Analytical Methods

RSC Publishing

PAPER

View Article Online
View Journal | View Issue

Cite this: *Anal. Methods*, 2013, **5**, 891

Determination of the empirical solvent polarity parameter $E_T(30)$ by multivariate image analysis

Fatemeh Shakerizadeh-Shirazi,[a] Bahram Hemmateenejad*[a] and Abdol Mohammad Mehranpour[b]

Received 9th August 2012
Accepted 26th November 2012

DOI: 10.1039/c2ay26185e

www.rsc.org/methods

The solvent polarity $E_T(30)$ scale has found wide-spread applications in studying chemical processes in solvents. This parameter is usually measured by vis spectrophotometric measurements of the long-wavelength intramolecular charge-transfer (CT) absorption band of Reichardt's pyridinium-*N*-phenolate betaine dye, *e.g.* the $E_T(30)$ dye, dissolved in the solvent or solvent mixture of interest. Recent advances in colorimetric measurements based on digital photo-capturing devices suggest these methods as a simple, cheap and fast alternative to spectrophotometric measurements in some analytical applications. In this work, we studied the feasibility of colorimetric measurements coupled with multivariate data analysis to determine the empirical solvent polarity parameter $E_T(30)$. The picture of the $E_T(30)$ dye dissolved in different solvents was captured by a digital camera and then color values in the RGB space were analyzed by the principal component analysis (PCA) method. PCA scores of the unfolded image were then used as input of multiple linear regression and an artificial neural network model to predict the $E_T(30)$ parameter. The ANN models were optimized to gain a model of lower prediction ability utilizing a cross-validation test. Then, this was used to predict $E_T(30)$ values for an external solvent test set. The generated model could explain and predict 99% of the variances in the polarity data and can predict $E_T(30)$ values with a root mean square error of 2.25 kcal mol^{-1} (in the $E_T(30)$ scale). The results suggest colorimetric measurements as a useful and practical alternative to the vis spectrophotometric measurements for determination of solvent polarity parameters derived from solvatochromic betaine dyes.

Introduction

Nowadays it is well known that solvents can have a strong influence on the position of chemical equilibria, on reaction rates, as well as on the position and intensity of spectral absorption bands (*e.g.* UV/Vis, IR, EPR, and NMR). Therefore, the selection of a proper solvent, suitable for the reaction under study, is of paramount importance for the success of this reaction. The effects can be explained in terms of 'solvent polarity'.[1] Solvent polarity is much better described by molecular–microscopic empirical solvent parameters derived from suitable solvent-dependent reference processes. In this case, the individual solvent molecules surrounding the ions or dipoles of the reference solute are arranged to a loose or tight solvation shell.

In the last 50 years, numerous solvent polarity scales have been proposed.[2] By virtue of their exceptionally large negative solvatochromism, zwitterionic pyridinium *N*-phenolate betaine dyes were introduced by Reichardt *et al.* in 1963 as solvent polarity indicators, which overcame some practical limitations of other solvatochromic dyes.[3] Since their long-wavelength solvatochromic absorption band lies within the visible region of the spectrum, even a visual estimate of solvent polarity can often be made with these betaine dyes.[4] The $E_T(30)$ values, empirically derived from solvatochromic measurements, are simply defined as the molar transition energies (in kcal mol^{-1}; 1 kcal = 4.184 kJ) of the standard betaine dye (Fig. 1), measured in solvents of different polarity at room temperature (25 °C) and normal pressure (1 bar), according to eqn (1):

$$E_T(30)\ (\text{kcal mol}^{-1}) = 28\,591/\lambda_{\max}\ (\text{nm}) \qquad (1)$$

where $\lambda_{\max}$ is the wavelength of the maximum of the long-wavelength intramolecular CT absorption band of the betaine

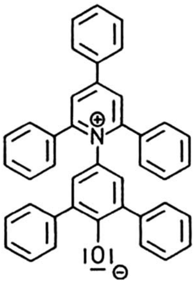

Fig. 1 Molecular structure of the negatively solvatochromic standard pyridinium *N*-phenolate betaine dye, no. 30.[3]

