DOI 10.5012/bkcs.2010.31.11.3151

Application of Clustering Methods for Interpretation of Petroleum Spectra from Negative-Mode ESI FT-ICR MS†

Injoon Yeo, Jae Won Lee,‡ and Sunghwan Kim*

*Kyungpook National University, Department of Chemistry, Daegu, Korea. *E-mail: sunghwank@knu.ac.kr*
‡Korea University, Department of Statistics, Seoul, Korea
Received August 14, 2010, Accepted September 30, 2010

This study was performed to develop analytical methods to better understand the properties and reactivity of petroleum, which is a highly complex organic mixture, using high-resolution mass spectrometry and statistical analysis. Ten crude oil samples were analyzed using negative-mode electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry (ESI FT-ICR MS). Clustering methods, including principle component analysis (PCA), hierarchical clustering analysis (HCA), and *k*-means clustering, were used to comparatively interpret the spectra. All the methods were consistent and showed that oxygen and sulfur-containing heteroatom species played important roles in clustering samples or peaks. The oxygen-containing samples had higher acidity than the other samples, and the clustering results were linked to properties of the crude oils. This study demonstrated that clustering methods provide a simple and effective way to interpret complex petroleomic data.

Key Words: Petroleomics, FT-ICR MS, Clustering analysis, ESI

Introduction

The world's remaining oil deposits are becoming heavier because sweet oil is being rapidly depleted. The deposits of shale oil, one form of heavy oil, are equivalent to the known resources of total conventional oil (~2.4 trillion barrels).[1] Despite their abundance, it is very difficult to use heavy crude oils because of their complexity and abundance of heavy heteroatoms. These heavy components can cause difficulties in processing of crude oils. Eventually, the ability to use heavy crude oils will become very important. There is a demand to understand the chemical compositions of heavy components to overcome processing difficulties. Petroleomics has been used to study petroleum composition. Heavy components of crude oil are studied using negative-mode electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry (ESI FT-ICR MS) in petroleomic studies.[1] FT-ICR MS was shown to be effective for identifying the molecular composition of crude oils.[2-12]

One of the most important goals of petroleomics is to predict the reactivities and properties of crude oils based on molecular information, which is obtained using high-resolution mass spectroscopy. Interpretation of high-resolution mass spectra can cause difficulties because of the large amounts of information contained in each spectrum. Each high-resolution mass spectrum of crude oil can routinely contain thousands of peaks. To find a link between complex mass spectra and the reactivity and properties of crude oil, it was necessary to develop and apply statistical tools to handle large amounts of data. For example, statistical analysis was successfully applied to find a correlation between the atmospheric pressure photoionization (APPI) mass spectra and the properties of 20 crude oils. In another study, principle component analysis (PCA) and hierarchical clustering analysis (HCA) were used to examine the overall distributions of peaks and samples and to determine the relationships between samples based on selected groups of peaks. PCA and HCA are regarded as clustering methods. Clustering is a process of grouping peaks or samples based on their similarities. As no prior knowledge about the data is required, clustering is known as an unsupervised method. Although the two clustering methods, PCA and HCA, were successfully used in previous study,[13] no systematic investigations to find the best method for petroleomic studies have yet been reported.

In this study, various statistical clustering techniques, including PCA, HCA, and *k*-means, were applied for comparative analysis of the electrospray ionization (ESI) mass spectra of 10 crude oil samples.

Experimental

Mass spectrometry. Ten oil samples with relatively high and low sulfur content or TAN value were selected for this study. Ten crude oil samples (labeled Crude01 – Crude10) and their bulk properties, including sulfur content and total acid number (TAN), are listed in Table 1. The sulfur contents were obtained by elemental analysis[14] and TAN values were determined by the commonly used titration method.[15] Samples were prepared by dilution to 0.5 mg/mL in a 50:50 v/v solution of toluene/methanol. The diluted samples were spiked with 4 μL of 30% NH_4OH to ensure adequate ionization efficiency for negative-ion ESI. The prepared samples were directly injected with a syringe pump (Harvard, Holliston, MA, USA) at a flow rate of 70 μL/h for ESI analysis. Analyses were performed using a 15 T FT-ICR mass spectrometer at the Korean Basic Science Institute (KBSI, Ochang-eup, Korea). The Apex hybrid Qq-FT instrument was equipped with a Bruker Apollo II dual source. Nitrogen was used as the drying and nebulizing gas. The operating parameters for ESI analysis were a drying gas temperature of 200 °C with

†This paper is dedicated to Professor Hasuck Kim for his outstanding contribution to electrochemistry and analytical chemistry.

