not only to estimate NIR spectral data from the IR spectra, but it is also possible to estimate an IR spectral dataset from the NIR spectra. Such an inter-conversion of spectra would be very useful, especially for estimating well-resolved (but experimentally more difficult to measure) IR spectra from highly overlapped (but much easier to measure) NIR spectra for a practically important class of materials, such as biological samples, agricultural products, alcohol, and Chinese medicines.

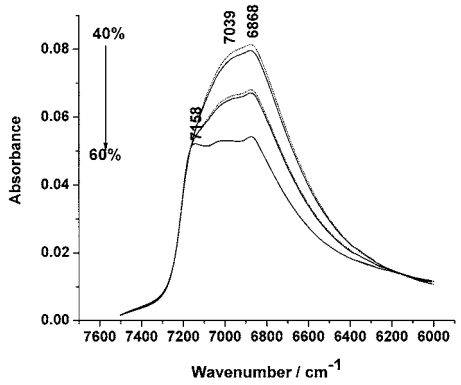


Fig. 2 NIR spectra predicted by PLS2 (dotted line) and measured by instrument (connected line) for the HFIP–water solution. The corresponding concentrations of HFIP are 40, 50, and 60%, respectively.

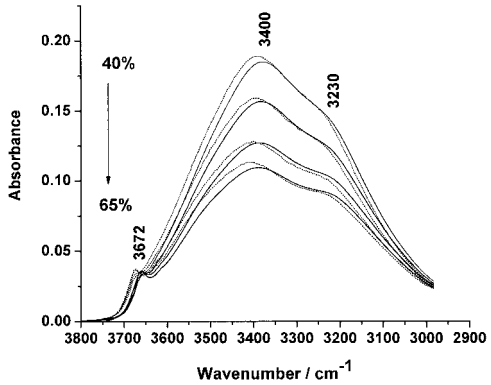


Fig. 3 IR spectra predicted by PLS2 (dotted line) and measured by instrument (connected line) for the HFIP–water solution. The corresponding concentrations of HFIP are 40, 50, 60, and 65%, respectively.

Furthermore, this technique will be very helpful in elucidating assignments of complicated NIR bands. From Fig. 3, it can be clearly seen that very close band features are achieved between the IR spectra based on the built model and the corresponding experimental measurements.

Judgment on the predicted results

Figure 2 shows the NIR spectra predicted by applying the above model to the experimental IR spectra data at 40, 50, and 60% HFIP and the corresponding ones that were measured by instrument. Even by a visual inspection it can be shown that the two spectra are very close regarding the peak shape, position and the absorbance intensity.

The **X**-data are used to predict, and the **Y**-data are used only to compare the predictions and the known **Y**-values. Plotting predicted *versus* measured reference values is a useful way to illustrate the validity of this procedure. Figure 4 plots (a) the NIR reference data against the IR predicted values; and (b) the IR reference data against the NIR predicted values in the independent validation set at two typical bands. Apparently, the predicted values and the measured ones are very close, and the errors of prediction are also evaluated by the MSEP using an equation reported in the literature;[18] the results are given in Table 2.

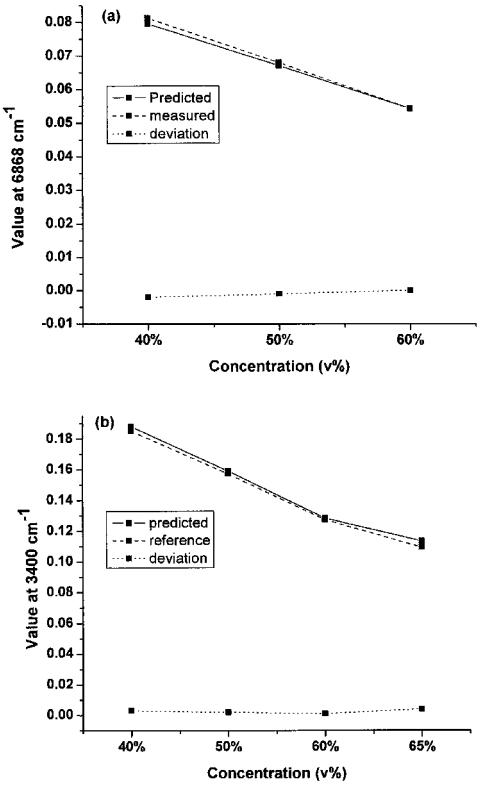


Fig. 4 Plot of (a) NIR predicted values against the measured NIR spectral values at a wavenumber of 6868 cm^{-1}, and (b) IR predicted values against the measured IR spectral values at a wavenumber of 3400 cm^{-1}.

