THE INTERNATIONAL JOURNAL OF SCIENCE & TECHNOLEDGE

Clustering and Analysis of the NMR Spectra for Selected Organic Compound Mixtures

Ugwuonah, Linda. A.
Full Time Lecturer, Chemical Sciences Department, Godfrey Okoye University, Enugu, Nigeria

Abstract:
Thirty mixtures (two–component) of NMR Spectra of four organic compounds (Allyl alcohol, Butanol, Carvone and Ethyl Salicylate) in different proportions were made. The mixtures were processed using MestReNova. The resulting data from the mixture was analyzed using two multivariate statistical methods (Cluster observations and Principal component analysis, PCA). The analysis aimed at clustering similar spectra using a simple multivariate statistical method which could be applied to a variable size of data set with a minimum of human input. Also, this work was aimed at identifying the various two-component mixtures that contain a certain concentration of the sample organic compounds which could be classified as a cluster.

Keywords: *NMR Spectra, organic compounds, Mestre Nova, cluster observations, principal component analysis*

1. Introduction

Over the past fifty years nuclear magnetic resonance spectroscopy, commonly referred to as NMR, has become the pre-eminent technique for determining the structure of organic compounds. It can provide two - and sometimes three - dimensional information on any type of molecule as long as the compound in question possesses NMR active nuclei – almost all do. Of all the spectroscopic methods, it is the only one for which a complete analysis and interpretation of the entire spectrum is normally expected.

Although larger amounts of sample are needed than for mass spectroscopy, NMR is non-destructive, and with modern instruments, good data may be obtained from samples weighing less than a milligram. The nuclei of many elemental isotopes have a characteristic spin (**I**). Some nuclei have integral spins (e.g. I = 1, 2, 3 ….), some have fractional spins (e.g. I = 1/2, 3/2, 5/2 ….), and a few have no spin, I = 0 (e.g. ^{12}C, ^{16}O, ^{32}S …). Isotopes of particular interest and use to organic chemists are ^{1}H, ^{13}C, ^{19}F and ^{31}P, all of which have I = 1/2.

This research work is very important to an area of science referred to as metabonomics. Metabonomics is a systems approach for studying in vivo metabolic profiles, which promises to provide information on drug toxicity, disease processes and gene function at several stages in the discovery and development- process. It involves the determination of systemic biochemical profiles and regulation of function in whole organisms by analysing biofluids and tissues.

Studying the effects of drugs on whole organisms by metabonomics relies on multiparametric measurement of alterations in metabolism over time in response to a stressor or intervention. This approach can also be readily adapted to investigate the functional consequences of genetic variation and transgenesis, which is potentially of great importance in the creation and validation of new models of human disease and efficacy.

To investigate the complex metabolic consequences of disease processes, toxic reactions and genetic manipulation, non-selective, but specific, information-rich analytical approaches are required (NMR spectroscopy, pattern recognition methods, supervised and unsupervised techniques, strategy for metabonomic analysis). Several spectroscopic methods in addition to NMR can produce metabolic signatures of biomaterials, including mass spectrometry (MS), gas chromatography/mass spectrometry (GC/MS), high-performance liquid chromatography (HPLC) and optical spectroscopic techniques. Bio analytically, NMR and MS are powerful means of generating multivariate metabolic data. NMR has the advantages of being non-destructive, applicable to intact biomaterials and intrinsically more information rich with respect to the determination of molecular structures, especially in complex-mixture analyses.

The reported analyses have used only Proton NMR. Proton NMR (also Hydrogen-1 NMR, or ^{1}H NMR) is the application of nuclear magnetic resonance in NMR spectroscopy with respect to hydrogen-1 nuclei within the molecules of a substance, in order to determine the structure of its molecules. Proton NMR spectra of most organic compounds are characterized by chemical shifts in the range +12 to -4 ppm and by spin-spin coupling between protons.

In this work, thirty mixtures (two–component) of NMR spectra of four organic compounds (Allyl alcohol, Butanol, Carvone and Ethyl Salicylate) in different proportions were made and processed using MestReNova. The resulting data from the mixture was analyzed using two multivariate statistical methods (Cluster observations and Principal component analysis, PCA). The aim of the analysis is to cluster together similar spectra using a simple multivariate statistical method which could be applied to any size of data set with a minimum of human input and to identify the various two component mixtures that contain a certain concentration of the sample organic compounds which could be classified as a cluster.

