Rapid Identification of Petroleum Products by Near-Infrared Spectroscopy

Hoeil Chung,* Hyuk-Jin Choi, and Min-Sik Ku

NIR Project Team, SK Corporation 110 Nam-Gu, Kosa-Dong, Ulsan 680-130, Korea

Received April 29, 1999

Near-infrared (NIR) spectroscopy has been successfully utilized for the rapid identification of six typical petroleum products such as light straight-run (LSR), naphtha, kerosine, light gas oil (LGO), gasoline, and diesel. The spectral features of each product were reasonably differentiated in the NIR region, and the spectral differences provided enough qualitative spectral information for discrimination. For discrimination, principal component analysis (PCA) combined with Mahalanobis distance was used to identify each petroleum product from NIR spectra. The results showed that each product was accurately identified with an accuracy over 95%. Most noticeably, LSR, kerosine, gasoline, and diesel samples were predicted with identification accuracy of 99%. The overall results ensure that a portable NIR instrument combined with a multivariate qualitative discrimination method can be efficiently utilized for rapid and simple identification of petroleum products. This is especially important when local at-site measurements are necessary, such as accidental petroleum leakage and regulation of illegal product blending.

Introduction

Petroleum products such as diesel and naphtha are important products in a humans daily life, as they have played a critical role in many industries as an energy and petrochemical source. There are several different petroleum products for diverse usage based on their physical and chemical properties.[1] Recently, the regulation related to these products has been gradually becoming more strict due to environmental and product quality issues. Environmentally, when oil leakage or other accidents happen, it is essential to identify a product as soon as possible in order to proceed with the proper following actions. On the other hand, there is a possibility of uncontrolled mixing from transportation using a long-distance pipeline, or illegal blending and selling of these products. In these cases, rapid and simple analytical instrumentation is necessary. Currently it requires several analytical instruments, including gas chromatography and ASTM D86 distillation analyzer to identify the products. Additionally, conventional analytical methods require long analysis time (more than 1 hour), high investment cost, and continuous maintenance.

Alternatively, near-infrared (NIR) spectroscopy[2,3] is an excellent analytical method for the identification of petroleum products because it is fast, rugged, and provides highly reproducible results with minimal maintenance. Additionally, a compact and portable NIR instrument is now commercially available. The rapid identification of petroleum products using NIR spectroscopy has been studied. Six different typical petroleum products of light straight-run (LSR), naphtha, kerosine, gas oil, gasoline, and diesel were examined in this study. These products are produced from the fractional distillation of crude oil or blending with other components. The physical appearance of these samples is similar, all as a clear liquid form. The corresponding physical properties are summarized in Table 1. There is another petroleum product from crude oil called atmospheric residue (AR, boiling temperature over 350 °C and main fuel for power stations, ships, heating installations), which is completely dark, viscous, and easy to distinguish, therefore it was not included in this study.

Principal component analysis (PCA)[4,5] combined with Mahalanobis distance[6,7] was used to discriminate petroleum products from their NIR spectra. PCA is the data reconstruction and reduction method using principal component (a.k.a. eigenvector, loading vector, spectral loading or factor) and score which is a scaling constant used to reconstruct the spectra. Usually the score is used for qualitative or quantitative analysis. Mahalanobis distance is useful and widely used method for the determination of the discrimination boundary. Mahalanobis distance accounts for the differences of a data cluster by the range of variability. That means it constructs a boundary space of discrimination that weights more where the larger variation in the data is present.

The spectral features of six petroleum products were reasonably different in the NIR region. The differences in physical properties and chemical compositions of these products resulted in identifiable spectral features. Those spectral fea-

Table 1. Typical properties of petroleum products used in this study

Products	Carbon Number	Boiling Range (°C)
LSR	C_5 - C_7	30- 75
Naphtha	C_6 - C_{10}	75 -190
Kerosine	C_9 - C_{15}	190 -250
LGO	C_{13} - C_{18}	250 -350
Gasoline	C_5 - C_{12}	30 -210
Diesel	C_{11} - C_{21}	200-370

*To whom correspondence should be addressed. E-mail: hoeil @skcorp. com Tel: 82-052-270-1301 Fax: 82-052-270-1309.

tures provided enough qualitative information for discrimination using multivariate techniques such as PCA. By using PCA combined with Mahalanobis distance, six different petroleum products were successfully classified and identified.

