Application of High-Resolution Computer Graphics to Pattern Recognition Analysis

B. K. Lavine* and A. B. Stine

Department of Chemistry, Clarkson University, Potsdam, New York 13699-5810

Howard Mayfield

Head Quarters AFCESA/RDVC, Tyndall Air Force Base, Tyndall, Florida 32403-6001

Robert Gunderson

Department of Electrical Engineering, Utah State University, Logan, Utah 84322

Received June 21, 1993®

False color data imaging has proven to be a powerful technique for the analysis and visualization of satellite data. This success suggests that high-resolution computer graphics can also play an important role in the analysis of multivariate data obtained from chemical instruments. The development of a graphics tool for visualization of multivariate data is described in this paper. The proposed graphics tool utilizes the fuzzy c-varieties clustering algorithm, principal components analysis and false color data imaging to generate maps of high informational density of the data space. The graphics tool has been tested successfully using data sets from the literature representative of the pattern recognition problems encountered by chemists.

INTRODUCTION

Low-cost, high-performance computers have made it possible for chemists to combine different analytical methods into so-called hyphenated techniques. Gas chromatography combined with mass spectrometry (GC–MS) and liquid chromatography combined with Fourier transform infrared spectroscopy (LC–FTIR) are just two examples of sophisticated measurement systems that combine the separation capability of chromatography with the capability for compound identification of spectroscopy. GC–MS and LC–FTIR have greatly increased the number of compounds that can be identified and quantified, even at trace levels, in the environment. However, these techniques also produce large quantities of data. Often, multivariate statistical methods are required for analysis of the data. Therefore, analytical chemists have turned to pattern recognition methods[1-3] to analyze the larger data sets generated in studies involving hyphenated techniques.

Pattern recognition has its origins in the field of image and signal processing. In a typical pattern recognition study, samples are classified according to a specific property by using measurements indirectly related to that property. For pattern recognition analysis, each sample is represented by a data vector $\mathbf{x} = (x_1, x_2, x_3, ..., x_j, ..., x_n)$, where component x_j is the value of the jth measurement. Such a vector can also be considered as a point in an n-dimensional measurement space—the distance between pairs of points in this n-space is inversely related to the degree of similarity between the samples. Therefore, points representing samples from one class will cluster in a limited region of the measurement space distant from the points corresponding to the other class. Pattern recognition is a set of methods for investigating data represented in this manner in order to assess its structure, which is defined as the overall relation of each object to every other object in the data set.

To characterize the multidimensional structure of a data set, i.e. to visualize the relative position of the sample points in the corresponding n-dimensional measurement space, a two- or three-dimensional representation of the measurement space is needed that faithfully reflects its high-dimensional structure. One approach to this problem is to use a technique called principal components analysis (PCA).[4] In PCA, the original measurement variables are transformed into new, uncorrelated variables called principal components. Each principal component is a linear combination of the original measurement variables. The most informative principal component is the first, and the least informative is the last. Using this procedure is analogous to finding a set of orthogonal axes that represent the directions of greatest variance in the data. For a data set with a large number of interrelated variables, PCA is a very powerful method for analyzing the structure and reducing the dimensionality of a data set.

With principal component analysis, the original data can be mapped onto a subspace defined by the two or three largest principal components. However, there is a problem associated with principal component maps. The spatial relationships between individual data points and between groups of points in the original measurement space are often distorted when the original data are mapped onto a subspace defined by the two or three largest principal components. Using the fuzzy c-varieties (FCV) pattern recognition clustering algorithm,[5] much of this missing spatial information can be restored through the use of color, a crucially important information dimension. In this article, an exploratory data analysis technique is described which utilizes the techniques of false color data imaging,[6] FCV clustering,[7] and principal components analysis to generate informative two- or three-dimensional colored maps of the high-dimensional measurement space. Two representative studies are presented that demonstrate the usefulness of this graphics tool.

® Abstract published in *Advance ACS Abstracts*, October 15, 1993.

THEORY

False Color Data Imaging. The technique of false color data imaging has long been used to analyze multispectral data collected by satellite electromagnetic scanner systems. In a false color data imaging experiment, each data vector is projected onto a three-dimensional coordinate system whose axes are defined by the three largest principal components of the original data, with each axis assigned a primary color, e.g. red, blue, and green. A given data point is then assigned a color which is a combination of the three primary colors, with the intensity of each color being inversely scaled according to the distance of the data point from the particular color axis in question. Therefore, a data point close to both the red and blue axes would be assigned the color purple, whereas a point projected onto a line equidistant from all three coordinate axes would be assigned a color composed of equal intensities of the three primary colors, i.e. some shade of gray.

FCV Clustering Algorithm. The FCV clustering algorithm attempts to fit each of the classes in the data set to a principal component model. A unique feature of the FCV clustering algorithm is that each data vector in the training set influences or contributes to the modeling of each of the classes within the data. The actual algorithm consists of four equations that are solved simultaneously using a Picard iteration procedure.[8]

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{c} (D_{ik}/D_{jk})^{1/(m-1)}} \tag{1}$$

$$v_i = \frac{\sum_{k=1}^{n} (\mu_{ik})^m x_k}{\sum_{k=1}^{n} (\mu_{ik})^m} \tag{2}$$

$$S_i = \sum_{k=1}^{n} (\mu_{ik})^m (x_k - v_i)(x_k - v_i)^T \tag{3}$$

$$D_{ik} = (|x_k - v_i|^2 - \sum_{j=1}^{r} \langle x_k - v_{ij}, d_{ij} \rangle^2)^{1/2} \tag{4}$$

The membership value of sample k with respect to class i ($i = 1, 2, 3, ...$) is μ_{ik}, and these values are subject to the conditions $0 < \mu_{ik} < 1$ and $\Sigma\mu_{ik} = 1$. D_{ik} is the distance between cluster center i and sample k, v_i is the center of cluster i, d_{ij} is a unit eigenvector corresponding to the jth largest eigenvalue of the fuzzy within cluster scatter matrix S_i, m is a fixed weighting exponent (usually selected to be 2), and r defines the shape of the cluster ($r = 0$ for round clusters, $r = 1$ for linear varieties, etc.)