2. Experimental

2.1. Sample Spectra Processing

The spectra of four compounds (represented as A for Ethyl Salicylate, B for Carvone, C for Butanol and D for Allyl Alcohol) were combined as shown in the table below:

Compounds	A	B	C	D
A	-	A + B	A + C	A + D
B	-	-	B + C	B + D
C	-	-	-	C + D
D	-	-	-	-

Table 1: Two-Component Mixtures from the Four Organic Compounds

The spectral mixture was made for the resulting six mixtures obtained as shown in the table. Five different concentrations were made for each (10:1, 3:1, 1:1, 1:3, and 1:10), resulting in a sample data set of thirty NMR spectra. The thirty mixtures were numbered 1-30 with a corresponding checklist to identify each mixture and the concentrations of the components.

The addition of the pure compounds to form mixtures and the processing of the thirty new spectra was done using Mest ReNova. MestReNova is a Nuclear Magnetic Resonance data processing, visualization, simulation, prediction, presentation and analysis software package. A summary of the steps for making the spectral mixture is described as follows:

- Open the MestReNova interface
- File - open spectra document
- Select the desired spectra and follow the menu Processing/Arithmetic
- In the Arithmetic dialog box, type in the formula for the spectrum to be added e.g. 10A + B means a mixture of A and B in the ratio of 10:1.
- Click Ok and save the resulting spectra (a .fid file)

The spectral mixture obtained was processed to fine tune the peak intensities. The following steps were taken to process the spectral mixture;

- Open .fid file (on MestReNova)
- Analysis – Reference ($CDCl_3$)
- Processing – Phase correction
  Choose “manual correction”
  (Pivot – biggest)
- Processing – Baseline – Baseline correction
  Choose “Full Automatic”
  Polynomial fit
  Polynomial order=10
  Deselect “autodetect”
- Processing – Binning
  Apply to: deselect “full spectrum”
  Choose from 0 to 10
  Choose 0.01 (high resolution)
- Processing – Normalize
  Choose “total area”
  Value, 100
- File – Save as
  File name: your sample name
  File type: ASCII file
- Open Excel software and Import ASCII file into excel. In the worksheet, you will have two columns (one for ppm values and the other for binned values)
  The binned values obtained for the sample data set was 998 over the region from 0 to 10ppm

- Set the binned value range for the chemical shift range of the solvent, $CDCl_3$ to zero (for this work the ppm range was 7.20 – 7.25)
- Set negative bin values to zero
- Transfer data to Minitab for statistical analysis by transposing excel worksheet.

2.1.1. Multivariate Statistical Analysis

All statistical analysis was performed using Minitab 15 for Windows(R) on a laptop computer, AMD Turion (tm) 64 X2 mobile technology under Windows Vista. The Excel file containing the binned values and the chemical shifts for the spectral data of the 30 mixtures was imported into minitab.

Table 2 shows the 30 spectral mixtures, the ratio of the organic compounds used to obtain the mixtures and Figure 1 shows the structure of the four organic compounds. The samples are simply tagged 1 to 30 for easy identification in the graphs obtained from the multivariate statistical analysis.

Samples	Compound Mixture	Concentration Ratio
1	Ethyl Salicylate + Allyl Alcohol	10:1
2	Ethyl Salicylate + Allyl Alcohol	3:1
3	Ethyl Salicylate + Allyl Alcohol	1:1
4	Ethyl Salicylate + Allyl Alcohol	1:3
5	Ethyl Salicylate + Allyl Alcohol	1:10
6	Ethyl Salicylate+ Carvone	10:1
7	Ethyl Salicylate+ Carvone	3:1
8	Ethyl Salicylate+ Carvone	1:1
9	Ethyl Salicylate+ Carvone	1:3
10	Ethyl Salicylate+ Carvone	1:10
11	Ethyl Salicylate + Butanol	10:1
12	Ethyl Salicylate + Butanol	3:1
13	Ethyl Salicylate + Butanol	1:1
14	Ethyl Salicylate + Butanol	1:3
15	Ethyl Salicylate + Butanol	1:10
16	Allyl Alcohol + Carvone	10:1
17	Allyl Alcohol + Carvone	3:1
18	Allyl Alcohol + Carvone	1:1
19	Allyl Alcohol + Carvone	1:3
20	Allyl Alcohol + Carvone	1:10
21	Allyl Alcohol + Butanol	10:1
22	Allyl Alcohol + Butanol	3:1
23	Allyl Alcohol + Butanol	1:1
24	Allyl Alcohol + Butanol	1:3
25	Allyl Alcohol + Butanol	1:10
26	Carvone + Butanol	10:1
27	Carvone + Butanol	3:1
28	Carvone + Butanol	1:1
29	Carvone + Butanol	1:3
30	Carvone + Butanol	1:10

