Preprints of the 10th IFAC International Symposium on Dynamics and Control of Process Systems
The International Federation of Automatic Control
December 18-20, 2013. Mumbai, India

Extraction of pure component spectrum from mixture spectra containing a known diluent

Abhishek K Baikadi [*] **Mandeep Kaur** [**] **Sreeja S** [***]
Guhan Jayaraman [****] **Shankar Narasimhan** [†]

[*] *Department of Chemical Engineering (email: abhishek.baikadi@gmail.com)*
[**] *Department of Bio Technology (email: mandeepkalsi5@gmail.com)*
[***] *Department of Bio Technology (email: s.sreejadoss@gmail.com)*
[****] *Department of Bio Technology (email: guhanj@iitm.ac.in)*
[†] *Department of Chemical Engineering (email: naras@iitm.ac.in)*
Indian Institute of Technology Madras, Chennai 600036, India.

Abstract: Multivariate data analysis techniques are widely used in getting better insight into the processes in the fields like chemometrics, speech processing, biomedical signal processing and astronomy. In the present study, the problem of extracting the spectrum of a pure component from Near Infrared (NIR) mixture spectra containing a known diluent is tackled. Different multivariate data analysis methods such as Ordinary Least Square (OLS), Principal Component Regression (PCR) and Non Negative Matrix Factorization (NMF) are modified to solve the problem. It is shown that including partial knowledge such as the spectrum of the known diluent in the data analysis techniques, accounting for errors in the absorbance measurements, and imposing non-negativity constraints on absorbance and concentrations estimates, results in better estimation of the pure component spectrum.

Keywords: Multivariate Analysis, NIR Spectra, Pure Spectrum, Diluent Spectra, Fermentation process, OLS, PCR, NMF.

1. INTRODUCTION

Near infra-red spectroscopy (NIRS) is a widely used spectroscopic technique as it offers advantages over its counterparts like Raman spectroscopy and mid infra-red spectroscopy (MIRS) in the context of online monitoring of industrial processes. The major advantages of NIR is the possibility of remote sensing using optic fibres and the ability to record absorbance spectra without any sample preparation (Siesler, 2007). Another advantage of NIR is its ability to measure the absorbance of varying density solutions and large sized particles like biomass. Despite these advantages NIR was not popular until the advancements of multivariate data analysis techniques. Unlike other spectroscopic techniques, absorbance data obtained from NIR does not contain absorbance peaks characteristic unique to each constituent in the mixture which makes qualitative analysis of the mixture difficult. Multivariate data analysis methods help in comprehending NIR data and quantifying the components in the sample (Tosi et al., 2003).

NIR measures the frequency (wavelength) and intensity (Absorbance) of bond vibrations. Overtones and combinations of fundamental molecular vibrations are measured in the wavelength range 800nm-2100nm, and molecules containing C-H, N-H, O-H, S-H bonds are strong absorbers (Celio, 2003). Depending on the energy provided by the light source, the molecules undergo molecular vibrations to varying energy levels, thereby causing a dip in the light intensity at particular wavelengths which is measured as absorbance at that particular wavelength.

2. MOTIVATION

In anaerobic bacterial fermentation, glycolysis is the fundamental pathway involved in energy production. During glycolysis the carbon source taken up by the microorganisms is converted to pyruvate which is further converted to products like acetate, lactate, formate and ethanol (Murray et al., 2006). Since a substantial proportion of the carbon source is directed to glycolysis, quantification of the end products of glycolysis helps in understanding the fate of glucose inside the cell and developing methods to divert the carbon flux from glycolysis to other desired pathways resulting in the product of interest. Since the intermediates and products involved in this pathway absorb in the near infrared range, it may be possible to use NIR absorbance spectra to monitor the fermentation reaction and estimate the change in the concentration of these species as the reaction proceeds. Based on these concentrations, it may ultimately be possible to develop a comprehensive metabolic flux analysis model. As a first step in this attempt, it is necessary to develop a multivariate calibration model between the NIR spectra and concentration of the species present in the fermentation mixture.

In a typical fermentation process, the mixture may contain hundreds of different species all of which absorb in the NIR range of interest. However, many of these are present in extremely low concentrations and may be treated as noise. About ten species are of interest whose concentrations we would like to track as the reaction proceeds. In order to develop a multivariate calibration model, it would be advantageous to obtain the pure NIR absorbance spectrum of each of these species. It is not possible to obtain pure solutions of all the species and they are available only as aqueous solutions. For example, aqueous solutions of glucose in water need to be prepared for measuring the NIR spectrum. Unfortunately, water also has an absorbance spectrum in the NIR range of interest and it is al dominant spectrum. Thus the problem is to extract the pure species spectrum (in this case glucose) by removing the effect of diluent spectrum (in this case, water) from the mixture spectra containing the diluent and the species of interest. In this work, different approaches to remove or eliminate the influence of NIR spectrum of the diluent (water) from the solution spectra are developed and assessed for their accuracy. The specific system used in this work is to obtain NIR spectrum of glucose from NIR absorbance measurements of dilute aqueous solutions of glucose.