Experimental Section

Sample Preparation. Three hundred seventy two samples of LSR, naphtha, kerosine, light gas oil (LGO), gasoline, and diesel were obtained over a 4 month period at SK Corporation, Ulsan, Korea. Over this long period, samples were cautiously collected to give more compositional and process related variations into the data set. Immediately after collection, samples were sealed and stored in a refrigerator at 4 °C to prevent evaporation of the hydrocarbons.

NIR Spectra. NIR spectra were collected over the 1100 to 2500 nm spectral region with a NIRSystems model 6500 spectrometer (Foss NIRSystems, Silver Spring, MD) equipped with a tungsten halogen lamp, PbS detector, and a fiber optic interactance/reflectance probe. The resolution of collected spectra was 10 nm with 2 nm data intervals. The fiber optic probe consisted of concentric rings of illuminating fibers (inner core, 210 fibers), receiving fibers (outer ring, 210 fibers), and a reflecting mirror. The length of the optical fibers from the probe to spectrometer was 2 m. The distance between the optical fibers and the reflecting mirror was 0.5 cm, resulting with an actual pathlength of 1 cm. NIR spectra were collected by positioning the fiber optic probe into each sample that was contained in a sealed bottle. Each NIR sample spectrum consisted of 16 co-added scans.

Data Set Preparation and Discriminant Analysis. The data set was prepared as described in Table 2. The division of samples was intended to assign roughly 75 to 80 percent of the total spectra into the calibration set. Each calibration model was developed using each calibration data set. A total of 99 spectra from 6 different petroleum products were combined into the prediction data set, which served as a validation set. Spectra in the calibration and prediction set were randomly chosen.

Discriminant analysis was performed using GRAMS/32 software with an add-on PLS algorithm (Galactic Industries Corporation, Salem, NH). NIR spectra were imported into GRAMS/32 before the discriminant analysis was performed.

Table 2. Description of data set preparation

Products	Number of total spectra	Number of spectra In calibration	Number of spectra in prediction
LSR	57	43	14
Naphtha	61	43	18
Kerosine	61	45	16
LGO	64	47	17
Gasoline	65	47	18
Diesel	64	48	16
Total	372		99

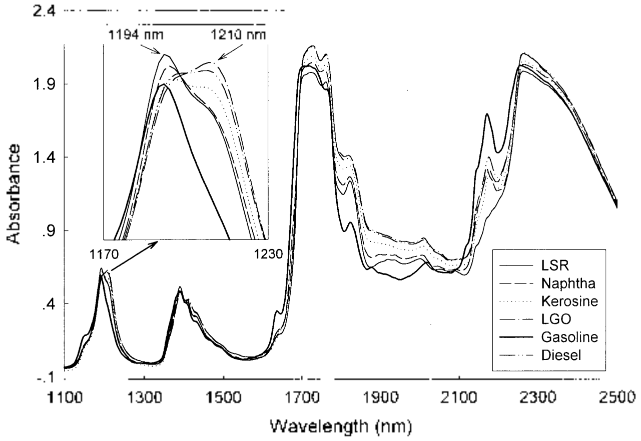

Figure 1. NIR spectra (1100-2500 nm) of each petroleum product. The spectra were averaged from all spectra in each calibration data set.

Results and Discussion

Spectral Features. For selective qualitative identification of each petroleum products it requires that at least minute spectral differences between each product should be recognized in the NIR region. NIR spectra (1100-2500 nm) of each petroleum product shown in Figure 1 are an average of all spectra in each calibration data set. Examination of the spectra reveals considerable spectral differences between each of the petroleum products. The most useful spectral information is located in the 1100 to 1650 nm and 1800 to 2100 nm spectral ranges. The spectral differences are more dominant in the 1800 to 2100 nm range compared to the 1100 to 1650 nm range. The 1650 to 1800 nm and 2100 to 2500 nm ranges contain no useful spectral information due to strong absorption of NIR radiation from the relatively long optical pathlength. Therefore, these ranges are excluded from further analysis in this study.