[a] Chemistry Department, Shiraz University, Shiraz, Iran. E-mail: hemmatb@sums.ac.ir
[b] Chemistry Department, Persian Gulf University, Bushehr, Iran

dye. According to eqn (1), high $E_T(30)$ values correspond to high solvent polarity.[5]

Multivariate image analysis (MIA) is a class of chemometric methods that deals with extracting meaningful information from chemical images, mainly spectroscopic images such as infrared, fluorescence, and Raman images.[6,7] MIA was born to deal with images that presented more than one measurement per pixel (related to the three RGB channels in color images or to a number of spectroscopic channels in multispectral and hyperspectral images).[8] MIA has been widely used to characterize different specimens such as bread,[9] fish,[10] skin,[11] cancerous tissues[12] and so on. Detection systems based on image analysis of a CCD camera have also found an application in analytical chemistry.[13]

Recently, our active research has been focused on the development of simple, inexpensive and fast methods and techniques to determine physical and chemical constants. In this regard, we use image analysis for its ability to perform fast and non-invasive low-cost analysis of products and processes. So, we used multivariate analysis of the images recorded from thin-layer chromatography sheets to develop an analytical method for simultaneous determination of components of highly overlapped spots[14] and for characterization of organic reactions.[15]

As explained previously, $E_T(30)$ is one of the mostly applied empirical parameter of solvent polarity, and according to eqn (1) it can be calculated from λ_{max}, which is determined experimentally by a spectrophotometer. Since color values in the RGB space are correlated with the absorbance spectra of the solutions,[16] we suggest using color values of the images recorded by a digital camera (instead of λ_{max}) as analytical parameters to calculate $E_T(30)$. However, since the relationship between $E_T(30)$ and RGB color values is not a simple linear relation, we use chemometric methods to make a connection between color values and $E_T(30)$. We use the principal component analysis (PCA)[17] method to model the color values in the RGB space and then demonstrate how artificial neural networks[18] and multiple linear regression can be used to model PCA scores to produce quantitative estimates of the $E_T(30)$ values. To the best of our knowledge, quantization of $E_T(30)$ by image analysis has not been addressed until now.

Material and methods

Reagents

Solvents were used as commercially supplied in the highest available quality (HPLC or extra pure grade). The list of the solvents used in this study is given in Table 1. Precautions were taken to avoid evaporation and contamination by humidity. A sample of the Dimroth-Reichardt dye [2,6-diphenyl-4-(2,4,6-triphenylpyridinium-1-yl)phenolate] was obtained from Professor Ch. Reichardt, Marburg, Germany, as a gift. Binary solvent mixtures were prepared gravimetrically. The concentration of the dye in all solvents was 3.0×10^{-4} M. The solutions in each individual solvent and binary solvent mixture were prepared just prior to use.

Table 1 The list of solvents and their corresponding $E_T(30)$ values used in this study[a]

Solvent	$E_T(30)$/(kcal mol^{-1})
Ethylene glycol	56.3
Methanol	55.4
Methanol/acetone (0.9/0.1)	54.8
Methanol/acetone (0.7/0.3)	54.1
Methanol/acetone (0.5/0.5)	53.0
Methanol/acetone (0.4/0.6)	52.7
Ethanol	51.9
Methanol/acetone (0.3/0.7)	51.7
1-Propanol	50.7
Methanol/acetone (0.2/0.8)	50.6
1-Butanol	49.7
Methanol/acetone (0.1/0.9)	47.8
Acetonitrile	45.6
DMSO	45.1
DMF	43.2
Acetone	42.2
Dichloromethane	40.7
Toluene/acetone (0.3/0.7)	40.7
Toluene/acetone (0.5/0.5)	39.6
Chloroform	39.1
Toluene/acetone (0.7/0.3)	38.5
Ethyl acetate	38.1
Toluene/acetone (0.9/0.1)	37.6
THF	37.4
o-Xylene	34.7

[a] For mixed solvents the values in parentheses denote mole fractions.