3. MULTIVARIATE DATA ANALYSIS TECHNIQUES

According to Beer-Lamberts law (Beer, 1852), the absorbance of a dilute mixture is a linear combination of the pure species absorbance. The absorbance spectrum of N mixtures containing different concentrations of a diluent and a species of interest measured at n wavelengths can be grouped in a data matrix Z, where each row represents the spectrum of a mixture. The relation between the mixture and pure species spectrum can be written as

$$Z = CS + E \tag{1}$$

C is a $N \times 2$ molar fraction matrix with first column corresponding to mole fraction of pure species and second column for diluent. S is $2 \times n$ pure spectrum matrix. First row of S corresponds to the spectrum of the pure species of interest and the second row is the spectrum of the pure diluent. The matrix E represents the errors in measurements of the mixture spectra.

3.1 Ordinary Least Square Analysis

If the spectrum of the diluent is known and the concentration of the different mixtures are also specified, then the standard method for determining the pure species spectrum is to eliminate the effect of diluent from the different mixtures (Billeter et al., 2009).

The adjusted measurements after removal of the diluent spectrum can be written as

$$\bar{Z} = Z - C_{.2}S_{2.} \tag{2}$$

where, $C_{.j}$ implies j^{th} column of C matrix and $S_{i.}$ the i^{th} row of S matrix. In obtaining Eq. (2), it is implicitly assumed that both the concentration and the absorbance measurement of the diluent are noise free.

The pure species spectrum can be obtained from the adjusted measurements using a ordinary least squares (OLS) or partial least squares (PLS) regression. For example, if OLS is used, then the pure species spectrum can be estimated as

$$\hat{S}_{1.} = (C_{.1}^T C_{.1})^{-1} C_{.1}^T \bar{Z}^T \tag{3}$$

Alternatively PLS can also be used to estimate the pure species spectrum (Ham et al., 1997).

3.2 Principal Component Regression

When the measured data is adjusted by using the spectrum of the diluent in the standard approach, it is implicitly assumed that the diluent spectrum is perfectly known and is free of errors. However, since the diluent spectrum is also measured using the same instrument, it will also contain same errors, which should be taken into account. We propose a modified Principal Component Regression (PCR) method which extract the pure species spectrum while simultaneously denosing the diluent spectrum.

The data matrix is first augmented using the diluent spectrum. The augmented matrix is given by

$$\tilde{Z} = \begin{bmatrix} Z \\ S_{2.} \end{bmatrix} \tag{4}$$

The concentration matrix is also augmented in a corresponding manner as follows

$$\tilde{C} = \begin{bmatrix} C_{.1} & C_{.2} \\ 0 & 1 \end{bmatrix} \tag{5}$$

Principal Component Analysis (PCA) is applied to the augmented data matrix and the scores corresponding to the first two Principal Components (PC) corresponding to the two largest eigenvalue values are obtained. For this purpose, the singular value decomposition of the augmented data matrix is computed as

$$\tilde{Z} = U_1 D_1 V_1^T + U_2 D_2 V_2^T \tag{6}$$

where U_1 and V_1 are the singular vectors corresponding to the first two largest singular values, and D_1 is a diagonal matrix containing the corresponding singular values (square root of eigenvalues). The first term in the right hand side of above equation can also be regarded as the denoised spectra of all mixtures including the diluent spectrum.

The scores matrix corresponding to the first two PCs is given by

$$T = \tilde{Z} V_1 \tag{7}$$

The second step is to use OLS regression to relate the concentrations to the scores matrix as follows

$$\tilde{C} = TB + \varepsilon \tag{8}$$

Two possible solutions to the regression matrix B can be obtained depending on the assumptions made. If we make the standard assumption that the scores are free of errors, but the concentration measurements contain errors, then the regression matrix B is given by

$$B = (T^T T)^{-1} T^T \tilde{C} \tag{9}$$

If on the other hand, we assume that the concentrations are free of errors and the scores contain errors, then the regression matrix is given by

$$B = \left[(\tilde{C}^T \tilde{C})^{-1} \tilde{C}^T T\right]^{-1} \tag{10}$$

In order to obtain estimates of the pure species spectrum, we define a mixture whose concentration vector is $C_{pure} = [1 \quad 0]$. This vector represents the concentration of a mixture containing only the pure species, without the diluent. The scores corresponding to this concentration are obtained using the PCR regression model as

$$T_{pure} = C_{pure} B^{-1} \tag{11}$$

The pure species spectrum is estimated by

$$S_{pure} = T_{pure} V_1^T \tag{12}$$

We refer to the use of Eq. (9) to obtain the regression matrix as PCR-S method, and the use of Eq. (9) for deriving the regression matrix as the PCR-C method.