The most systematic spectral variations are observed 1170 to 1230 nm range. The spectral variation based on CH_2/CH_3 ratio in normal linear hydrocarbons and C_6 structural isomers has been systematically investigated in our research group.[8,9] As increasing the chain length in normal linear hydrocarbons, which corresponds to increasing the number of CH_2 groups with a fixed number of CH_3 groups, the CH_2 second overtone band at 1210 nm became more dominant while CH_3 second overtone band at 1194 nm became less dominant. On the contrary, between C6 isomers, increasing the number of CH_3 groups (more branched structure) in a given molecular weight, CH_2 second overtone band at 1210 nm became less dominant while CH_3 second overtone band at 1194 nm became more dominant. The trend exactly matches with that of the petroleum products, macroscopically. As shown in Table 1, the molecular weights and number of CH_2 groups relative to CH_3 groups are increasing in the order of LSR, naphtha, kerosine and LGO, respectively. The CH_3 band at 1194 nm is decreasing while the CH_2 band at 1210 nm is increasing from LSR to LGO (shown in Fig. 1). Gasoline and diesel are blended products. Gasoline, especially, is produced from several different petrochemical feed stocks such as reformate, which contains high concentration of aromatic hydrocarbons. Therefore, spectral features of gasoline are significantly different from other petroleum products, such as the aromatic CH band at 1140 and 1630 nm. The major components of diesel are LGO (mainly) and kerosine. Therefore, the corresponding spectral features are between LGO and kerosine, albeit closer to LGO. The magnitude of absorption in the 1800-2100 nm region is increasing from LSR to LGO due to higher molar absorptivity from higher molecular weight compounds.

Overall, unique spectral features of each petroleum products exist and can provide selective information for the qualitative discrimination of each product. It is expected that there will be spectral overlaps between products with similar physical properties, such as LGO and diesel. However, by using multivariate discrimination techniques such as Principal Component Analysis (PCA) combined with Mahalanobis distance, minute spectral differences can be discriminated.

Preliminary Qualitative Examination. Before developing qualitative calibration models for each product, PCA was performed on the combined data set (total 273 spectra) of all calibration spectra from each product to examine the discrimination feasibility. PCA is the data reconstruction method using principal component (a.k.a. eigenvector) and scores, which is a scaling constant used to reconstruct the spectra. Figure 2 shows the score cluster plot using the first and second principal component (PC). The first PC describes the greatest variation in the spectral data set and the second PC describes the second greatest variation. The top and bottom plots show the score scatter of the raw and second derivative spectra, respectively. It is well known that second derivatization of raw spectra helps to enhance the spectral features and remove (or at least reduce) baseline variation.[10] Better discrimination is achieved using second derivative spectra by removing unrelated spectral variation, such as baseline variation. The points from a given product are more closely located to each other and better segregation between products is shown with second derivative spectra. Especially in kerosine, there are some overlaps with diesel in the raw spectra, but are clearly discriminated when using the second derivative spectra. Because the composition of gasoline is quite different from other petroleum products, the points are fairly remote from other products and closely located to each other. There are some overlaps in data points between LSR and naphtha, diesel and LGO, which are due to compositional overlap as shown in Table 1. However, the data points are reasonably discriminated in the PC score space. The overall preliminary results present successful discrimination of each petroleum products when using PCA.

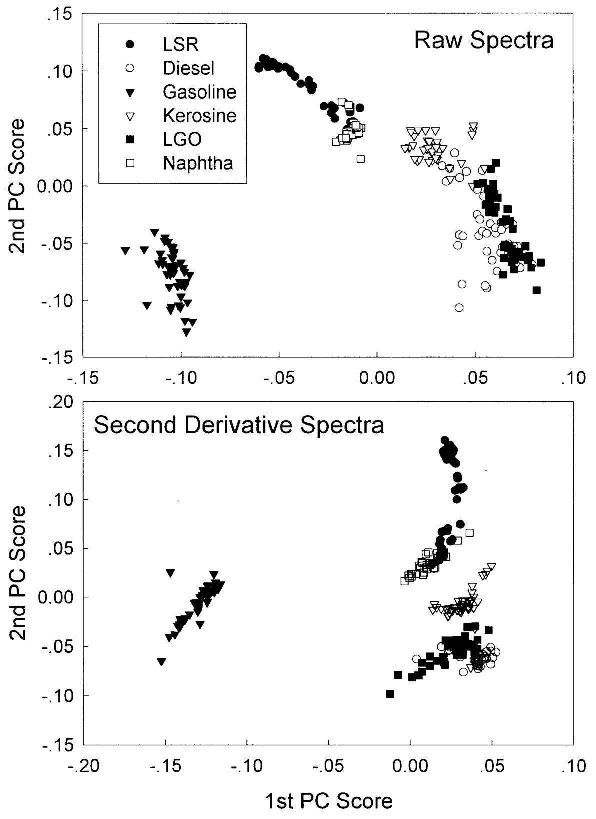

Figure 2. Score cluster plot using first and second principal component (PC). PCA was performed on the combined data set (total 273 spectra) of all calibration spectra from each product.