Procedure

A 3.0 mL portion of the betaine dye solution in each solvent was transferred into a glass cell and pictures of them were recorded by a camera for each sample (Fig. 2). In order to obtain high quality and reproducible images, radiation source equipment composed of a two-dimensional array of LED lamps covered with a flexi glass sheet was constructed. The cells were placed in front of the radiation source. A SONY-H5 digital camera with the capability of taking images with 7.2 Mega pixel resolution was used and fixed in front of the cells. The employed camera has different modes of an imaging system. Any mode has a special

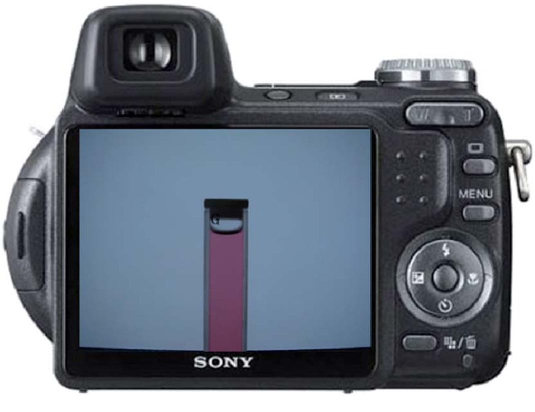

Fig. 2 An example of the captured image of the betaine dye dissolved in one solvent by a digital camera.

characteristic; in this work, the auto-adjustment mode and macrophotography, which is close-up photography, were used.

Data analysis

All MIA calculations were performed in the MATLAB environment (Mathwork Inc.). The images, saved in JPEG format, were imported into MATLAB and then they were digitized using the "imread" function. The output of this function is three matrices of color values for red (R), green (G), and blue (B) parts of the images. Thus, for each picture three data matrices composed of color values of R, G, and B were obtained. However, to avoid edge effect of the pictures, only 25 pixels around the center of the images (from top, down, left and right) were selected. Thus, the size of each R, G or B matrix was (51 × 51) pixels. Then, each ($I \times J$) matrix was unfold to row vectors of length of IJ, where I and J are the number of pixels in the direction of lengths and widths of the images, respectively. Then, all R, G and B vectors were stacked beside each other to form a larger vector of length of IJK, where K is the number of color elements (here $K = 3$). For a series of M solvents, the data can be organized as a data matrix D of the size of ($M \times IJK$), where M is the number of used solvents (here $M = 25$). It should be noted that unlike conventional multivariate image analysis where pixels are taken as objects, here we used solvents as objects and pixels were taken as variables. So, color information in the pixels was used to model the polarity scale of solvents.

In the image analysis field, principal component analysis (PCA)[19] has become the most common method[20] to reduce the dimension of the datasets by keeping the relevant information. The extracted principal components (PCs) are new uncorrelated and approximately normally distributed variables that provide faithful representations of the image, which can be used later on as input information for exploration, segmentation, classification and other purposes.[8] In this work, since the number of variables is much larger than the number of objects and also since the color values are correlated, PCA was run on matrix D to produce uncorrelated PCs of much lower dimension than the original variables. Then, the scores of the PCA were used as input variables of MLR or ANN models.

Multiple linear regression (MLR) and artificial neural networks (ANN)[18] were employed as linear and nonlinear regression methods, respectively. Moreover, the generated models were validated by leave-one-out cross-validation and y-scrambling methods.[21] Final model performances were evaluated using ten randomly selected solvents as an external prediction set.

Results and discussion

In this study, we used an image-capturing device as an alternative to spectrophotometers for the calculation of $E_T(30)$ values of solvents, as one of the most important scales of solvent polarity. We correlated the color values of the images taken from betaine dye solutions in different solvents and the previously determined $E_T(30)$ values. The literature values of $E_T(30)$, determined by the spectrophotometric method, were used for individual and binary solvent mixture systems.[22] Table 1 gives the $E_T(30)$ values of the studied solvent systems.

The recorded images and absorbance spectra of $E_T(30)$ solutions in some representative solvents are given in Fig. 3a. At first, the relationship between color values (obtained with the digital camera) and λ_{max} (obtained by UV/Vis spectrophotometric measurements) was investigated. As it is observed from Fig. 3b, none of the RGB values represent a simple and linear relationship with λ_{max}. However, the R values exhibit relatively linear relationships with λ_{max} in two different wavelength intervals. Similar trends were observed between $E_T(30)$ and RGB color values.