To obtain a solution to this set of four equations, the user must supply the starting cluster centers: class membership values, the within cluster scatter matrix, and the distance between each sample and each cluster in the data are computed in rapid succession. New cluster centers are recomputed, and the algorithm uses these new cluster centers as the starting point for a second iteration. This process continues until the algorithm converges: the number of iterations necessary for convergence depends on the prespecified change criterion for the class membership value.

FCV–False Color Data Imaging. The first step in a FCV–false color data imaging[9–10] experiment is to display the multivariate data by projecting it onto a suitable two- or three-dimensional subspace. The two or three largest principal components of the original data are a good choice for this subspace because one has reasonable assurance of capturing most of the variance or "action" in the measurements. Since this can be easily accomplished with microcomputer graphics even for large data sets, it is a reasonable first step in any analysis. However, there is a problem with PCA. The spatial relationships between the data points in n-space are often distorted in the projective process. It is at this point where the proposed method departs substantially from other graphical display techniques. By taking advantage of the membership coefficients generated by the FCV clustering algorithm, much of the spatial information lost during projection can be restored through the use of color, a crucially important information dimension.

With the FCV clustering algorithm, the data can be fitted to a user-specified number of linear disjoint principal component models. Each model, which represents a different cluster of data points, is assigned a different contrasting basic color: red, green, blue, etc. Data points can then be displayed as mixtures of these colors, with the amount of any one color in the mixture determined by the sample's membership value for that particular class. Interpretation of the resulting color images can provide valuable insight into the data structure. For example, two projected data points do not lie close to one another in the high-dimensional space (and hence are not similar) unless they are displayed in nearly the same color, or a solid group of data appearing in a nonprimary color suggests the presence of an unsuspected class, and so forth.

SOFTWARE IMPLEMENTATION

It was proposed that a software system be created that would allow the user to perform the tasks of editing, analysis, and visualization of data. In designing the system, a number of specific goals were established from the outset. First, the system should be easily portable to a wide variety of hardware/software platforms including workstations, personal computers, and graphics terminals. Second, the system should employ an existing and widely used graphics Application Programming Interface (API) such as Graphical Kernel System (GKS), Programmers Hierarchical Interactive Graphics System (PHIGS), PostScript, or X Window System (X). Third, the system should be implemented using languages that are available over numerous platforms, e.g. C and FORTRAN. Fourth, the system should be designed and written in a modular fashion. Fifth, the system should be able to utilize distributed computing resources. And finally sixth, the system should integrate off the shelf components whenever possible and practical, e.g. spreadsheets.

To address the aforementioned goals of the proposed FCV–false color data visualization system, one must consider the graphics technology available at the time of the start of the project which was 1987. Of the available graphics API's, only GKS seemed to provide sufficient portability across platforms of interest. Other graphics API's that were investigated included PHIGS (insufficient industry support), X (insufficient industry support at the time), proprietary windowing systems such as SunView, VWS, or MacIntosh (not portable).

Initial prototyping was performed using a VAXstation 2000 with four-plane color. Some basic graphics modules were completed as a proof of concept, in order to validate the design and the overall system concept. During the course of this work, we learned that GKS lacked the graphical user interface (GUI) functionality necessary for the system. Evidently, GKS was not capable of providing an elegant user interface.

Furthermore, GKS implementations could not be found to cover a wide range of platforms (workstations, personal computers, etc). Typically, the best implementations were available for the workstations and the least suitable were those on personal computers. For personal computers, Microsoft Windows appeared to have some merit, although, at the time (pre-Windows 3.0), it was not considered to be stable and hence was not widely used.

By 1989, graphics API technologies had sufficiently matured to prove applicable for the system. For version 1.0 of the system, it was determined that X could provide the necessary graphics on workstation platforms. While X was only a 2D graphics system at the time (in contrast, GKS had a 3D standard defined), the 3D requirements of the system were limited and hence were feasible for implementation using X.

Work for version 1.0 of the system was performed on a VAXstation 3100 Model 38 with eight-plane graphics running VAX/VMS and DECwindows V1.0 (X11 Release 2 based). At that time, the only X Window toolkit supported by Digital Equipment Corp. (DEC) under VMS was the XUI (DECwindows) toolkit, which was only available on DEC workstations (VMS and Ultrix). Although the system could have been written utilizing only Xlib functions which would make it portable to any workstation supporting X, it was decided that doing this would greatly increase both the time to complete the system and the overall complexity of the system by requiring basic GUI functions to be written from scratch. Several public-domain toolkits were researched as to their applicability for the system, including the Athena Widget Set, the Andrew Widget Set, and HP's X Widget Set. Some time was spent porting all of these widget sets to VMS in order to more fully investigate them. During this time period, the Open Software Foundation was formed and announced that it would use a combination of the DEC XUI toolkit and the HP widget set to form a new GUI standard, called Motif. This new GUI would be available for all vendors to license and distribute which would make it more widely available, fulfilling another one of our requirements. Therefore, we decided to begin development of the system using the XUI toolkit, as it provided the most functionality of any of the toolkits. Hence, the system could be readily ported to Motif when it became available.

Unlike other toolkits, DEC's XUI had the added advantage of user interface language (UIL). UIL enabled the rapid prototyping of the user interface and reduced the amount of user interface specific C code which needed to be written. Overall, about half of the system was written in C and the other half in UIL.

The other advantage of UIL was that it could be changed independent of the C code, so we could define such things as labels, titles, behaviors, etc., in the UIL code and customize them at a later date without necessarily requiring the original C code to be modified. For the system it meant that the time required to change an aspect of the system was greatly reduced, as the entire application did not require recompiling/relinking.