3.3 Non Negative Matrix Factorization

The approach described in the preceding two subsections will not ensure that the estimated spectrum contains non-negative absorbances for all wavelengths. This gives rise to physically unrealistic spectrum which have to be corrected using ad-hoc measures such as scaling the estimated absorbances between 0 and 1. In order to overcome this problem, methods such as non-negtave matrix factorization can be used.

Non-negative matrix factorization (NMF) is a matrix decomposition approach that factorizes a non-negative matrix into two low-rank non-negative matrices (Lee and Seung, 1999). It has made great success in biological data mining. In chemometrics, NMF has a long history under the name Self modelling Curve Resolution (Lawton and Sylvestre, 1971). Also early work on non-negative matrix factorizations was performed by a Finnish group of researchers in the middle of the 1990's under the name positive matrix factorization (Paatero and Tapper, 1994; Anttila et al., 1995). It became more widely known as non-negative matrix factorization after Lee and Seung (1999) investigated the properties of the algorithm and published some simple and useful algorithms for two types of factorizations.

Given the absorbance measurements (always positive) of mixtures containing a known number of species, NMF can be used to obtain both the relative concentrations of species in the different mixtures as well as the pure species spectrum. Furthermore, the method ensures that the estimated concentrations and absorbances are non-negative. The use of NMF to estimates pure species spectrum from mixture spectra is also known as the blind source separation. It may be noted that NMF does not use any knowledge of the mixture concentrations or the pure component spectrum in deconvolution of the mixtures.

In this work, the augmented mixture matrix (Eq. 4) is subjected to PCA and first two PC's are extracted. These PC's are used in NMF algorithm available in MATLAB NMF Toolbox to extract the source spectrum for the diluent ($S_{2.}$) as well as pure species ($S_{1.}$). The inclusion of the diluent absorbance spectra in the data matrix, utlizes all the available information in the analysis, except the measured concentration of the diluent.

The different approaches for extraction of pure spectrum of glucose are evaluated using NIR absorbance measurements of different concentrations of aqueous solutions of glucose

4. MATERIALS AND METHODS

Glucose powder was procured from HiMedia (Mumbai, India). Solutions were made in double distilled water to avoid the influence of any other substrates in present in water. Around 6 - 7 mixtures of varying concentration starting as low as 0.01 g/l to 50 g/l were prepared. Foss Near Infra Red Systems (Model 6500) was used to obtain the NIR absorbances of the mixtures at path length of 1 mm. Absorbance measurements in the range 800-2200 nm were made at 0.5 nm intervals. Absorbance spectrum of pure water and a 60 % glucose aqueous solution is shown in Fig. 1.

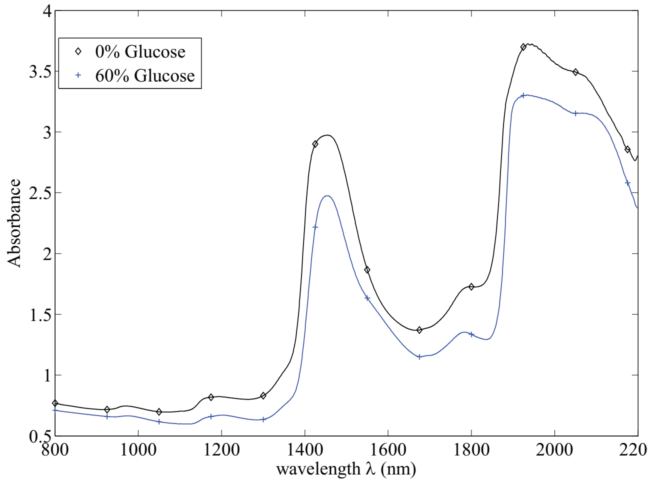

Fig. 1. NIR Spectra of Glucose diluted in Water

5. RESULTS AND DISCUSSION

The different multivariate data analysis methods described in section 3 were applied to the NIR absorbances measurements to obtain the pure spectrum of glucose. The estimated glucose spectrum by different methods is shown in Fig. 2. The PCR and NMF approaches also provide an estimate of water spectrum. The experimental measured water spectrum as well as those estimated using PCR and NMF are shown in Fig. 3. For comparison, the spectrum of water and glucose estimated by Ham et al. (1997) using a Partial Least Squares (PLS) method on NIR data of aqueous mixtures is shown in Fig. 4. In these figures, the spectra are scaled such that the maximum absorbance is unity.