Table 2: Components of mixture for the 30 samples and their concentrations

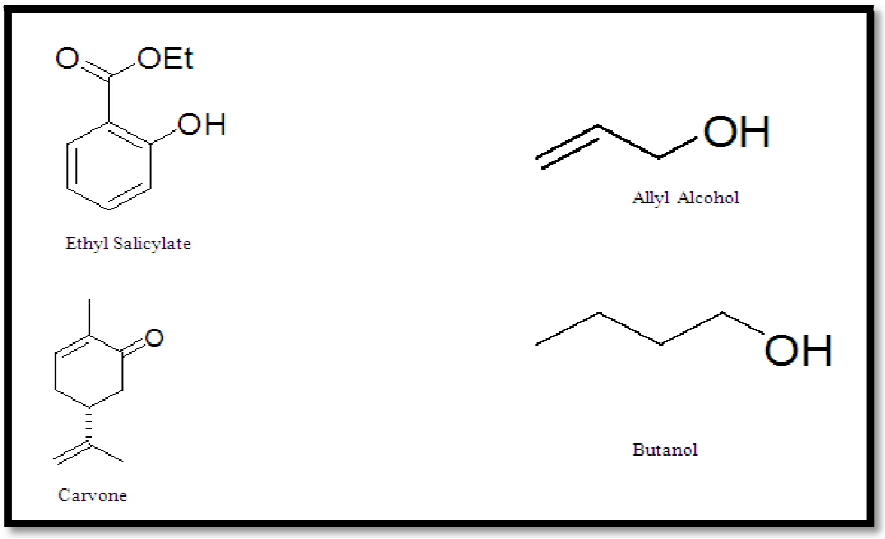

Figure 1: Structures of the Four Organic Compounds

The imported excel sheet into Minitab was transposed before any multivariate statistical analysis was done. The resulting data were subjected to two multivariate statistical analysis, Cluster observations and Principal component analysis, PCA.

2.1.2. Cluster Observations

Cluster observations employed an agglomerative hierarchical method that begins with all observations being separate, each forming its own cluster. In the first step, the two observations closest together are joined. In the next step, either a third observation joins the first two, or two other observations join together into a different cluster. This process will continue until all clusters are joined into one.

Cluster observations analysis in this work involved the following steps;

- Open the Minitab software
- File > Open worksheet > Data > Transpose columns
- Stat > Multivariate > Cluster Observations
- Variables > Linkage Method (single) > Distance measure (Euclidean) > Number of clusters (4) > Show Dendogram

2.1.3. Principal Component Analysis

PCA is a useful statistical technique that has found application in fields such as chemometrics, engineering, computer vision and other applied sciences. It is a common technique for finding patterns in data of high dimension and is used to form a smaller number of uncorrelated variables from a large set of data. The goal of principal components analysis is to explain the maximum amount of variance with the fewest number of principal components. Principal components analysis is commonly used as one step in a series of analyses. Two types of matrix can be used when calculating principal components, the correlation matrix which is used if it makes sense to standardize variables (the usual choice when variables are measured by different scales) and the covariance matrix which is used if you do not wish to standardize variables.

The covariance matrix was used for the analysis of the sample data set in this project. Covariance values can range from negative infinity to positive infinity. Positive covariance values indicate that above average values of one variable are associated with above average values of the other variable and below average values are similarly associated. Negative covariance values indicate that above average values of one variable are associated with below average values of the other variable.