Table 1

	Crude01	Crude02	Crude03	Crude04	Crude05	Crude06	Crude07	Crude08	Crude09	Crude10
Sulfur (%)	2.87	3.57	2.01	3.53	0.18	4.79	3.77	0.13	0.25	4.5
TAN (mg KOH/g)	0.29	0.45	0.18	0.47	0.79	0.27	0.3	0.58	3.15	3.5
Nitrogen (ppm)	1645	4019	4011	3123	3532	2136	1620	949	4405	-

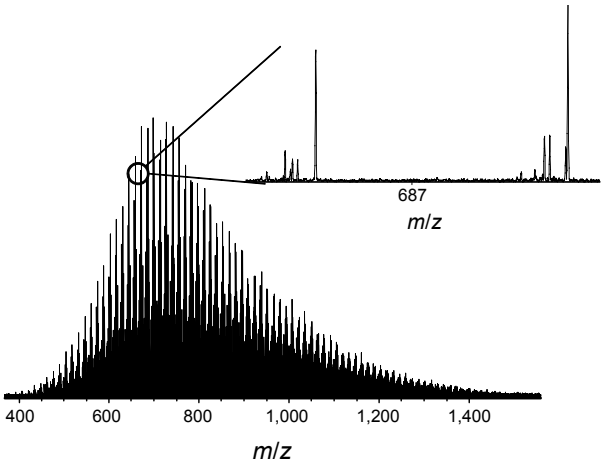

Figure 1. (-) mode ESI spectrum of crude oil sample 3 and the expanded view at 687 *m/z* region.

a flow rate of 1.5 L/min and a spray voltage of 3000 - 4000 V; the skimmer voltage was set to 13.0 V to minimize in-source fragmentation. Ionized samples were stored in an argon-filled collision cell for 1 s and transferred to the ICR cell with a 2 ms time-of-flight window. Both sidekick and gated trapping approaches were used. A sidekick voltage of 20 V was used to initially trap the ions. After transferring the ions to the ICR trap, the trap voltage was raised to 3 V and ramped down to 1.5 V for detection. At least 100 scans were accumulated and averaged to improve the signal-to-noise ratio of the resulting spectra. For each spectrum, at least 2×10^6 data points were recorded. A resolving power greater than 300,000 at ~400 *m/z* was routinely achieved for the obtained spectra. An example of mass spectrum and its expanded view are displayed in Figure 1.

Mass spectra collected in this study were calibrated in two steps. In the first step, an internal calibrant material (G2421A electrospray "tuning mix" from Agilent, Santa Clara, CA, USA) was used as a mass internal calibration standard. The internal standard was added to the samples just before analysis. In the second step, the samples were analyzed without the internal calibrant to eliminate possible peak contributions from the calibrant material. The *m/z* scale of the resulting spectra was first calibrated using an external calibrant material (G2421A electrospray "tuning mix" from Agilent) and second, using the exact *m/z* numbers of major sample peaks obtained from the first spectrum.

Approximately 3000 - 4000 peaks were found in the mass spectra of each sample. Initially, the threshold for peak picking was a signal-to-noise ratio higher than 4.5. An automated peak-picking algorithm was later implemented for more reliable and faster results. After peak picking, elemental formulae were calculated and assigned based on *m/z* values. Peak assignments were performed within an error range of 1 ppm. Normal conditions for petroleum data ($C_cH_hN_nO_oS_s$, *c* unlimited, *h* unlimited, $0 \le n \le 5, 0 \le o \le 10, 0 \le s \le 2$)[12] were used in these calculations. Typically, more than 98% of the observed peaks were successfully assigned with elemental formulae.

