Discrimination analysis of gallstones by near infrared spectrometry using a soft independent modelling of class analogy

Sang Hak Lee,[a] Bum Mok Son,[a] Ju Eun Park,[a] Sang Seob Choi[b] and Jae Jak Nam[c]

[a]*Department of Chemistry, Kyungpook National University, Taegu 702-701,Korea*

[b]*Department of Environmental Management, Andong Science College, Andong 760-820, Korea*

[c]*National Institute of Agricultural Science and Technology, Suwon 441-707, Korea*

Introduction

Classification of gallstones has been based mostly on morphology by visual inspection.[1] Classification by chemical analysis has improved the results but it has been limited by the dissolution of the sample and the insolubility of many of the components. Discrimination of gallstones presented a problem primarily because of the presence of the bile pigment and inorganic compounds. Various instrumental methods have been proposed but are not complete in themselves alone. However, the application of solid state analytical methods such as infrared spectroscopy and X-ray diffraction has permitted the identification of the type of gallstones and to determine the contents of insoluble components.[2,3]

Recently, the application of near infrared (NIR) spectroscopy to quantitative and qualitative analysis of organic compounds is finding increasing use since the NIR technique allows rapid analysis of powdered samples with little sample preparation.[4] The analysis and interpretation of NIR spectra require a variety of chemometric tools. The identification of the NIR spectra needs a suitable chemometric classification method to correctly identify unknown samples. Several methods for this purpose have been reported.[5–8] A soft independent modelling of class analogy (SIMCA) is a classification technique which gives a distinct confidence region around each class after applying principal components analysis (PCA).[9–12] New measurements are projected into each principal component's (PCs) space that describes a certain class to evaluate whether they belong to it or not.

In the present work a method to discriminate human gallstones by NIR spectrometry using SIMCA has been studied. The NIR spectra of 150 gallstones in the wave number range from 4500 to 10,000 cm^{-1} were measured. The 150 gallstone samples were classified to three classes (cholesterol stone, calcium bilirubinate stone and calcium carbonate stone) according to the contents of major components in each gallstone. The training set which contains objects of the different known classes was constructed using 120 NIR spectra and the test set was made with 30 different gallstone spectra. The number of important PCs to describe each class was determined by cross-validation in order to improve the decision criterion of the SIMCA for the training set. For each class the score plot of the objects in the

Near Infrared Spectroscopy: Proceedings of the 10th International Conference

Table 1. The construction of class data set.

Class	Type of gallstone	Number of samples (Training set)	Number of samples (Test set)
A	calcium bilirubinate stone	40	10
B	calcium carbonate stone	40	10
C	cholesterol stone	40	10

training set belonging to the other classes was inspected. The critical distance for each class was computed using both Euclidean distance and Mahalanobis distance at an appropriate level of significance (α). Two methods were compared with respect to classification and their robustness towards the number of PCs selected to describe different classes.

Experimental

Sample and instrumentation

The gallstone samples studied in this work were kindly provided by the Kaemyung University Hospital (Taegu, Korea). The NIR spectra were collected in the reflectance mode with an InfraProverII FTIR spectrometer (Bran+Luebbe). Before the data acquisition, a successful system suitability test (wavenumber scale, absorbance scale and noise) was performed. Each spectrum used for SIMCA is the average spectrum of 32 scans. The spectral range used for the data analysis is from 4500 to 10,000 cm^{-1}. The samples were classified into three classes; cholesterol stone, calcium bilirubinate stone and calcium carbonate stone. The composition of experimental design employed in this study is listed in Table 1.

Procedure of SIMCA

A training matrix contains objects of different known classes and the submatrix contains *n* training objects belonging to a class that were measured at *p* variable. Each class is modelled separately, based on the similarity of objects within the class. The singular value decomposition (SVD) can be used to

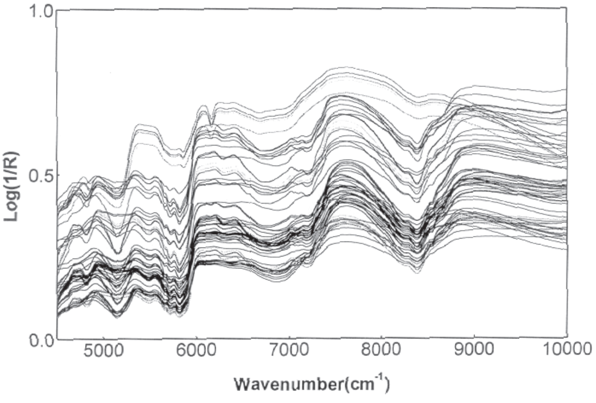

Figure 1. NIR spectra of gallstones taken from Korean patients.

perform PCA after column centering. In this work the number of PCs was selected according to the percentile of the total variance that is expressed by each PC. The PCs containing more than 1% of total variance were arbitrarily chosen for modelling.[12] The class boundaries or confidence limits are then constructed around the PC model. They are based on the distribution of distances (Euclidean distance, ED or Mahalanobis distance, MD) between the objects and the origin in the space of the residual PCs. With the help of an F-Test the critical distance can further be computed at a certain level of significance (a). To predict whether a certain object belongs to a certain class, it is projected on the space defined by the selected PCs of the training set of that class. After the model has been developed on the training set, new objects can be classified. For the identification of such a new object it is projected into the PC space defined by the PCA model and its distance from the class model is compared to the critical distance. If the distance is large than the critical distance, the object is considered a part of the class for which the model was established.