Version 2.*x* of the system was implemented using Motif. With the aid of DEC's Motif Developers Kit, the process of porting the system from XUI to Motif was greatly simplified. The use of Motif for the GUI made the porting of the system to non-DEC platforms possible. Initially, the system was ported to the SUN SPARCstation (using DECwindows OSF/Motif for the SPARCstation). The current dependency on the DECwindows Motif toolkit on the SPARCstation is due to the use of DEC's *Help* widget, which is an extension to the standard Motif toolkit as distributed by OSF. Future versions of Motif may include a Help widget (as part of the standard), so this dependency on the DECwindows Motif toolkit is not expected to continue. This will allow the system to be readily moved to other platforms, including IBM RS/6000, HP, etc., in the near future.

Several techniques from the prototype and V1.0 of the system were further refined for V2.*x* including the enhancement of the coloring methods, addition of viewport pan and zoom, enhanced printscreen function, and inclusion of more sophisticated data management. In addition to workstation support added for V2.0, a port of the system was done for the PC using Microsoft Windows V3.0. Since OSF and Microsoft are cooperating in the refinement of their respective GUI's in order to make them as compatible as possible, the port of the system to the DOS platform was reasonably straightforward. Microsoft Windows V3.0 provided many of the same widgets that are available in Motif, and the appearance and behavior was generally compatible.

For version 2.*x*, the platform requirements (workstation) are a UNIX or VMS based workstation (10 SPEC89 or better), X11R3 or higher, OSF/Motif 1.1 or higher, eight-plane color graphics, minimum 16M of memory, and a minimum of 20M disk space for the system. The minimum requirements for a PC platform are a 386 running at 25 MHz with 8M of memory, eight-plane color graphics, MS-Windows 3.*x*, and 20M of disk space.

SOFTWARE SYSTEM

The FCV–false color data imaging system is divided into four main components: (1) data editing, (2) principal components analysis, (3) cluster analysis, and (4) visualization. Data editing includes data entry, data format conversion, and data scaling and is implemented by two programs: xfcv_spreadsheet, and xfcv_scale. The principal components analysis module (xfcv_plot) computes the scores for each sample. The cluster analysis module (xfcv_clustering) computes the class membership values for each sample. The visualization module (xfcv) uses both the scores from xfcv_plot and the class membership values from xfcv_clustering to perform a color rendering of the data.

The spreadsheet component of the system is implemented using a modified version of the public domain spreadsheet, sc, on Unix and VMS based systems. On personal computers using MS-Windows, a commercial spreadsheet (Lotus 1-2-3 or Quattro) is used. For all platforms, there are front-end conversion routines to make the data set readable by the spreadsheet being used.

Input of data into the spreadsheet can be in one of a number of formats. In the raw format, data are formatted as rows of data samples, with each column representing a measurement variable. The counted rows/counted columns format is similar to the raw format but there is a count of the rows, columns or both in the first line of the data set. Finally, there is the xfcv ready format. Data in this format have been processed by the conversion routines. The rows and columns are counted and the counts entered in the first line of the data file. In addition, each data sample (row) is labeled using a unique integer which represents the sample number.

The data scaling component, xfcv_scale, allows the user to carry out various scaling transformations on the data, e.g. logarithmic, autoscaling, and normalization. Although this