As it was explained previously, the recorded image of each solution can be arranged into a two-dimensional array of RGB color values according to the procedure described in the experimental section. In the next step, we used the principal component analysis method to convert the pictured images of the $E_T(30)$ solutions to the scores suitable for linear and nonlinear calibrations. So, the three factors or PCs were selected by cross-validation and used as input variables of MLR and ANN

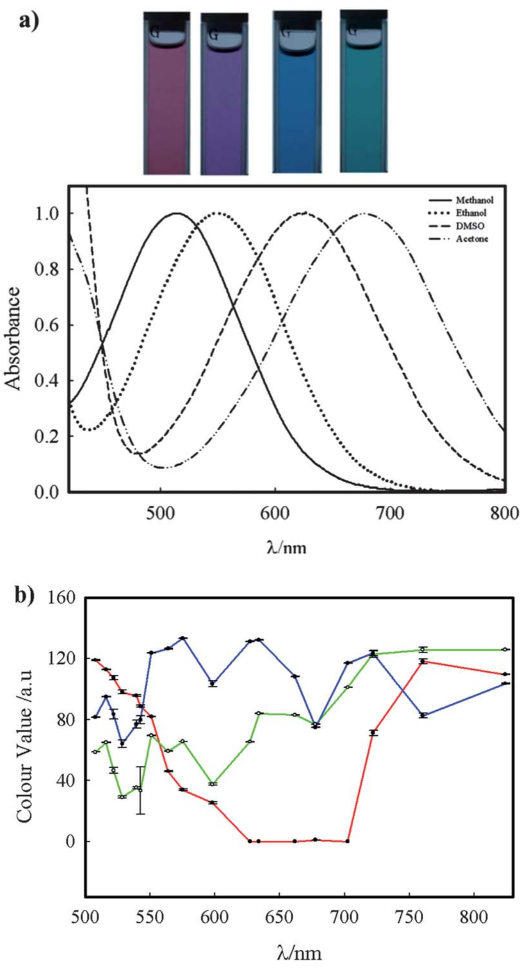

Fig. 3 (a) Image and vis absorption spectra of the standard betaine dye for four representative solvents. (b) Relationship between the color values (red line: R, green line: G, blue line: B) of the images and lambda (max) values for the eighteen solvents and solvent mixtures studied.

models. It was found that the data matrix D can be represented by 3 PCs, capturing 99% of variances.

Now, a solvent-related PCA score matrix of dimension (25×3) is available as the representation of the images recorded from the $E_T(30)$ solutions in different solvents. Each column of this matrix is considered as one input variable of MLR and ANN models. In the next, these scores are represented by S_1, S_2 and S_3, respectively. Firstly, simple multivariate calibration methods, MLR, were used to predict $E_T(30)$ values of solvents employing S_1, S_2 and S_3 as independent or predictor variables. Also to consider nonlinearity in the relationship between $E_T(30)$ and the image, second terms and interactions of PCs were also included. To do so, a multiparametric equation of the form of eqn (2) was considered:

$$E_T(30)\ (\text{kcal mol}^{-1}) = b_0 + b_1S_1 + b_2S_2 + b_3S_3 + b_{11}{S_1}^2 + b_{22}{S_2}^2 + b_{33}{S_3}^2 + b_{12}S_1S_2 + b_{13}S_1S_3 + b_{23}S_2S_3 + b_{123}S_1S_2S_3 + e \quad (2)$$

where the constants b_i are the regression coefficients and e is the noise (non-modeled part of $E_T(30)$).

The dataset was partitioned into a training set (15 samples) and a prediction set (10 samples). By regression of $E_T(30)$ over the PCA scores utilizing stepwise selection of variables the following equation was obtained:

$$E_T(30)/(\text{kcal mol}^{-1}) = 44.70(0.74) - 0.01(0.007)S_2 + 5.60 \times 10^{-6}(2.25 \times 10^{-6}){S_2}^2 + 1.10 \times 10^{-5}(9.12 \times 10^{-7})S_1S_3 \quad (3)$$

The values in parentheses are the standard deviation of the regression coefficients. It is observed that the second PC represents a quadratic effect on $E_T(30)$, whereas only the interaction term of the first and third PCs is appeared in eqn (3). The more significant contribution of S_2 over S_1 is not so strange. It is correct that the first PC possesses the largest variances of original data; however, this does not mean that this PC is essentially the best for calibration. This has been addressed in our previous publication.[23]