Statistical analysis. The 10 resulting peak lists with elemental formulae assignments were merged into a single table for statistical analysis. Detailed information on data preprocessing can be found in the previous literature.[13] Briefly explained, each sample was analyzed in triplicate for statistical significance. Triplicate spectra were later combined into a single spectrum, resulting in 10 mass spectra and an equal number of peak lists. The relative abundance of individual peaks was normalized to the summed relative abundance of each peak list. The resulting peak lists with elemental formulae assignments were merged into a single table for statistical analysis. Data merging was performed based on assigned elemental formulae, which were equivalent to theoretical mass values.

Two normalization methods (quantile and central tendency normalization) were applied to remove systematic experimental variation.[16] Then, the peaks that best explained sample variance were selected for further statistical analyses. Spectral interpretation and statistical computations were performed with R version 2.9.0.[17]

PCA, HCA, and *k*-means clustering were applied to the processed data. Detailed descriptions of the methods have been reported previously.[13,18] Briefly, principle component analysis is used to reduce the dimensionality of data while retaining the variation in the original data as much as possible. Principal components (PCs) are generated by linear combination of the original data set. The results of PCA are usually represented by score and loading plots. In a score plot, the distribution of objects is presented based on PCs, while a loading plot displays the distribution of variables. Clustering can be done based on the distributions of objects and variables in each plot.

In hierarchical clustering, a progressive combination of variables that are most similar is performed. The similarity is calculated from the distance between all data points using distance metrics. The results are usually plotted in a dendrogram representing the clusters and relations between them.

In *k*-means clustering, the number of clusters is first estimated by the user. The results of the *k*-means algorithm can depend on the initial clusters. Therefore, it is very important to choose optimum values for the clusters. Often, another algorithm is employed to find the optimum number (*k*). In *k*-means clustering analysis, the algorithm progressively moves each point to each cluster to maximize the differences among clusters. It then assigns each observation to clusters based on the mean of the cluster. The cluster mean is then recomputed, and the process begins again. The *k*-means clustering algorithm is one of the simplest and fastest clustering algorithms.

Results and Discussion

PCA of crude oil spectra. PCA was performed on 10 crude oil spectra. The loading and score plots are shown in Figures 2a and b. Score and loading plots are typically used to display the results of PCA. From the loading plot (Fig. 2a), it was clear that samples could be divided into three groups: Groups I, II, and III. Group I was composed of crude05, crude08, crude09, and crude10 samples. The score plot from PCA of peaks in this study is presented in Figure 2b. In this study, each class was represented by showing only hetero atoms in the compounds. For example, O and NO classes in the Figure 1 each represent compounds with chemical formulae of C_cH_nO and C_cH_nNO. In the figure, each quadrilateral represents the elemental composition of 10 crude oil samples, and they were color-coded according to the class of compounds. For example, the O class is shown in red, the OS class in black, the NO class in pink, the NS class in green, and the NOS class in light blue. In Figure 1b, most of the O class peaks were located in the A cluster. The B and C clusters also had O class compounds. Sulfur-containing classes (e.g., OS, NS, NOS) were mainly located in the B cluster. Comparison between loading and score plots helped to visualize the relationships between the samples and the observed peaks.