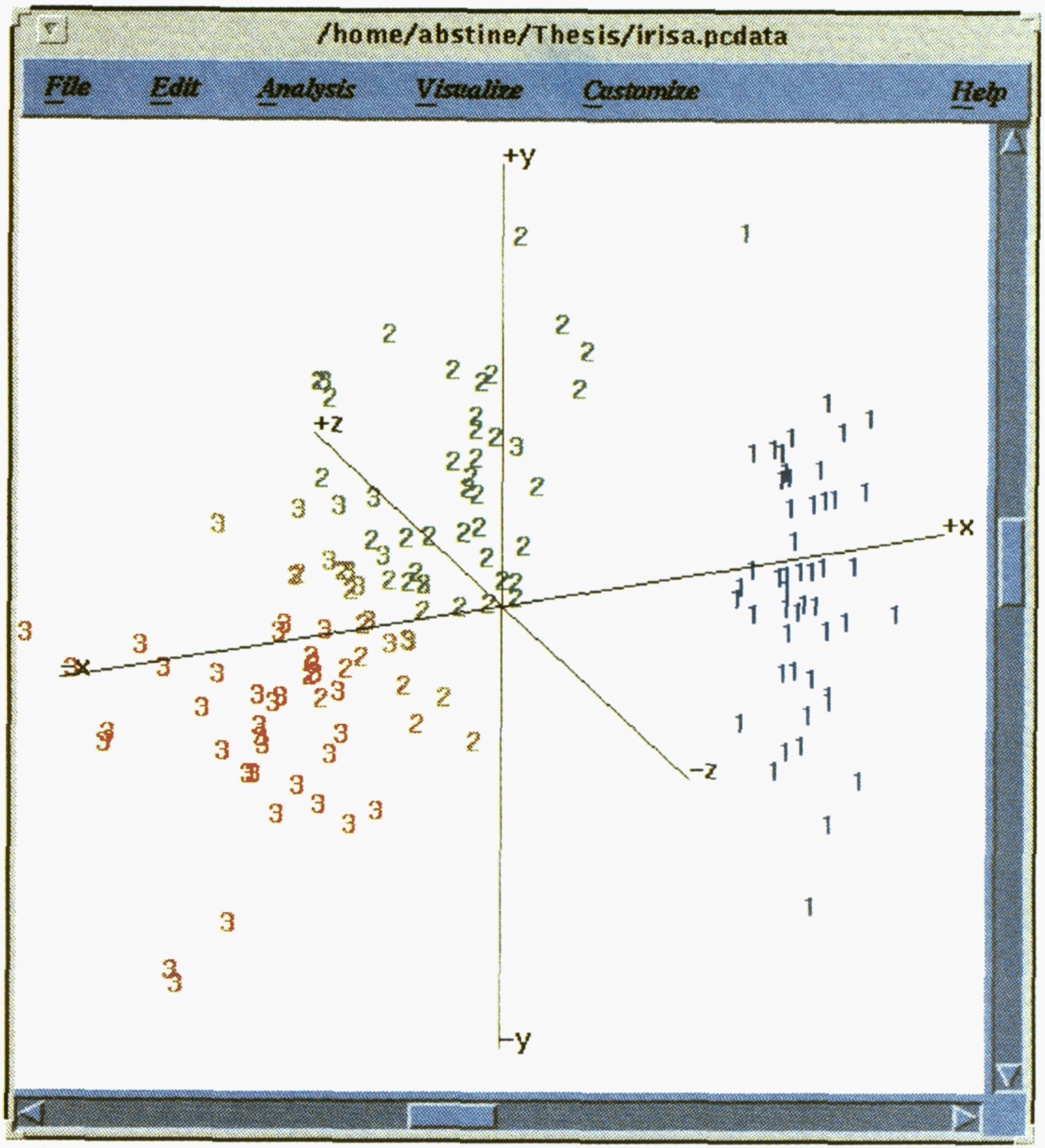

Figure 1. Three-dimensional FCV–false color data image plot of the Iris data set: 1 = Setosa; 2 = Versicolor; 3 = Virginica.

component is at present a separate program, it will become integrated with the spreadsheet in a future release.

The principal components plotting module takes as input the data set after editing/scaling and produces a data file that has the *x*, *y*, and *z* coordinates of each sample in the principal components subspace which is necessary for display of the data using the visualization component. The cluster analysis module also takes as input the data set after editing/scaling and produces two key output files: the class membership file, which is used for the color rendering in the visualization component, and the analysis report, which contains a summary of the parameters used and the results obtained from the FCV clustering algorithm.

The visualization component, *xfcv*, functions both as an integrated back-plane for accessing the other components and also as a visualization engine for the system. Some of the features of the visualization component include (1) sample labeling (sample number, class/cluster number, or some other user defined tag), (2) X, Y, Z rotations, (3) zoom/pan, (4) display animation and magnification, (5) data point editing, and (6) on-line help. Future work will be directed toward enhanced output support, context sensitive help, and data/display clipping.

APPLICATIONS OF FCV FALSE COLOR DATA IMAGING

Two data sets were used in this study to test the FCV–false color data imaging concept: (1) a set of Iris data[11–12] and (2) gas chromatographic data of neat jet fuels.[13] These data sets were chosen because of their availability and also because of the results of prior studies with which our results could be compared. FCV–false color data imaging is a method which utilizes eigenanalysis to investigate the structure of multivariate data. Hence, the scaling of the data will be critical since it will affect both the value of the within cluster scatter matrix and the variance–covariance matrix of the data. In the two studies described in this article, the data were autoscaled[14] prior to analysis; i.e., each measurement variable was adjusted so that it had a mean of zero and a standard deviation of unity. This scaling technique removes any inadvertent weighting of the variables that would otherwise arise due to differences in magnitude among the various measurements. After autoscaling, all of the measurements have equal weight and therefore equal effect in the analysis.

Starting centers for FCV–false color data imaging were chosen on the basis of *a priori* knowledge about the problem.