The statistical parameters of the obtained MLR models are summarized in Table 2. Obviously, the quality of the generated model is not so high but the RMS errors of both calibration and prediction are acceptable (the relative RMS errors of training and prediction are 2.81% and 4.94%, respectively) compared to the average value of the $E_T(30)$ of the training set samples. The plot of the calculated $E_T(30)$ by MIA *versus* spectrophotometrically measured $E_T(30)$ is shown in Fig. 4. It is observed that the data are scattered around a straight line with a slope of 0.99 and an intercept of 0.32, which are close to the ideal values of one and zero, respectively.

Table 2 Correlation coefficients and root mean square errors of calibration, prediction and cross-validation (RMSE$_C$, RMSE$_P$, and RMSE$_{CV}$, respectively) for the determination of $E_T(30)$ values by MLR

Calibration		Prediction		
RMSE$_C$	R^2	RMSE$_P$	R^2	RMSE$_{CV}$
1.31	0.96	2.30	0.90	2.02

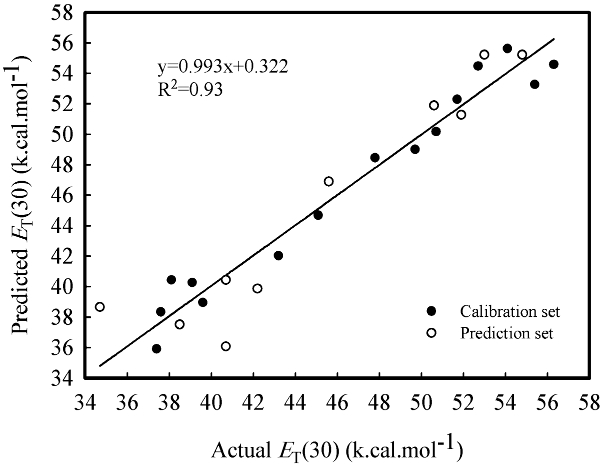

Fig. 4 Relationship between the actual values and predicted values (obtained by MLR) of $E_T(30)$ in a calibration set (filled circles) and a prediction set (empty circles).

In the next step, an artificial neural network (ANN) was employed to predict the $E_T(30)$ values based on the training data. The artificial neural network used in this work was a simple feed-forward neural network trained with a back-propagation of error, usually referred to simply as the back-propagation neural network. The general topology of the networks used in this work is shown in Fig. 5. The network possesses three inputs corresponding to the PCA scores S_1, S_2 and S_3. It should be noted that since ANNs consider a nonlinear relationship in its structure, we only considered the linear terms as the input of ANNs. The hidden layer nodes employed sigmoidal transfer functions which facilitated non-linear modeling. The number of nodes in the hidden layer was optimized to gain a lower RMSE of cross-validation. The output layer of the network combines the outputs of the hidden layer and another fixed input which provides an offset. Although a sigmoidal transfer function could be used in the output node, a simple linear function was found to work well in this application.

The specifications for the networks created for the training of the data are listed in Table 3. The parameters listed in Table 3 were found to be optimal for fast learning with low prediction errors of cross-validation.

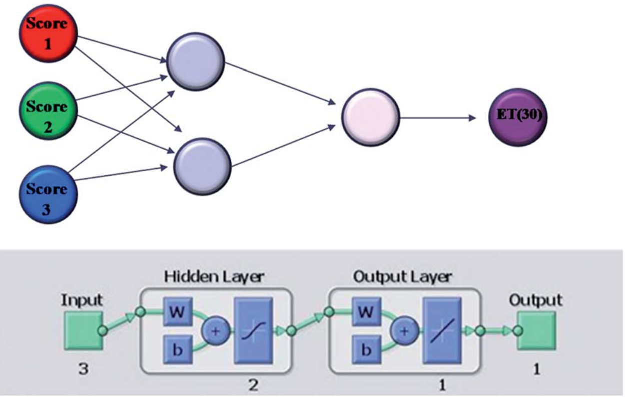

Fig. 5 Artificial neural network topology used in this work.