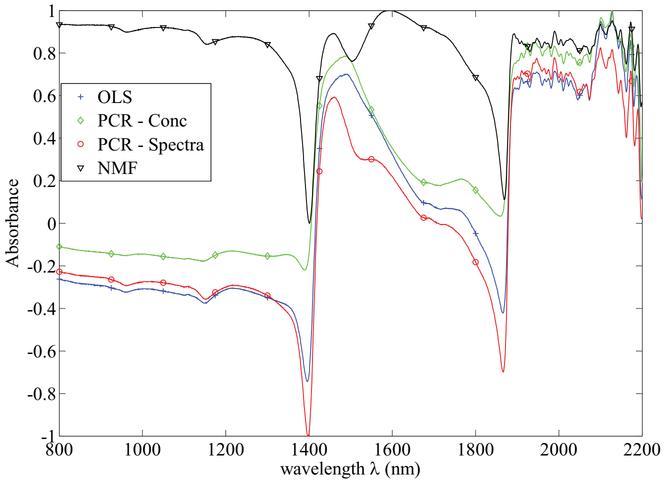

Fig. 2. Pure Spectra of Glucose

From Fig. 2 it is observed that the estimated spectrum of glucose using all methods except NMF have negative absorbances in the wavelength range 800-1400 nm. NMF ensures that the estimated absorbance for the entire wavelength range is non-negative. A comparison of the glucose spectrum with that obtained by Ham et al. (1997)(Fig.4) shows that the spectrum obtained using NMF has good qualitative agreement in the common range 1500-2200 nm.

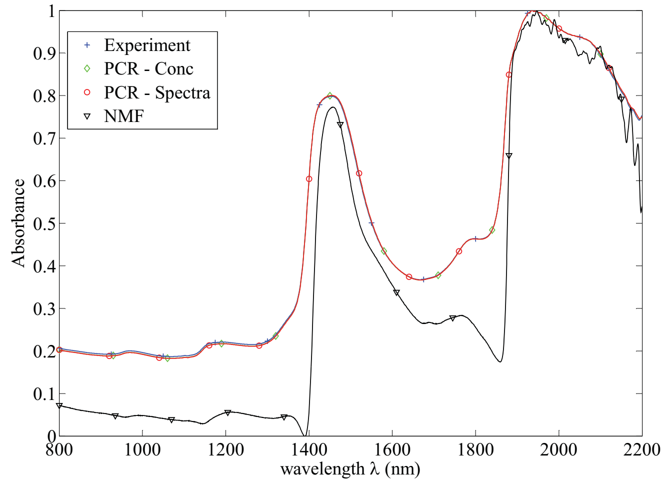

Fig. 3. Pure Spectra of Water

The performance of different approaches is quantitatively assessed by using Leave-one-out-cross-validation (LOOCV) procedure. In this procedure, the measurements corresponding to one of the mixtures is dropped (Validation data) and the method is applied using the remaining mixture measurements(Training data). The concentration of glucose in the mixture left out is predicted using its absorbance spectrum and compared with the concentration used in preparing the mixture.This procedure is repeated such that each observation in the sample is used once as the validation data. The average percentage error in the predicted concentration obtained using different methods are shown in Table 1.

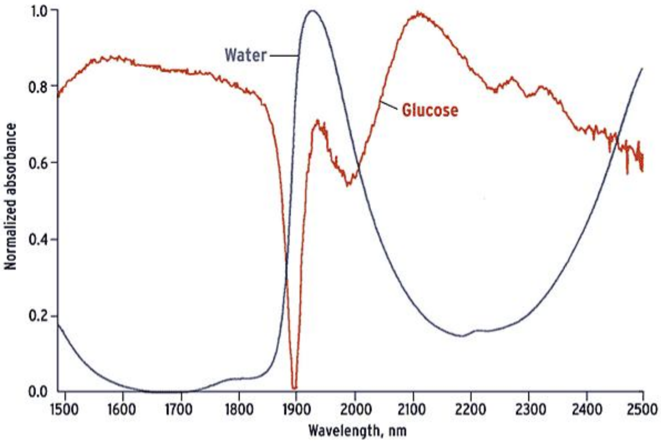

Fig. 4. Pure Spectra of Glucose and Water obtained by Ham et al. (1997)

In Table 1, ”% Error in 0.01g/l” signifies the error in predicting the concentration for a mixture spectrum of concentration 0.01g/l glucose using the pure spectra of glucose and water obtained leaving out 0.01g/l glucose spectrum in estimation. It can be seen that error in predicting lower concentration is high in comparison to higher concentration. This is due to the fact that the effect of lower glucose is masked by water spectrum.