Table 2 Mean square error of the prediction (MSEP)

Predicted NIR	Predicted IR
1.4×10^{-6}	7.5×10^{-6}

Conclusions

This work demonstrates the application of PLS2 for estimating spectral dataset of HFIP–water mixtures between IR and NIR. It is a simple and easy technique that can be used to estimate any spectral dataset where an intrinsic relationship exists among them. The results demonstrate that, after building a suitable model, not only the NIR spectra, but also well-resolved IR spectra of HFIP–water mixtures can be properly estimated in this way. It also illustrates that the use of IR and NIR spectroscopy together with multivariate techniques (PLS2 regression) can alleviate laborious and costly measurements, and also provide easier assignments of generally weak and highly overlapped NIR spectral bands.

Acknowledgements

The present study was supported by the Project of NSFC (Nos. 20373017, 20473028), the Major State Basic Research Development Program (2007CB808000), Program for New Century Excellent Talents in University (NCET), the Program for Changjiang Scholars and Innovative Research Team in University (IRT0422) and the 111 project (B06009), which are gratefully acknowledged.

References

1. "*Multivariate Analysis in Practise*," ed. K. Esbensen, **1994**, Wennbergs Trykkeri As, Trondheim, Norway.
2. D. Cozzolino, I. Murray, A. Chree, and J. R. Scaife, *Food Sci. and Tech./LWT*, **2005**, *38*, 821.
3. L. Nørgaard, M. T. Hahn, L. B. Knudsen, I. A. Farhat, and S. B. Engelsen, *Int. Dairy J*., **2005**, *15*, 1261.
4. J. Moros, F. A. Iñón, S. Garrigues, and M. D. L. Guardia, *Anal. Chim. Acta*, **2005**, *538*, 181.
5. R. Lew and S. T. Balke, *Appl. Spectrosc*., **1993**, *47*, 1747.
6. C. E. Miller, *Spectrochim. Acta, Part A*, **1993**, *49*, 621.
7. D. Hong, M. Hoshino, R. Kuboi, and Y. Goto, *J. Am. Chem. Soc*., **1999**, *121*, 8427.
8. M. O. Buck, *Rev. Biophys*., **1998**, *31*, 297.
9. S. J. Wood, B. Maleeff, and R. Wetzel, *J. Mol. Biol*., **1996**, *256*, 870.
10. S. Kuprin, A. Gräslund, A. Ehrenberg, and M. H. J. Koch, *Biochem. Biphys. Res. Commun*., **1995**, *217*, 1151.
11. K. Yoshida, T. Yamaguchi, T. Adachi, T. Otomo, D. Matsuo, T. Takamuku, and N. Nishi, *J. Chem. Phys*., **2003**, *119*, 6132.
12. B. Czarnik-Matusewicz, K. Murayama, Y. Wu, and Y. Ozaki, *J. Phys. Chem. B*, **2000**, *104*, 7803.
13. A. J. Barnes and J. Murto, *J. Chem. Soc., Faraday Trans. 2*, **1972**, *68*, 1642.
14. H. Schaal, T. Häber, and M. A. Suhm, *J. Phys. Chem. A*, **2000**, *104*, 265.
15. G. W. Robinson and C. H. Cho, *Biophys. J*., **1999**, *77*, 3311.
16. B. Czarnik-Matusewicz, S. Pilorz, and J. P. Hawranek, *Anal. Chim. Acta*, **2005**, *544*, 15.
17. Y. R. Shen, *Proc. Natl. Acad. Sci. U. S. A*., **1996**, *93*, 12104.
18. E. Vigneau, M. F. Devaux, E. M. Qannari, and P. Robert, *J. Chemom*., **1997**, *11*, 239.