Table 3 Artificial neural network specifications and parameters

Parameter	Data
Hidden nodes	2
Learning rate	0:00:19
Momentum	0.1
Gradient	0.00012
Input-layer transfer function	Linear
Hidden-layer transfer function	Sigmoid
Output-layer transfer function	Linear

The corresponding prediction results for calibration and prediction sets are summarized in Table 4. The high correlation coefficients (0.99 for the prediction and 0.89 for the test set) and very low average relative errors (1.39% for the calibration set and 4.74% for the test set) reveal the capability of the obtained model for the prediction of $E_T(30)$ values. It should be noted that leave-one-out cross validation was run at five different weight initialization trials. Also, the prediction set samples did not have contribution in the network training and the best network was selected using cross-validation.

In order to further evaluate the ANN model obtained in this work, the chance effect was also checked. We have examined the performance of networks designed to carry out ANN modeling by using random numbers as input data and a single, random, continuous target. Detailed results from these experiments are given in the last columns of Table 4. The root mean square errors are worse for the calibration and prediction sets confirming that the obtained model is not chancy. To examine the predictive ability of the ANNs, it is useful to look at plots of the predicted property value *versus* the measured value. The corresponding plots for the training and test set are shown in Fig. 6. The data are distributed around a straight line with slope and intercept very close to the ideal values. An interesting point is the homogenous scattering of training and test samples. This suggests that the obtained model is not only optimum for self-prediction but it can also accurately predict the $E_T(30)$ values of the external samples.

Table 4 Correlation coefficients and root mean square errors of calibration, prediction and cross-validation ($RMSE_C$, $RMSE_P$ and $RMSE_{CV}$, respectively) for determination of the $E_T(30)$ values by ANN

Calibration		Prediction		
$RMSE_C$	R^2	$RMSE_P$	R^2	$RMSE_{CV}$
Artificial neural network				
0.65	0.99	2.21	0.89	2.25
ANN chance effect				
4.8	0.50	10.10	0.36	10.12

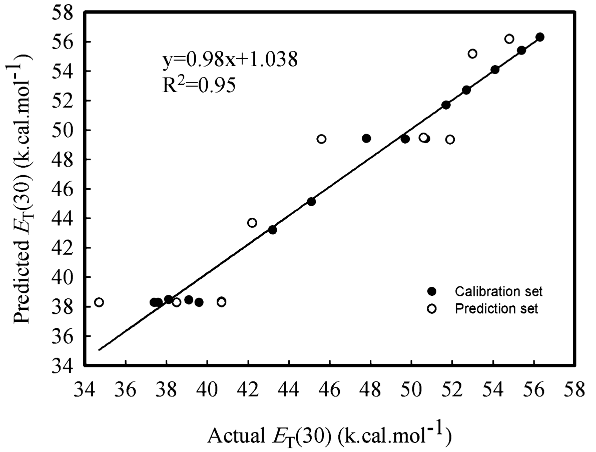

Fig. 6 Relationship between the actual values and predicted values (obtained by ANN) of E_T^N in the calibration set (filled circles) and prediction set (empty circles).

Conclusion

The proposed method was successfully applied for the estimation of $E_T(30)$ values using the color of solution instead of the wavelength of maximum absorbance. RGB color values were first processed by the PCA method and the scores of PCA were modeled using artificial neural network (ANN) and multiple linear regression (MLR) methods. By including nonlinear terms in the MLR model, it represented the results similar to ANN. However, MLR is preferred over ANN for its simplicity and higher interpretability of the generated models. In comparison with the conventional spectrophotometric method, the proposed method is more simple and cheap from the instrumental point of view. However, it is mathematically more complex. In addition, in the spectrophotometric method, only λ_{max} is measured and tailing or broadening of the spectrum is not considered. However, since peak broadening and tailing affect the observed color of the solution, in the imaging method all of the peak characteristics are included.

Acknowledgements

We gratefully acknowledge a generous gift of the betaine dye and worthwhile advices from Prof. Ch. Reichardt, Marburg, Germany. We are also thankful to the reviewers for their valuable comments.