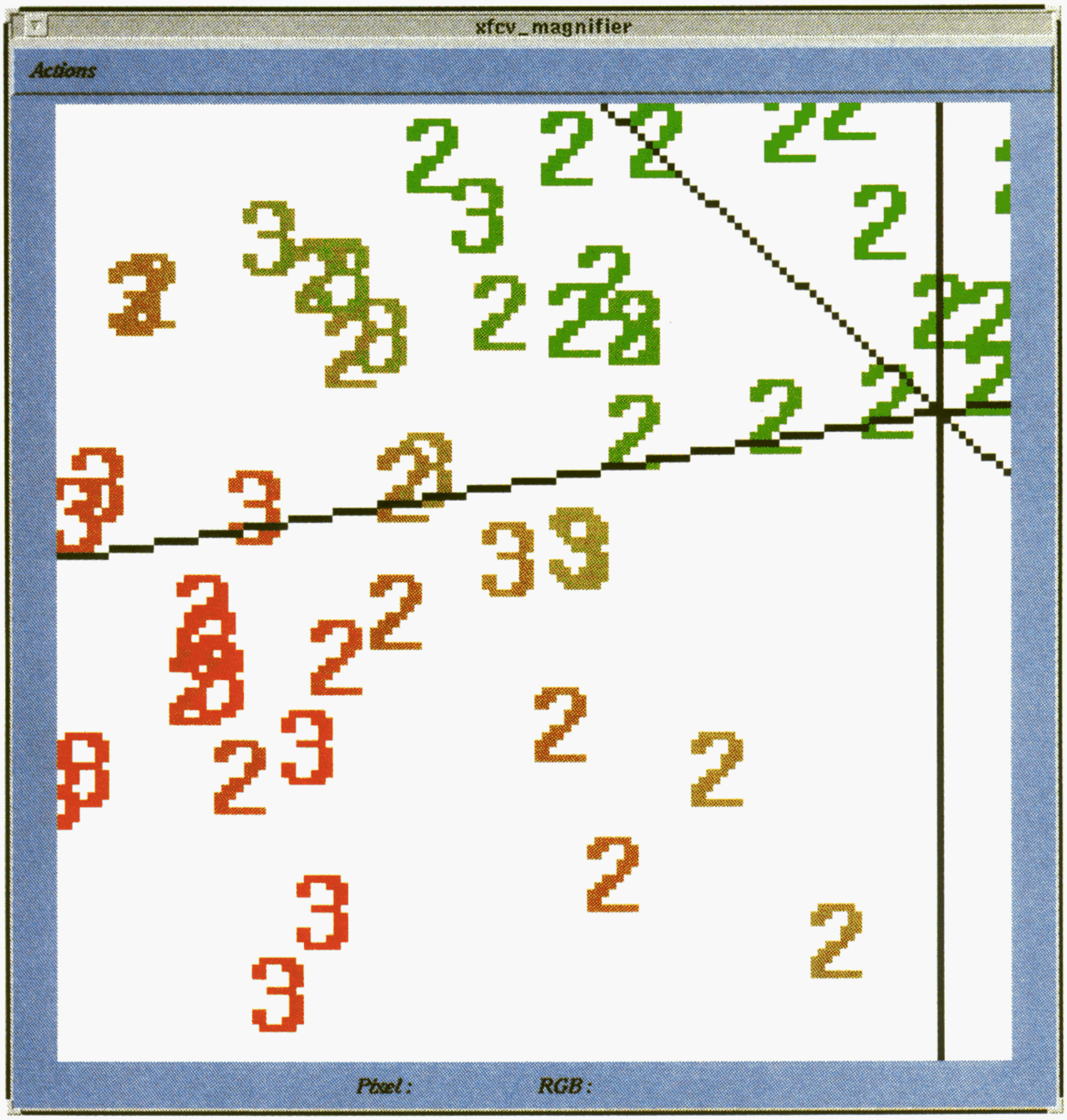

Figure 2. Magnification of the region of overlap between the 2's (Versicolor) and the 3's (Virginica) in the 3D FCV–false color data image plot. The colors brown, orange, and olive green are indicative of overlap of varying degrees between the probability distribution functions of Virginica (red) and Versicolor (green).

Samples in the data set representative of prototypical class vectors were usually selected as starting centers. In the imaging experiments, *r* was set equal to zero, *m* was set equal to 2, and the minimum prespecified change criterion for the class membership value was set equal to 0.005. These conditions were determined from our previous experience with the FCV clustering algorithm.

Iris Data. The Iris data consist of four measurements on 150 flowers. Septal and petal lengths and widths were determined for 50 flowers from each of the three varieties of Iris. The data were collected by Anderson[11] and first analyzed by Fisher[12] using linear discriminant analysis. The Iris data set has been analyzed by many other researchers over the years, making it a good data set for comparing the performance of one multivariate method against another.

Figure 1 shows a FCV–false color data image print of the Iris data set. The number of clusters searched for in the data was three, which is equal to the number of varieties of Iris; the colors assigned to the clusters were red, blue, and green. Three samples representative of the different varieties of Iris were used as starting centers in this imaging experiment. From the plot, it is evident that the 1's (Setosa) are well separated from the 2's (Versicolor) and the 3's (Virginica) in the map. However, there is overlap between Versicolor and Virginica in the map (see Figure 2), and the colors of some of the Versicolor and Virginica samples are indicative of the overlap. For example, brown and olive green are formed by mixing green (Versicolor) with red (Virginica), and samples with these colors are present in the map. The brown samples are closer to the Virginica cluster (red), and the olive green samples are closer to the Versicolor cluster (green). Evidently, the coloration in the map is actually a color image of the empirical probability density function developed for each cluster from the linear principal component models. (Each PC model

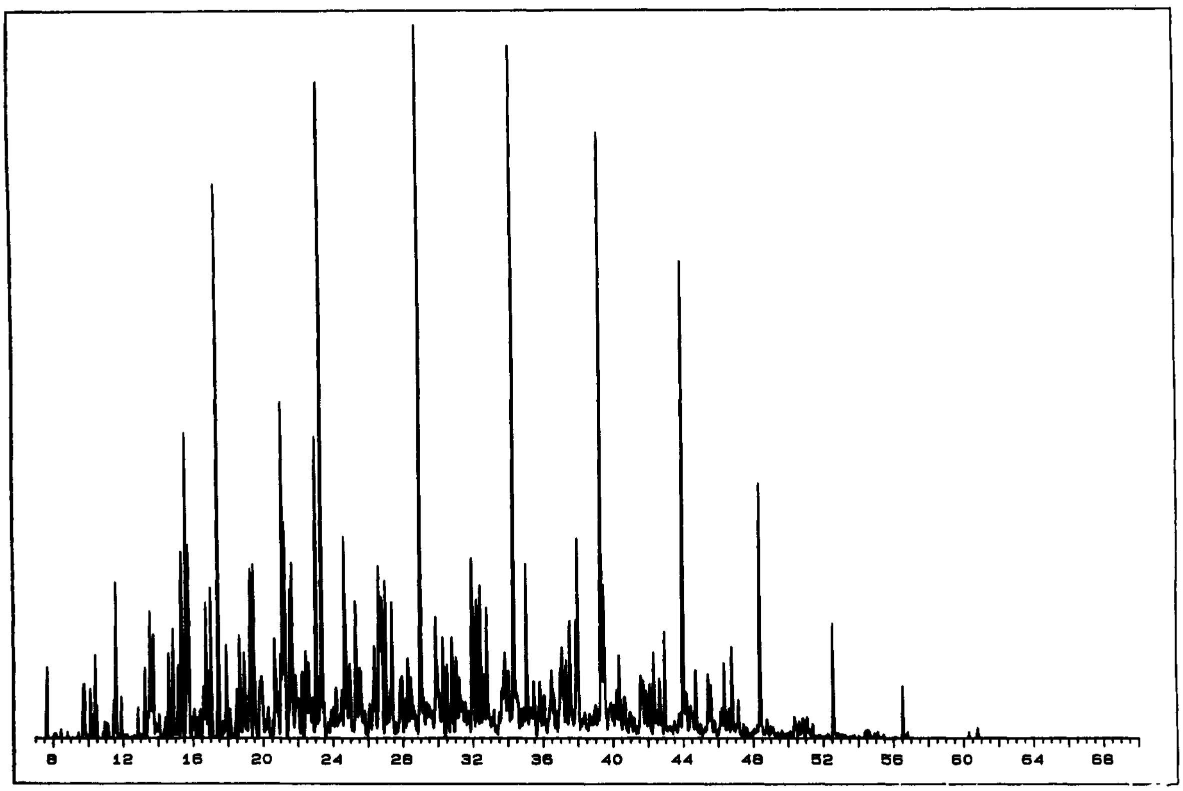

Figure 3. Total ion chromatogram representative of Jet-A fuel.

approximates the data in the same manner that a polynomial approximates bivariate data in a limited interval;[15] in other words, a PC model can be viewed as a Taylor series expansion of the data. Hence, we can treat each PC model as a probability density function describing the variation in the data of the samples comprising the cluster. The form of the probability density function used to describe the data is a set of eigenvectors that effectively span the space of the point cluster.) The mapping of probability density functions onto graphs that already convey information about the spatial distribution of the sample points in the measurement space has yielded a display of high information density.