Principal components analysis in this work involved the following steps;

- Open Minitab software
- File > Open worksheet > Data > Transpose columns
- Stat > Multivariate > Principal Components
- Variables > Number of components to compute (2) > Type of Matrix (covariance) > Graphs (scree plot and score plot)

Further analysis was performed on the principal components (PC1 and PC2) obtained from the above procedure. The steps involved are as follows;

- Copy text results (variable and PC1) from PCA work project window
- Paste result in a new minitab worksheet
- Graph > Bar Chart > Bars represent (values from a table) > Simple > Graph variables (PC1) > Categorical variable (variable)

- Repeat the procedure for variable and PC2

Two different charts were obtained for PC1 and PC2.

3. Results and Discussion

This study used the proton nuclear magnetic resonance spectra of four pure organic compounds. Two component mixtures of the spectra were made for five different concentrations. A total of 30 new spectral mixtures were obtained on which processing and multivariate statistical analysis (cluster observations and principal component analysis) were performed.

Spectra processing included setting all negative bin values and bin values for the solvent region to zero. For the multivariate statistical analysis results obtained showed that the criterion for the clustering of the sample spectral mixtures was based on the concentration of organic compound component in the mixture and the nature of the organic compound. Thus, compounds with a particular concentration ratio of one of the four organic compounds were found to be in the same cluster.

3.1. Cluster Observations

Figure 2 shows the dendrogram for the cluster observations multivariate statistical analysis obtained for the sample 30 spectral mixtures. The cluster size chosen for the analysis was four, corresponding to the four different background colours for the clusters on the dendrogram (Red, Green, Orange and Blue).

The y-axis of the graph represents similarity level. This is an estimation of the percentage of closeness or similarity of the mixtures that belong to a cluster while the x-axis shows the clustering of sample mixtures 1 to 30 whose components have been identified in figure 2.

For the first cluster (red background colour), the similarity level for the two samples is about 78% and the samples 1 and 2 corresponds to a mixture of Ethyl salicylate and Allyl alcohol in the ratios 10:1 and 3:1 respectively. In other words, the two samples have higher concentrations of Ethyl salicylate in their mixture.

The second cluster (green background colour) contains seven samples (3, 4, 5, 16, 21, 17 22). The general similarity level for all seven compounds is about 80% and they contain a high concentration of Allyl alcohol. Samples 3, 17, 22 stand on their own in the cluster, corresponding to three different mixtures of different concentrations. Samples 4 and 5 belong to a cluster (concentration ratio for allyl alcohol is 3 and 10 respectively) while samples 16 and 21 belong to a cluster (concentration ratio of allyl alcohol for both is 10).

The third cluster (orange background) contains 17 samples (8,9,10,20,26,19,27,13,14,15,25,30,24,29,23,28,18). The general similarity level of the cluster is about 74%. Samples 8,9, 10, 20,26,19,27 are all in a group within the cluster and they have higher concentrations of carvone in common as part of their mixtures.