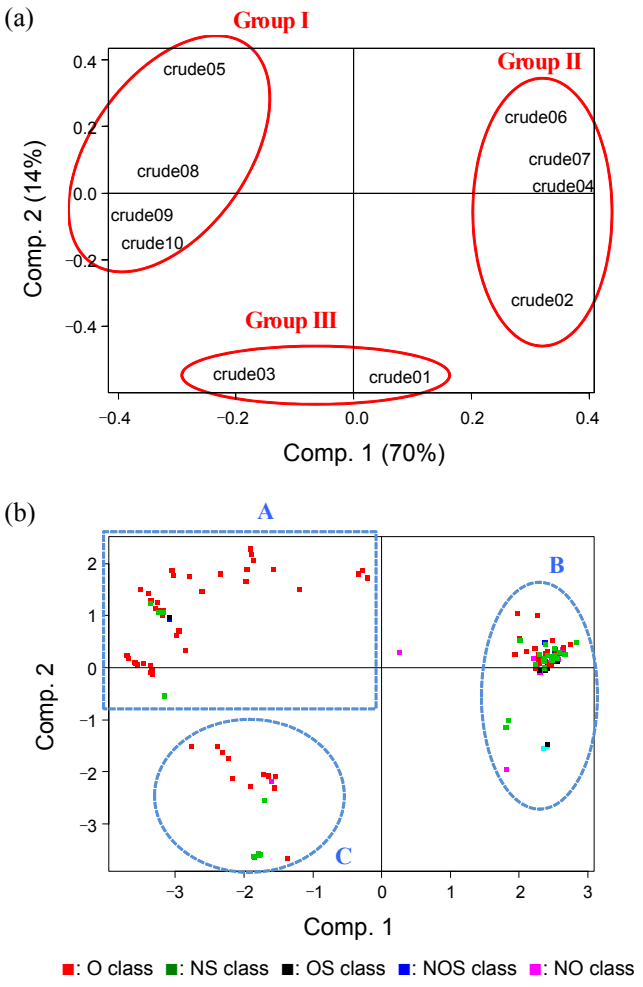

Figure 2. Score (a) and loading (b) plots are shown for PCA of 10 ESI negative-mode spectra.

Thus, it was clear that samples in Group I were abundant in O class compounds, while Groups II and III were abundant in sulfur-containing compounds.

Table 1 shows chemical properties of crude oils related to classes of compounds and samples shown in Figure 1. The samples in Group I had TAN values larger than 0.5 (refer to Table 1), while those in Groups II and III had TAN values lower than 0.5. In addition, samples in Groups II and III had higher sulfur contents than those in Group I. The only exception was the crude10 sample-this sample was included in Group I despite its very high sulfur content. The separation between Groups II and III was attributed to sulfur content. The crude01 and 03 samples in Group III had sulfur contents between 2 and 3. The crude02, 04, 06, and 07 samples in Group II all had sulfur content higher than 3. Thus, it was clear that the TAN value and sulfur content played important roles in grouping crude oil spectra obtained using negative mode ESI.

In a previous study of positive and negative-mode atmospheric pressure photo ionization spectra,[13] sulfur and nitrogen contents and TAN values were important in grouping samples. In this study, nitrogen content was not important for grouping samples. The crude01, 07, and 08 samples had the lowest nitrogen content. However, the three samples were scattered into three groups in Figure 1a, and this was attributed to the ionization method used in this study. In general, negative-mode ESI is not sensitive toward basic nitrogen compounds. Nitrogen compounds are typically observed as NO, NS, or NSO classes in negative-mode ESI. Thus, nitrogen compounds alone did not differentiate samples into groups.

HCA of crude oil spectra. In this study, hierarchical clustering was performed on 10 crude oil negative-mode ESI spectra, and results are shown as a heatmap and dendrogram in Figures 3a and b. The Spearman correlation coefficient was used as the basis for HCA. The coefficient was calculated using the following equation

$$R = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)} \quad (1)$$

where d^2 is the square of the difference between the ranks of the two values and n is the number of variables. The coefficient was used to calculate distance between variables. The distance was used as a measure to determine similarities and differences between variables and to determine which variables were grouped together.

The abscissa of the heatmap in Figure 3a represents 10 crude oil samples. The clustering between crude oil samples was accomplished by comparing similarities and differences in peak distributions. The results are presented as a dendrogram located at the top of the heatmap. In this analysis, the samples were divided into three groups, and the grouping was identical to that observed using PCA. In HCA, the peaks observed in the samples were also grouped, as shown in the dendrogram of Figure 3b. Chemical components of crude oil samples were separated into two clusters. Closer examination of the elemental formulae of the peaks showed that grouping occurred because of the oxygen and nitrogen-containing compounds. This also