Jet-A Fuel Data. The test data consisted of gas chromatograms of Jet-A fuel samples taken from fuel batches purchased by the Department of Defense. These fuel samples (49 in all) were actually splits taken from regular quality control standards collected over a 3-year period at Wright-Patterson Air Force Base (Dayton, OH) or Mukilteo WA Energy Management Laboratory. Prior to gas chromatography–mass spectrometric (GC–MS) analysis, the individual fuel samples were diluted with methylene chloride and spiked with anthracene-d_{10}, which served as an internal standard. The diluted fuel samples were injected directly onto the DB-5 capillary column that was temperature programmed. An HP-1000-F minicomputer was used to collect and store the GC–MS data. Further details about the collection of the GC–MS data can be found elsewhere.[16]

The chromatograms (see Figure 3) were standardized using a computer program[17] that correctly matched peaks by dividing each chromatogram into intervals defined by peaks always present (so-called marker peaks) and linearly scaling the retention times of the peaks within each interval for best fit with respect to a reference chromatogram. Prior to running the program, the marker peaks in each chromatogram were peak matched using mass spectral data. Peaks in each interval would only be matched with those in the reference chromatogram if the difference in the adjusted retention times for a given pair of peaks fell within a tolerance window which was specified by the user.

The computer program used for peak matching yielded a set of 76 standarized retention time windows. Hence, each ion chromatogram was initially represented as a 76-dimensional data vector $\mathbf{X} = (x_1, x_2, x_3, ..., x_j, ..., x_n)$ where x_j is the area of the jth peak. The set of data—49 gas chromatograms of 76 peaks each—were normalized to constant sum using the total integrated peak area and autoscaled to ensure that each chromatogram and feature (i.e. peak) had equal weight in the analysis.

In Figure 4 a 2D FCV–false color data image plot of the Jet-A fuel data is shown. For this experiment, $c = 3$, which is equal to the number of suspected Jet-A subclasses. The colors red, blue, and green were assigned to the principal component models developed from the data. The 1's in Figure 4 are chromatograms of fuel samples obtained from Wright-Patterson, and the 2's in Figure 4 are chromatograms of fuel samples obtained from Mukilteo. Even though the 1's and 2's were run at the beginning of the study, they were run on different days. (The 1's were run on one day and the 2's were run on another day.) The 3's in Figure 4 are chromatograms of fuel samples run near the end of the study, when a stricter quality control regime was imposed; i.e., a sample would be rerun if the internal standard did not appear as a well-defined peak in the chromatogram. Jet-A fuel samples comprising this group were obtained from both Wright-Patterson and Mukilteo.

The 1's, 2's, and 3's are separated from one another in the map. Since the category designations 1, 2, and 3 for the Jet-A fuel samples also denote the time period when the analysis was performed, the clustering of the data points in the map suggests that instrumental drift presumably due to aging of the capillary column is a serious problem. The confounding of an experimental artifact, e.g. column aging, with class information about the sample, e.g. type of jet fuel, is of concern