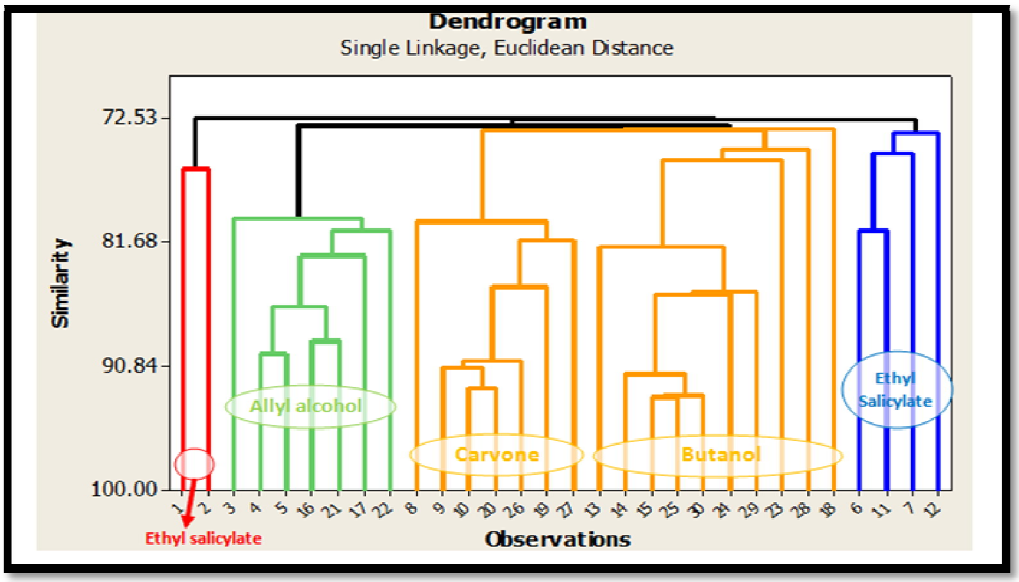

Figure 2: Cluster Observations Dendrogram

Samples 13,14,15,25,30,24,29 as a group within the cluster all have higher concentrations of butanol as part of their mixtures. However, samples 23, 28 and 18 stand on their own within the cluster but each contains one of carvone, butanol or allyl alcohol as part of their mixture. Worthy of note is that for these three compounds, the ratio of the two components of the mixture is 1:1 respectively.

The fourth cluster contains four samples (6, 11, 7, and 12) all of which contains higher concentrations of Ethyl salicylate. Sample 6 and 11 are in one group within the cluster because they both contain the same ratio of concentration of Ethyl salicylate, precisely 10 while samples 7 and 12 stand on their own within the cluster.
In general, there is a marked criterion for the clustering of the spectral mixture, considering the sample components or the concentration of the sample organic compound contained in the spectral mixture.

3.2 Principal Components Analysis (Score Plot and Scree Plot)

The principal components analysis score plot obtained for the sample data set is shown in Figure 3. The score plot is a graph of the second principal component versus the first principal component. The red dots represent the position of each sample on the graph while the number on each point represents the sample. The samples that are closer together are classified to be more similar than the ones that are farther apart.

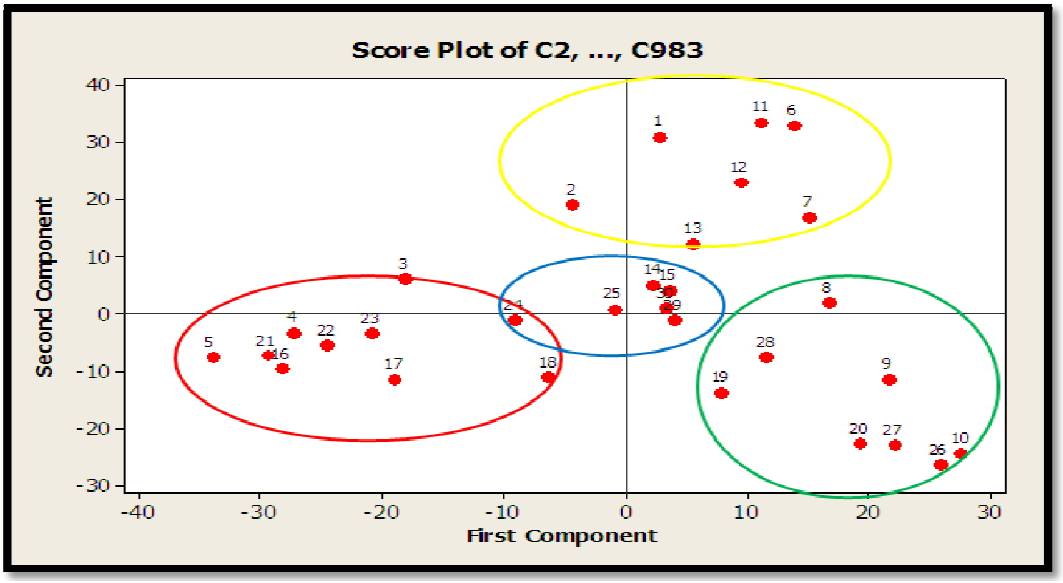

Figure 3: Principal Components Analysis (Score Plot)

The criteria for this classification are based on the principal components 1 and 2 for the sample. Thus, some samples almost form clusters as shown on the score plot. The sample cluster within the red, blue, yellow and green circle corresponds to high concentrations of allyl alcohol, butanol, ethyl salicylate and carvone respectively. In general, however, the similarity level of the whole 30 samples is comparable.