The error in predicting 1g/l (100mg/dl) of glucose is in coherence to Ham et al. (1997), which is typical blood sugar concentration.

Table 1. Error in predicted Glucose concentration using different multivariate calibration methods

	OLS	PCR - C	PCR - S	NMF
% Error (average)	14.14	7.37	8.18	7.45
% Error in 0.01g/l	23.41	16.57	17.36	15.32
% Error in 1g/l	19.13	8.87	9.02	7.63
% Error in 60g/l	8.2	5.4	5.32	4.88

The results show that as compared to the standard OLS method, the proposed approaches provide more accurate estimates. The reason for this can be attributed to the use of denoised water spectrum in the proposed methods in contrast to the OLS method which assumes that the measured water spectrum is exact. PCR-C (PCR assuming no error in concentration) has a better results over PCR-S (PCR assuming no error in scores) as variability in measurement of spectra is higher in comparison to sample preparation. Among the proposed methods the performance of NMF is marginally better than the others methods

6. CONCLUSION

Improved methods for extracting a pure species spectrum from mixture absorbances containing a known diluent are proposed in this work. The importance of denoising the diluent spectrum in improving the concentration predictions is brought out. Among the proposed methods NMF provides physically realistic spectrum of the pure species and also provides marginally more accurate estimates of the concentrations. It may be possible to further improve the performance of NMF by utilizing the available information regarding the mixture concentration in the analysis. Prediction may also improve considering the a short range of wavelength where the effect of particular bond vibration is significant.

REFERENCES

Anttila, P., Paatero, P., Tapper, U., and Jrvinen, O. (1995). Source identification of bulk wet deposition in finland by positive matrix factorization. *Atmospheric Environment*, 29(14), 1705 – 1718. URL http://www.sciencedirect.com/science/article/pii/135223109400367T.

Beer, A. (1852). Bestimmung der absorption des rothen lichts in farbigen flssigkeiten. *Annalen der Physik*, 162(5), 78–88. URL http://dx.doi.org/10.1002/andp.18521620505.

Billeter, J., Neuhold, Y.M., and Hungerbhler, K. (2009). Systematic prediction of linear dependencies in the concentration profiles and implications on the kinetic hard-modelling of spectroscopic data. *Chemometrics and Intelligent Laboratory Systems*, 95(2), 170 – 187. URL http://www.sciencedirect.com/science/article/pii/S0169743908001974.

Celio, P. (2003). Near infrared spectroscopy: Fundamentals, practical aspects and analytical applications. *J. Braz. Chem. Soc.*, 14, 198–219. URL http://dx.doi.org/10.1590/S0103-50532003000200006.

Ham, F., Kostanic, I., Cohen, G., and Gooch, B. (1997). Determination of glucose concentrations in an aqueous matrix from nir spectra using optimal time-domain filtering and partial least-squares regression. *Biomedical Engineering, IEEE Transactions on*, 44(6), 475–485. URL http://dx.doi.org/10.1109/10.581938.

Lawton, W.H. and Sylvestre, E.A. (1971). Self modeling curve resolution. *Technometrics*, 13(3), 617–633. URL http://www.tandfonline.com/doi/abs/10.1080/00401706.1971.10488823.

Lee, D.D. and Seung, H.S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755), 788–791. URL http://dx.doi.org/10.1038/44565.

Murray, R.K., Granner, D.K., Mayes, P.A., and Rodwell, V.W. (2006). *Harper's Illustrated Biochemistry (Harper's Biochemistry)*. McGraw-Hill Medical, 27 edition. URL http://www.worldcat.org/isbn/0071461973.

Paatero, P. and Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2), 111–126. URL http://dx.doi.org/10.1002/env.3170050203.

Siesler, H.W. (2007). *Introduction*, 1–10. Wiley-VCH Verlag GmbH. URL http://dx.doi.org/10.1002/9783527612666.ch01.

Tosi, S., Rossi, M., Tamburini, E., Vaccari, G., Amaretti, A., and Matteuzzi, D. (2003). Assessment of inline near-infrared spectroscopy for continuous monitoring of fermentation processes. *Biotechnology Progress*, 19(6), 1816–1821. URL http://dx.doi.org/10.1021/bp034101n.