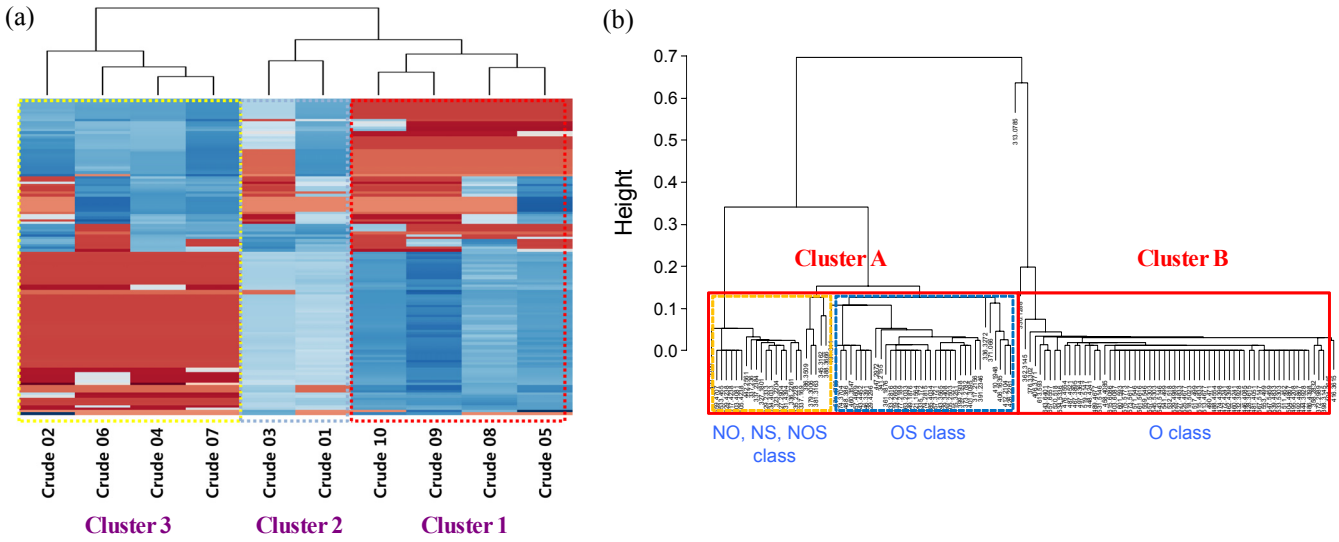

Figure 3. Heatmaps and dendrograms show the HCA results showing clustering of the samples (a) and peaks (b).