The scree plot (Figure 4) was obtained for the principal components of the sample data set. Scree plot displays the eigenvalues associated with a component or factor in descending order versus the number of the component or factor. The eigen analysis of the covariance matrix obtained for the sample shows that the first five eigenvalues accounts for the total variance (100%)

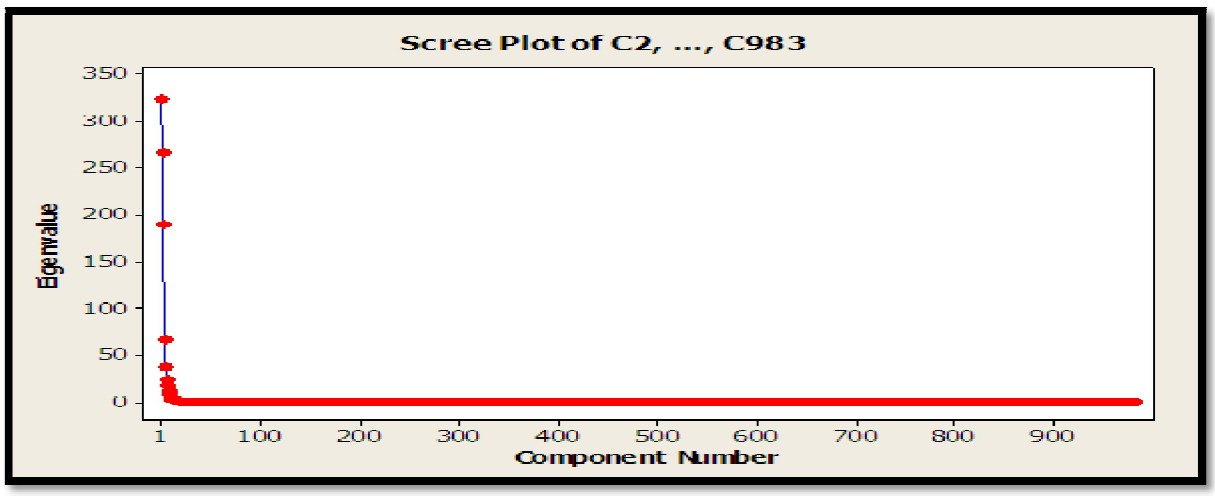

Figure 4: Principal Components Analysis (Scree Plot)

From the eigen analysis of the covariance matrix, it is observed that for the first principal component, the eigenvalue is 76.122 and accounts for 39.6% of the total variance, the second principal component has an eigenvalue of 69.520 and accounts for 36.1% of the total variance, the third eigenvalue is 45.172 and accounts for 23.5% of the total variance, the fourth eigenvalue is 1.254 and accounts for 0.7% of the total variance, the fifth eigenvalue is 0.373 and accounts for 0.2% of the total variance. The remaining principal components account for a very small proportion of the variability and are probably unimportant. Thus, most of the data structure can be captured in four or five underlying dimensions This information is shown visually on the screen plot, where the first five red dots are apart visible while the others are at one point on the graph and are clustered.

Principal Component Analysis:

C2, C3, C4, C5, C6, C7, C8, C9, C10, C11, C12, C13, ...

Eigen value	**76.122**	**69.520**	**45.172**	**1.254**	**0.373**
Proportion	0.396	0.361	0.235	0.007	0.002
Cumulative	0.396	0.757	0.991	0.998	1.000

Table 3: Eigen analysis of the Covariance Matri

Figures 5 and 6 respectively show the spectral pattern for principal components 1 and 2 respectively. Analysis of this spectral pattern showed that the peaks of the spectra correspond to that of the four organic compounds from which the sample data set were obtained. An NMR pattern for the two PC charts is shown in Figure 7.