References

1 C. Reichardt, *Angew. Chem.*, 1979, **91**, 119; C. Reichardt, *Angew. Chem., Int. Ed. Engl.*, 1979, **18**, 98; C. Reichardt, *Chem. Rev.*, 1994, **94**, 2319; C. Reichardt, S. Asharin-Fard, A. Blum, M. Eschner, A.-M. Mehranpour, P. Milart, T. Niem, G. Schäfer and M. Wilk, *Pure Appl. Chem.*, 1993, **65**, 2593; C. Reichardt, *Green Chem.*, 2005, **7**, 339.

2 C. Reichardt, *Pure Appl. Chem.*, 2004, **76**, 1903; C. Reichardt, *Pure Appl. Chem.*, 2008, **80**, 1415.

3 K. Dimroth, C. Reichardt, T. Siepmann and F. Bohlmann, *Justus Liebigs Ann. Chem.*, 1963, **661**, 1.

4 C. Reichardt, *Chem. Soc. Rev.*, 1992, **21**, 147.

5 C. Reichardt and E. Harbusch-Görnert, *Liebigs Ann. Chem.*, 1983, 721.

6 R. Wolthuis, A. Travo, C. Nicolet, A. Neuville, M. P. Gaub, D. Guenot, E. Ly, M. Manfait, P. Jeannesson and O. Piott, *Anal. Chem.*, 2008, **80**, 8461.

7 J. M. Emory and S. A. Soper, *Anal. Chem.*, 2008, **80**, 3897.

8 J. M. Prats-Montalbán, A. de Juan and A. Ferrer, *Chemom. Intell. Lab. Syst.*, 2011, **107**, 1.

9 K. Kvaal, J. P. Wold, U. G. Indahl, P. Baardseth and T. Naes, *Chemom. Intell. Lab. Syst.*, 1998, **42**, 141.

10 T. Brosnan and D. W. J. Sun, *Chilton's Food Eng.*, 2004, **61**, 3.

11 J. M. Prats-Montalban, A. Ferrer, R. Bro and T. Hancewicz, *Chemom. Intell. Lab. Syst.*, 2009, **96**, 6.

12 R. S. Kakonen, J. P. Rissanen, M. I. Suominen and J. N. Halleen, *J. Bone Miner. Res.*, 2008, **23**, S187; L. Mulrane, E. Rexhepaj, S. Penney, J. J. Callanan and W. M. Gallagher, *Expert Rev. Mol. Diagn.*, 2008, **8**, 707.

13 T. Hayakawa and M. Hirai, *Anal. Chem.*, 2003, **75**, 6728; N. Maleki, A. Safavi and F. Sedaghatpour, *Talanta*, 2004, **64**, 830; A. Safavi, N. Maleki, A. Rostamzadeh and S. Maesum, *Talanta*, 2007, **71**, 498.

14 B. Hemmateenejad, N. Mobaraki, F. Shakerizadeh-Shirazi and R. Miri, *Analyst*, 2010, **135**, 1747.

15 B. Hemmateenejad, M. Akhond, Z. Mohammadpour and N. Mobaraki, *Anal. Methods*, 2012, **4**, 933.

16 B. Haifa, V. Bacarea, O. Iacob, T. Calinici and A. Schiopu, *Appl. Med. Inf.*, 2011, **28**, 29.

17 I. T. Jolliffe, *Principal Component Analysis*, Springer, 2002.

18 J. R. Long, V. G. Gregoriou and P. J. Gemperline, *Anal. Chem.*, 1990, **62**, 1791.

19 J. E. Jackson, *A User's Guide to Principal Components*, Wiley, New York, 1991.

20 A. K. Jain, R. P. W. Duin and J. Mao, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2000, **22**, 4.

21 P. Gramatica, *QSAR Comb. Sci.*, 2007, **26**, 694.

22 C. Reichardt and T. Welton, *Solvents and Solvent Effects in Organic Chemistry*, Wiley-VCH, Weinheim/Germany, 4th edn, 2010, ch. 7.4, p. 448ff; P. M. E. Mancini, A. Terenzari, M. G. Gasparri and L. R. Vottero, *J. Phys. Org. Chem.*, 1995, **8**, 617.

23 B. Hemmateenejad, *J. Chemom.*, 2004, **18**, 475.