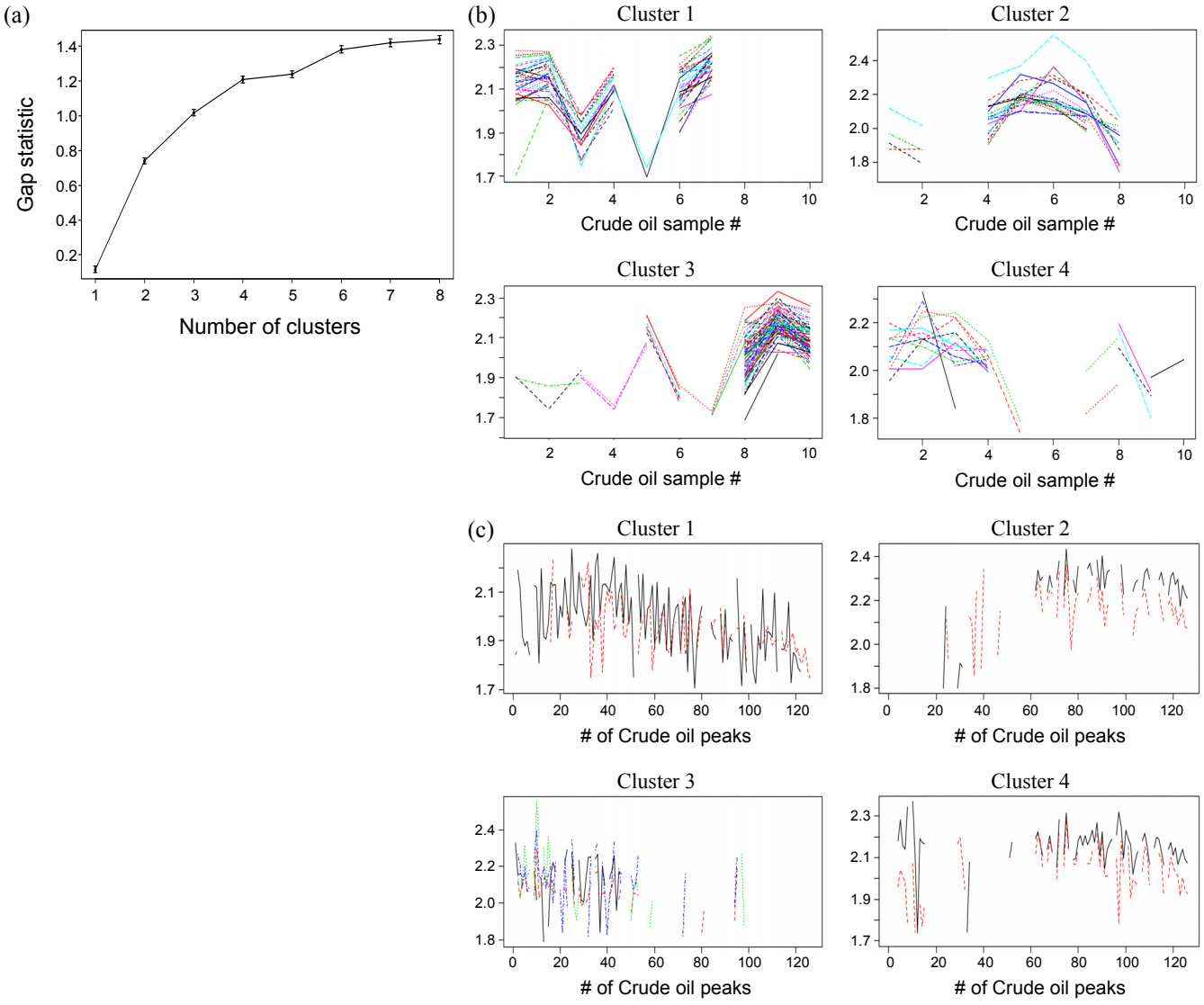

Figure 4. Gap curve (a) to determine the number of clusters and result of *k*-means clustering analysis showing clusters of peaks (b), and samples (c).

agreed well with results obtained using PCA.

Gap statistic and *k*-means clustering. In *k*-means clustering, the number of clusters is predetermined and specified by the user. In this study, the number of clusters was determined using the gap statistic,[19] which is commonly used to decide the number of groups for *k*-means clustering. The screen plot obtained using gap statistical analysis is shown in Figure 4a. The abscissa of the graph is the number of clusters, and the ordinate indicates the gap statistic value calculated using R. The gap curve became nearly flat after four clusters. The large change in the gap value occurred either between two and three or three and four clusters. Thus, the data set of the 10 crude oils was best represented with four clusters.

The results of *k*-means clustering are shown in Figures 4b and c. Figure 4b shows the clustering of peaks, and Figure 4c shows the clustering of samples. In Figure 4b, each of the four rectangular boxes contains a group of peaks with a similar distribution among 10 crude oil samples. In the first cluster, the peaks were attributed to sulfur-containing classes (OS and NS). The third cluster was mainly composed of O class compounds. It was clear that the peaks in each group had a similar distribution among 10 crude oil samples. The second cluster and fourth cluster were mainly composed of peaks in the NO and NOS classes. Thus, the clustering pattern was very similar to that obtained previously using PCA.[13]

The clustering of samples is shown in Figure 4c. In the first cluster, each of the black and red lines represents crude01 and 03 samples, respectively. The second cluster shows crude09 and 10 samples. The black, red, green, and blue lines in cluster 3 indicate crude02, 04, 06, and 07 samples, respectively. Cluster 4 represents crude05 and 08 samples. These results indicated that the *k*-means clustering algorithm was more accurately correlated with the bulk properties of 10 crude oil samples compared with PCA and HCA.