PCA Calibration Model Combined with Mahalanobis Distance Method. In the application of a multivariate calibration method such as PCA, it is generally known that the spectral range and number of principal components (PCs) are critical parameters. It has been previously determined that the quantitative calibration performance depends on the spectral range utilized.[10] In this study, the whole spectral range which combining 1100-1650 and 1800-2100 nm regions were examined. The spectral differences from each petroleum product were present throughout the entire NIR region, so the whole spectral range was used to incorporate more qualitative spectral information.

The number of PC was identified as the number of PCs that gave a Total Percent Variance (TPV) over 99.9%. TPV is an indicator of how much variation is accounted for by PCs. PCs represent the variation in the spectral data set, while eigenvalues are the relative weights of each individual PC. By summing eigenvalues and representing as a percentage, it can be estimated how much variance is described by the PCs. Figure 3 shows TPVs plotted as a function of the number of PCs for the discrimination of naphtha using second derivative spectra. The TPV increases sharply with the initial PCs and gradually increases as more spectral variation is incorporated into the calibration model. The majority of the spectral variation is described within the initial 4 PCs and the remainder of the variation is accounted for by the following PCs. In this case, 6 PCs were chosen to describe the spectral variation over 99.9%. The optimal number of PCs for other products was determined by generating and examining the same type of plot.

Mahalanobis distance was used to set up the discrimination boundary of the score cluster. Mahalanobis distance is an ellipsoidal boundary that circumscribes a data cluster. In

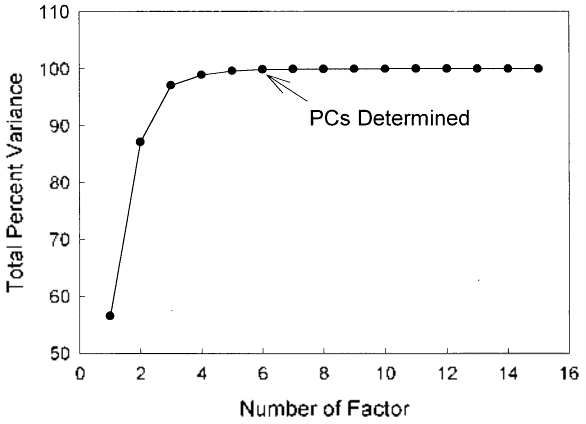

Figure 3. TPV (Total Percent Variance) plotted as a function of the number of PCs for the discrimination of naphtha using second derivative spectra.

this method, the Mahalanobis matrix (M) is calculated initially:

$$M = \frac{(X-\bar{X})'(X-\bar{X})}{n-1}$$

where X is the matrix of the location of data point, $\bar{X}$ is the matrix of the mean of cluster. To predict the Mahalanobis distance of a sample, the following equation is used:

$$D^2 - (X-\bar{X})'M^{-1}(X-\bar{X})$$

where D^2 is the square of the Mahalanobis distance of the sample from the mean of the cluster. Consequently, the calculated ellipse boundary around a cluster is the standard deviation (σ) from the mean. In general, a sample with a Mahalanobis distance of 3 (3σ) or greater (probability of 0.01 or less) can be identified as non-member of a group. In this study, both Mahalanobis distances of 2 and 3 were used as the discrimination criteria.

Table 3 shows the calibration results of PCA combined with Mahalanobis distance using the second derivative spectra. Calibration models of each product were developed using each calibration set, and then samples in the prediction set were predicted by each appropriate calibration model. The discrimination percent accuracy is shown in Table 3. Overall, the accuracy of prediction from each calibration model is excellent (over 95%). Relatively, the poorest results were observed for LGO and diesel, in comparison compared to the other products. The results from LGO and diesel are understandable because there is considerable compositional overlap and the resulting spectral features are too similar to each other, as shown in Table 1 and Figure 1. There is no difference in accuracy between Mahalanobis distance limits of 2 and 3, except for LGO and diesel. The score clusters of LSR, naphtha, kerosine and gasoline are fairly apart from each other, therefore the magnitude (standard deviation) of the circumscribing boundary does not affect the performance of discrimination. However, performance of discrimination for LGO and diesel changes with changing Mahalanobis distance limit. Since these two products are similar and the corresponding scores are closely located to each other, the narrower boundary (standard deviation of 2) gives the better discrimination results.