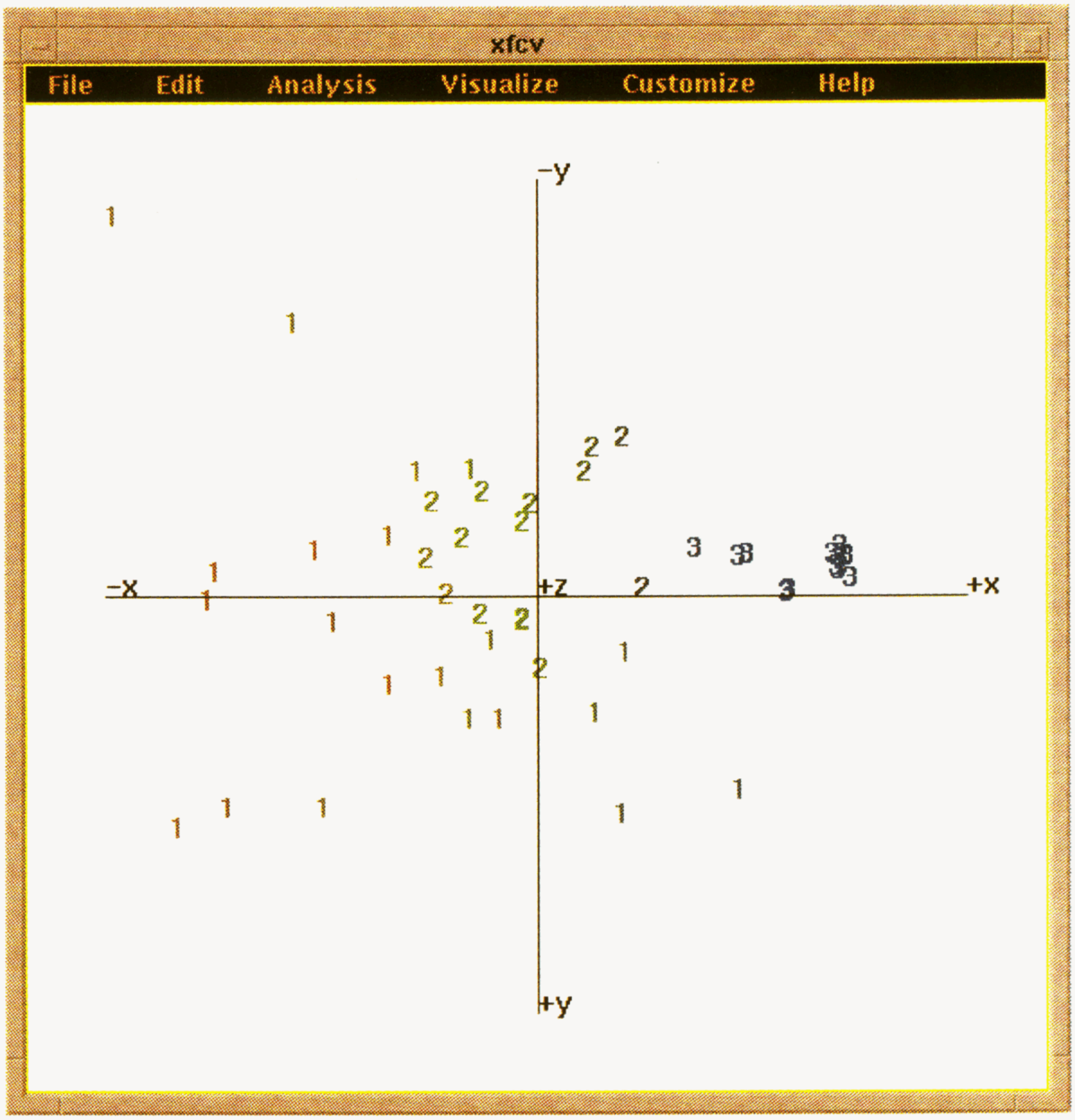

Figure 4. Two-dimensional false color data image plot of the Jet-A fuel chromatograms. The 1's and the 2's were run at the beginning of the study, and the 3's were run at the end of the study. The 1's and the 2's were run on different days.

since it can have a deleterious effect on our ability to use GC profile data to classify objects.[18]

The clustering of the data points in a principal component map according to the date when the analysis was performed is considered by some workers to be a strong indication of real differences in the GC profiles due to changes in the operating conditions of the column.[19] However, the 1's and the 2's in the map have a similar color (which is a mixture of red and green) suggesting considerable overlap between the two sets of data points in the multidimensional measurement space. Therefore, another false color data imaging experiment was performed to better understand the structure of the data. Figure 5 shows a 3D FCV–false color data image plot with rotation of the same Jet-A fuel data set. (Again, $c = 3$, and the colors red, blue, and green are used for each of the principal component models.) In this map, the location of the data points is consistent with their color. There is considerable overlap between the 1's and the 2's in the map. Furthermore, the 3's are separated from the 1's and the 2's in the map and form a compact cluster of points. The fact that the 3's are separated from the 1's and the 2's and also have a very different color is not unexpected in view of the fact that these chromatograms were run under more stringent operating conditions. In other words, the observed differences between the Jet-A fuel chromatograms are not due to aging of the GC column; rather they arise as a consequence of the adoption of a more rigorous protocol for the GC–MS analysis introduced at the end of the study. Although principal component analysis had actually overexaggerated differences in the concentration patterns between chromatograms during the projection process (see Figure 4), we were able to restore missing spatial information between individual sets of points in the high-dimensional measurement space (and also avoid making erroneous conclusions about our data) through the use of color which was provided by the membership coefficients of the FCV clustering algorithm.

CONCLUSION

FCV–false color data imaging can be an important tool for uncovering obscure relationships that are often present in complex multivariate data sets. The resulting graphic can also be used to establish the validity and relationships between