Conclusions

In this study, three clustering algorithms were successfully applied to compare 10 crude oil spectra obtained using negative-mode ESI. Oxygen and sulfur-containing peaks played important roles in clustering the negative-mode ESI spectra. As a result, the clustering of the data showed close correlations with TAN values and sulfur content of crude oils. It is interesting to note that the three different algorithms used in this study (PCA, HCA and *k*-means clustering) produced very similar results. Therefore, the choice of clustering method between the three methods does not appear to be crucial. The consistency between results obtained by different statistical methods shows credibility of applying statistical methods for interpretation of high resolution mass spectra of crude oil. However, our preliminary study showed the results obtained using another clustering technique were not consistent with those obtained using other methods. Therefore, selecting and using an appropriate statistical method can be important. Further study is being planned to explain why SOM analysis showed different result and provide guide line for the selection of appropriate statistical method.

Acknowledgments. This work was supported by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korean government (MEST) (2010-0027557). Jae won Lee was supported by Mid-career Researcher Program through NRF grant funded by the MEST (No. 20090079095).

References

1. Comisarow, M. B.; Marshall, A. G. *Chem. Phys. Lett.* **1974**, *26*, 489-490.
2. Qian, K.; Rodgers, R. P.; Hendrickson, C. L.; Emmett, M. R.; Marshall, A. G. *Energy Fuels* **2001**, *15*(2), 492-498.
3. Hughey, C. A.; Rodgers, R. P.; Marshall, A. G. *Anal. Chem.* **2002**, *36*(16), 4145-4149.
4. Hughey, C. A.; Rodgers, R. P.; Marshall, A. G.; Walters, C. C.; Qian, K.; Mankiewicz, P. *Org. Geochem.* **2004**, *35*(7), 863-880.
5. Fu, J. M.; Kim, S.; Rodgers, R. P.; Hendrickson, C. L.; Marshall, A. G.; Qian, K. N. *Energy Fuels* **2006**, *20*(2), 661-667.
6. Klein, G. C.; Kim, S.; Rodgers, R. P.; Marshall, A. G.; Yen, A. *Energy & Fuels* **2006**, *20*(5), 1973-1979.
7. Schaub, T. M.; Jennings, D. W.; Kim, S.; Rodgers, R. P.; Marshall, A. G. *Energy Fuels* **2007**, *21*(1), 185-194.
8. Smith, D. F.; Rahimi, P.; Teclemariam, A.; Rodgers, R. P.; Marshall, A. G. *Energy Fuels* **2008**, *22*(5), 3118-3125.
9. Smith, D. F.; Rodgers, R. P.; Rahimi, P.; Teclemariam, A.; Marshall, A. G. *Energy Fuels* **2009**, *23*(1), 314-319.
10. Headley, J. V.; Peru, K. M.; Barrow, M. P.; Derrick, P. J. *Anal. Chem.* **2007**, *79*(16), 6222-6229.
11. Panda, S. K.; Andersson, J. T.; Schrader, W. *Angew. Chem. Int. Ed.* **2009**, *48*(10), 1788-1791.
12. Kim, S.; Rodgers, R. P.; Marshall, A. G. *Int. J. of Mass. Spec.* **2006**, *251*(2-3), 260-265.
13. Hur, M.; Yeo, I.; Park, E.; Kim, Y. H.; Yoo, J.; Kim, E.; No, M.-h.; Koh, J.; Kim, S. *Anal. Chem.* **2009**, *82*(1), 211-218.
14. ASTM D7455 - 08 Standard practice for sample preparation of petroleum and lubricant products for elemental analysis. American Society for Testing and Materials (ASTM) International, West Conshohocken, PA.
15. ASTM D974 - 08e1 Standard test method for acid and base number by color-indicator titration. American Society for Testing and Materials (ASTM) International, West Conshohocken, PA.
16. Sohn, I.; Kim, S.; Hwang, C.; Lee, J. W. *Comput. Stat. Data Anal.* **2008**, *52*(8), 4104-4115.
17. Team, R. D. C. *R Foundation for Statistical Computing*, 2009.
18. Krzywinski, M.; Schein, J.; Birol, I.; Connors, J.; Gascoyne, R.; Horsman, D.; Jones, S. J.; Marra, M. A. *Genome Res.* **2009**, *19*, 1639-1645.
19. Tibshirani, R.; Walther, G.; Hastie, T. *J. Royal. Stat. Soc.: Series B (Statistical Methodology)* **2001**, *63*(2), 411-423.