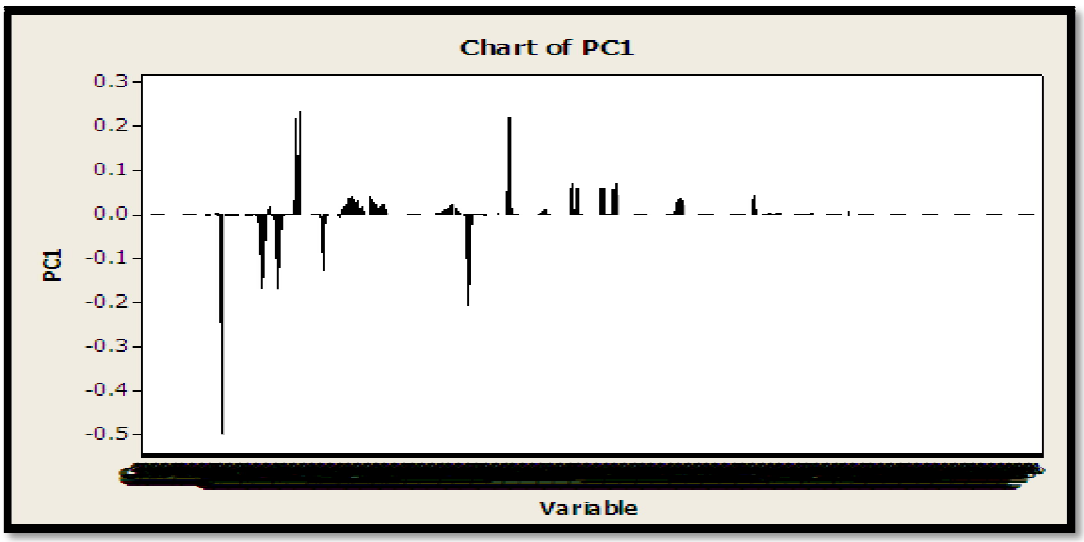

Figure 5: Chart for PC1

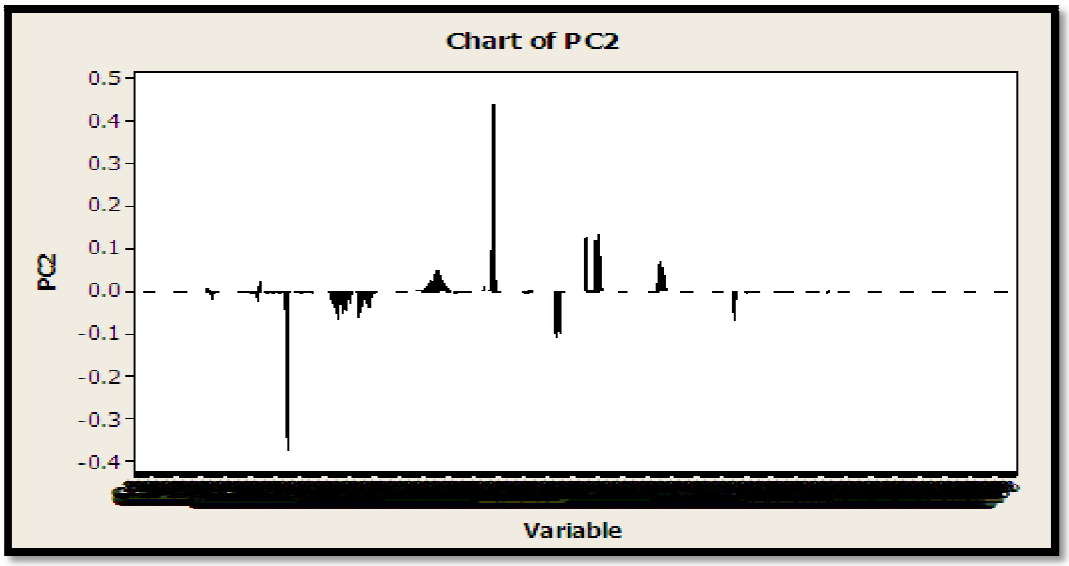

Figure 6: Chart for PC2

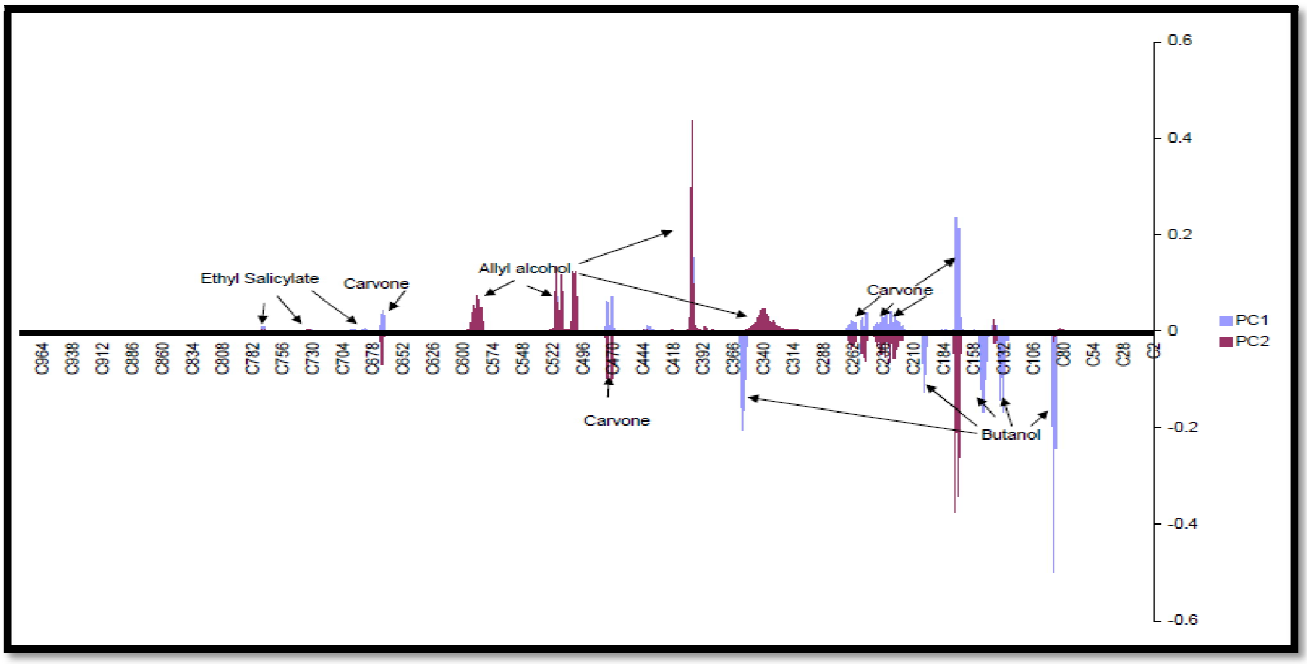

Figure 7: NMR Interpretation for PC1 and PC2

4. Conclusion

In comparison to a similar work that has been done in this area, a multivariate statistical method has been found to cluster NMR data. The two multivariate statistical analysis methods used for this study were useful in terms of clustering the NMR data of the mixtures for similarity or diversity purposes. The method does not need prior knowledge of the data and the results were statistically valid. The validity of the results is based on the regular patterns observed for the two multivariate statistical methods used for the analysis of the sample data set.

The cluster observations were useful in clustering the mixtures based on the uniformity of the concentration of the components as well as the specific sample organic compound available as part of the mixture. The principal components analysis (score and scree plot) was useful in assessing the variability of the data set (bins).