Table 3. Percent accuracy of discrimination of each model using PCA combined with Mahalanobis distance with second derivative spectra

	Number PCs	Mahalanobis distances under 2	Mahalanobis distances under 3
LSR	4	99.0%	99.0%
Naphtha	6	98.0%	98.0%
Kerosine	5	99.0%	99.0%
LGO	5	94.9%	93.9%
Gasoline	6	99.0%	99.0%
Diesel	9	99.0%	84.8%

Figure 4 shows the actual predicted Mahalanobis distances of samples in the prediction data set using the LGO calibration model. As expected, Mahalanobis distances of gasoline, LSR, and naphtha are amazingly high. Additionally, predicted Mahalanobis distances of kerosine appear small due to the scale of the plot, however the actual distances range from 33 to 117. The predicted Mahalanobis distances of diesel and LGO are magnified inside the plot. Only one sample from LGO is over the Mahalanobis distance of 2 and three samples from diesel are under Mahalanobis distance of 2. However, the predicted values of LGO are more closely located to each other, compared to diesel, which is ranging from 1.8 to 31.1. Overall results show that the discrimination of each product has been successfully accomplished using PCA combined with Mahalanobis distance.

Conclusion

Six typical petroleum products have been clearly discriminated using NIR spectroscopy. By combining PCA and Mahalanobis distance, the products with similar property such as LGO and diesel were efficiently identified. With the

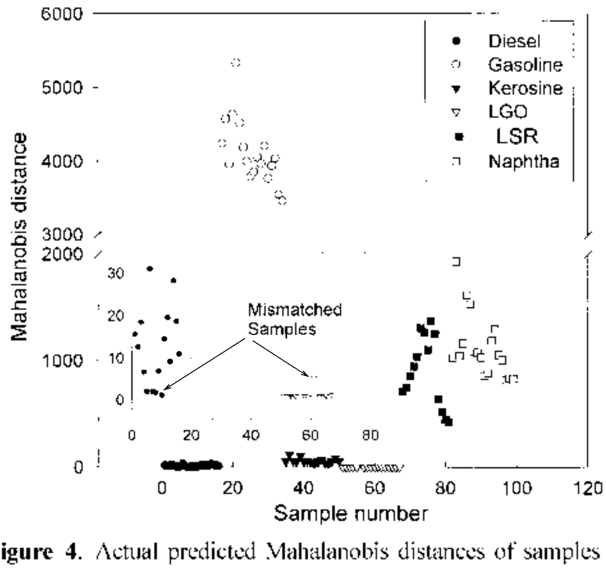

Figure 4. Actual predicted Mahalanobis distances of samples in the prediction data set using the LGO calibration model.

help of fast NIR analysis, petroleum products can be identified in less than 1 minute without high investment as with conventional analyzers. In practice, there are more petroleum products such as lube oil, solvent, and etc. The future study will incorporate more petroleum products for the qualitative discrimination and a portable NIR instrument will be utilized for the practical at-site measurements. Additionally, the same vibrational techniques of IR and Raman spectroscopy will be evaluated for the same qualitative discrimination.

References

1. Wiseman, P. *An Introduction to Industrial Organic Chemistry*; Applied Science Publishers: London, U. K., 1976; p 34.
2. Burns, D. A.; Ciurczak, E. W. *Handbook of Near-Infrared Analysis*; Marcel Dekker: New York, U. S. A., 1992.
3. Wetzel, D. L. *Anal. Chem.* **1983**, *55*, 1165A.
4. Martens, H.; Naes, T. M. *Multivariate Calibration*; John Wiley and Sons: New York, U. S. A., 1989; p 116.
5. Beebe, K. R.; Pell, R. J.; Seasholtz, M. B. *Chemometrics: A Practical Guide*; John Wiley and Sons: New York, U. S. A., 1998; p 81.
6. Mahalanobis, P. C. *Proc. Natl. Inst. Of Science of India* **1936**, *2*, 49.
7. Mark, H. L. *Anal. Chem.* **1987**, *59*, 790.
8. Ku, M. S.; Chung, H.; Lee, J. S. *Bull. Korean Chem. Soc.* **1998**, *19*, 1189.
9. Lee, J. S.; Chung, H. *Vibrational Spectrosc.* **1998**, *17*, 193.
10. Chung, H.; Lee, J. S.; Ku, M. S. *Appl. Spectrosc.* **1998**, *52*, 885.