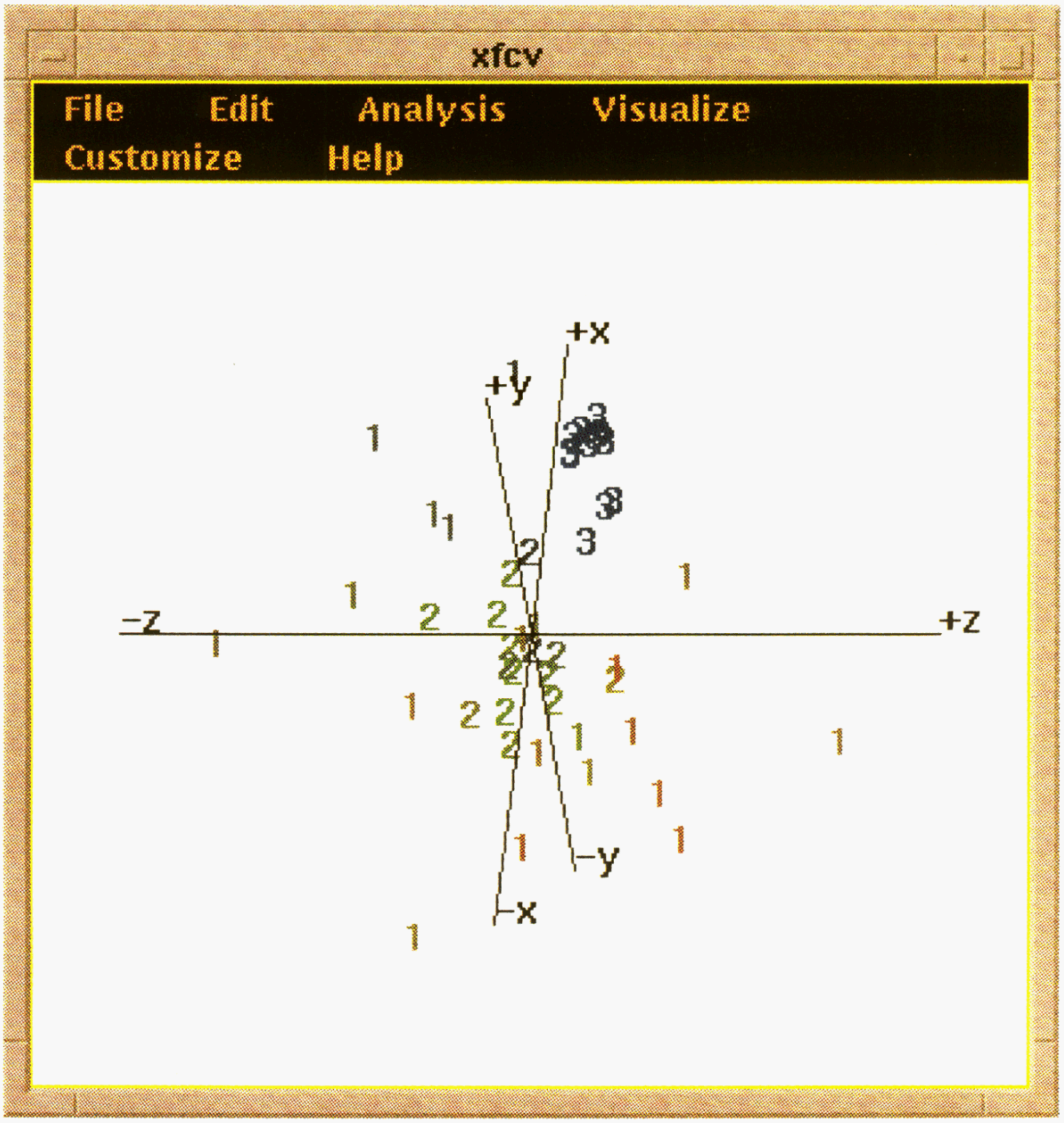

Figure 5. Three-dimensional false color data image plot with rotation of the Jet-A fuel chromatograms. The 1's and the 2's were run at the beginning of the study, and the 3's were run at the end of the study. The 1's and the 2's were run on different days.

competing partitions of the data and for understanding the basic between-class and within-class structure of a particular configuration. By examining visually a three-dimensional map that adequately represents the distribution of the data points in the high-dimensional measurement space, the chemist can assess the structural characteristics of a data set by organizing it into subgroups, clusters, or hierarchies. An advantage of the proposed methodology is that it relies heavily on graphics for the presentation of results. Our own experience has shown that graphic based methods constitute a powerful approach to data analysis because they extend the ability of human pattern recognition, allowing the chemist to play a more interactive role in the data analysis.

ACKNOWLEDGMENT

This study was supported by Contract F08635-90-C-0105 between Clarkson University and the United States Air Force. The authors thank Steven Brown (University of Delaware) and Charles Mann (Florida State University) for many helpful discussions.

REFERENCES AND NOTES

(1) Coomans, D.; Broeckaert, I. *Potential Pattern Recognition*; Research Studies Press: Letchworth, Hertfordshire, England, 1986.

(2) Tou, J. T.; Gonzalez, R. C. *Pattern Recognition Principles*; Addison-Wesley Publishing Co.: Reading, MA., 1974.

(3) Fukunaga, Keinosuke. *Introduction to Statistical Pattern Recognition*, 2nd ed.; Academic Press, Inc.: New York, 1990.

(4) Jackson, J. Edward. *A User's Guide to Principal Components*; John Wiley & Sons: New York, 1991.

(5) Bezdek, J. C.; Coray, C. R.; Gunderson, R. W.; Watson, J. D. Detection and Characterization of Cluster Substructure. I. Linear Structure: Fuzzy c-Lines. *SIAM J. Appl. Math.* **1981**, *40*, 339–357.

(6) Ekftrom, M. P. *Digital Imaging Techniques*; Academic Press, Inc.: Orlando, FL, 1984.

(7) Bezdek, J. C.; Coray, C. R.; Gunderson, R. W.; Watson, J. D. Detection and Characterization of Cluster Substructure. II. *SIAM J. Appl. Math.* **1981**, *40*, 358–372.

(8) Kreyszig, E. *Advanced Engineering Mathematics*; 4th ed.; John Wiley & Sons: New York, 1979; p 54.

(9) Lavine, B. K.; Qin, X.; Stine, A.; Mayfield, H. T. Application of Pattern Recognition Techniques to Problems in Advanced Pollution Monitoring. *Process Control Quality* **1992**, *2*, 347–355.

(10) Lavine, B. K.; Vander Meer, R. K.; Morel, L.; Gunderson, R. W.; Han, J. H.; Stine, A. B. False Color Data Imaging: A New Pattern Recognition

Technique for Analyzing Chromatographic Profile Data. *Microchem. J.* **1990**, *41*, 288–295.

(11) Anderson, E. The Irises of the Gaspe Peninsula. *Bull. Am. Iris. Soc.* **1935**, *59*, 2–5.

(12) Fisher, R. A. The Use of Multiple Measurements in Taxonomic Problems. *Ann. Eugen.* **1936**, *7*, 179–188.

(13) Lavine, B. K. Environmental Applications of Pattern Recognition Techniques. *Chemolab* **1992**, *15*, 219–230.

(14) Jurs, P. C.; Isenhour, T. L. *Chemical Applications of Pattern Recognition*; John Wiley & Sons: New York, 1975; p 30.

(15) Wold, S. A Theoretical Foundation of Extrathermodynamic Relationships (linear free energy relationships). *Chem. Scr.* **1974**, *5*, 97–106.

(16) Mayfield, H. T.; Henley, M. V. In *Classification of Jet Fuels using High Resolution Gas Chromatography and Pattern Recognition*; Hall, J. R., Glysson, G. D., Ed.; Monitoring Water in the 1990's: Meeting New Challenges; American Society for Testing Materials: Philadelphia, PA, 1991; pp 578–597.

(17) Mayfield, H. T.; Bertsch, W. An Algorithm for Rapidly Organizing Gas Chromatographic Data into Data Sets for Chemometric Analysis. *Comput. Appl. Lab.* **1983**, *1*, 130–137.

(18) Blomquist, G.; Johnson, E.; Soderstrom, B.; Wold, S. Classification of Fungi by Means of Pyrolysis-Gas Chromatography-Pattern Recognition. *J. Chromatogr.* **1979**, *173*, 19–32.

(19) Pino, J. A.; McMurray, J. E.; Jurs, P. C.; Lavine, B. K.; Harper, A. M. Application of Pyrolysis/Gas Chromatography/Pattern Recognition to the Detection of Cystic Fibrosis Heterozygotes. *Anal. Chem.* **1985**, *57*, 295–302.