Further work on this study is recommended in which case other multivariate statistical methods can be applied to more complex mixtures (three or four components mixtures) and other natural product extracts.

5. References

i. Gregory K Pierens, Meredith E Palframan, Carolyn J Tranter, Anthony R Carroll and Ronal J Quinn, A robust clustering approach for NMR spectra of natural product extracts. 2005, 43, 359-365

ii. Henning Risvik, Principal Components Analysis (PCA) & NIPALS Algorithm. 2007, 1-6 Retrieved April 2, 2009 from http://folk.uio.no/henninri/pca_module/pca_nipals.pdf

iii. http://www.ch.ic.ac.uk/local/organic/tutorial/steinke/ "Foundation Course NMR Workshop.pdf" (2003).

iv. Jeremy K Nicholson, John Connelly, John C Lindon and Elaine Holmes, Metabonomics: a platform for studying drug toxicity and gene function.2002, 1, 153-160

v. Lindsay I Smith, A Tutorial on Principal components analysis. 2002, 1-11Retrived April 2, 2009 from http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

vi. MestRe Nova – 5.3.0 – 4385_Manual.Pdf, 2008 Mestre Lab Research Minitab Help

vii. "Wikipedia, The free encyclopedia" (2010). http://en.wikipedia.org/wiki/Proton_NMR