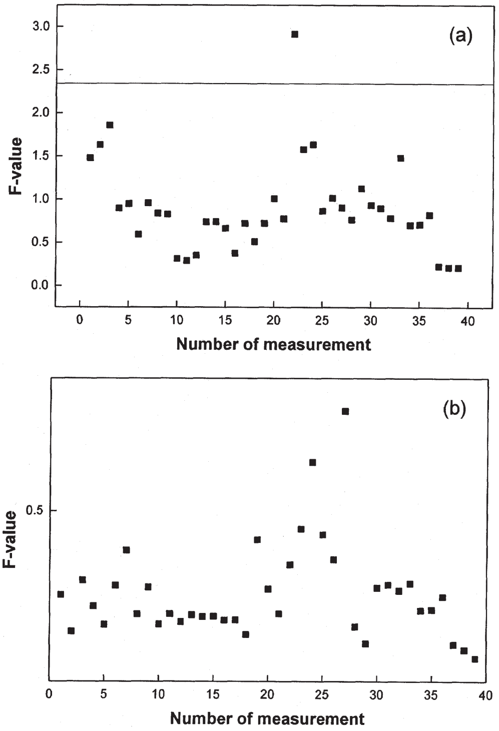

Figure 2. Calculated F-values based on Euclidean distance (a) and Mahalanobis distance (b) for the class A. Solid line indicates the critical F-value.

Result and discussion

Figure 1 shows the mean spectra of all classes. To determine the number of significant PCs for the training set, the SCREE and LEV plot method and the Malinowski IND function were used. The SCREE and LEV plot indicated eight significant PCs. Malinowski's IND function selected 11 PCs. The selection of all PCs that explain more than 1% of total variance within the data gave five significant PCs. The loadings on the early number PCs attribute weight to the different ranges of spectra. This indicates that the first eight PCs contain information that needs to be included in the model. From PC Number 9, the loadings are all

Table 2. Classification results of SIMCA based on Euclidean distance (ED) and Mahalanobis distance (MD).

Class	ED				MD			
	Training set		Test set		Training set		Test set	
	Sample	Outlier	Sample	Outlier	Sample	Outlier	Sample	Outlier
A	40	1	10	0	40	0	10	1
B	40	2	10	1	40	1	10	2
C	40	0	10	1	40	1	10	2

near to zero. From the result of PC selection methods and the investigation of the loadings, the selection of five to eight significant PCs seems to be reasonable. In this work we have decided to retain the PCs containing more than 1% of the total variance for our modelling. The remaining, residual PCs were used to build the confidence interval around the model.

Figure 2 shows the SIMCA results of class A for F-values obtained by using ED(a) and MD(b) for the training set at a significance level of 0.05. The SIMCA using MD seems to lead to fewer outliers. Table 2 shows the results of discrimination analysis of the test set. The SIMCA results using ED seem to have fewer outliers. The results of this work indicate that human gallstones can be successfully classified by NIR spectrometry using SIMCA.

Acknowledgement

This research was supported by Kyungpook National University Research Fund, 2000.

References

1. J.Y. Chu, I.H. Kim, T.J. Lim, H.J. Ryu, S.B. Kim and S.H. Lee, *J. of Korean Surgical Society* **51,** 88 (1996).
2. P.F. Malet, M.A. Dabezies, G. Huang, W.B. Long, T.R. Gadacz and R.D. Soloway, *Gastroenterology* **94,** 1217 (1988).
3. R.D. Soloway, B.W. Trotman and J.D. Ostrow, *Gastroenterology* **72,** 167 (1977).
4. B.M. Smith and P.J. Gemperline, *Anal. Chim. Acta* **423,** 167 (2000).
5. D. González-Arjona and A.G. González, *Anal. Chim. Acta* **363,** 89 (1998).
6. D. Jouan-Rimbaud, B. Walczak, D.L. Massart, I.R. Last and K.A. Prebble, *Anal. Chim. Acta* **304**, 285 (1995).
7. W.J. Welsh, W. Lin, S.H. Tersigni, E. Collantes, R. Duta, M.S. Carey, W.L. Zielinsk, J. Brower, J.A. Spencer and T.P. Layloff, *Anal. Chem.* **68,** 3473 (1996).
8. R. De Maesschalck, A. Candolfi, D.L. Massart and S. Heuerding, *Chemom. Intell. Lab. Syst.* **47,** 65 (1997).
9. A. Candolfi, R. De Maesschalck, D.L. Massart, P.A. Hailey and A.C.E. Harrington, *Chemom. Intell. Lab. Syst.* **19,** 923 (1999).
10. S. Wold and M. Sjöström, in *Chemometrics: Theory and Application*, Ed by American Chemical Society, Washington, DC, USA, p. 243 (1977).
11. M.P. Derde and D.L. Massart, *Chemom. Intell. Lab. Syst.* **4,** 65 (1988).
12. B. Mertens, M. Thompson and T. Fearn, *Analyst* **119,** 65 (1988).