Spectral Range Optimization for the Near-Infrared Quantitative Analysis of Petrochemical and Petroleum Products: Naphtha and Gasoline

YOUNGBOK LEE, HOEIL CHUNG,* and NAKJOONG KIM

Department of Chemistry, College of Natural Sciences, Hanyang University, Haengdang-Dong, Seoul, Korea 133-791

The proper selection of the spectral range in partial least squares (PLS) calibration is critical when highly overlapping spectra from compositionally complex samples are used, such as naphtha and gasoline. In particular, the relevant spectral information related to a given property is frequently localized in a narrow range, and the most selective region may be difficult to locate. We have presented the importance of range optimization in near-infrared (NIR) spectroscopy for the analyses of petrochemical and petroleum products that are generally highly complex in composition. For this purpose, the determination of a detailed compositional analysis (so called PIONA) and the distillation temperature of naphtha were evaluated. In the same fashion, the research octane number (RON) and Reid vapor pressure (RVP) were selected for gasoline. By optimizing the range using moving window (MW) PLS, the overall calibration performance was improved by finding the optimal spectral range for each property. In particular, for a detailed compositional analysis of naphtha, it was effective to search for localized spectral information in a relatively narrow range with fewer factors.

Index Headings: Near-infrared spectroscopy; NIR spectroscopy; Partial least squares; PLS; Moving window PLS; Range optimization; Naphtha; Gasoline.

INTRODUCTION

Near-infrared (NIR) spectroscopy has been extensively utilized in the petrochemical and petroleum industries over the past fifteen years.[1,2] The advantages of using NIR spectroscopy in these fields are based on its ability to yield fast and simultaneous multi-compositional analyses in an on-line manner that can result in a huge economic benefit when harmonized with advanced process control (APC).[3,4] Due to the compositional complexity of petroleum-driven products, their NIR spectral characteristics (overtone and combination bands) are inherently complicated, overlapping, and very difficult to interpret. For the quantitative analysis of NIR spectra with complex spectral features, partial least squares (PLS) regression[5–7] is frequently adopted as a multivariate calibration method in many real application fields.

In the petrochemical and petroleum fields, NIR analyses for naphtha[8–13] and gasoline[14–19] are the most demanding and widely investigated. Naphtha is one of the most important materials in the petrochemical industry and is a very complex mixture that contains C5 to C9 hydrocarbons.[20,21] Additionally, the structures of the individual components are fairly similar to each other. The most conventional analysis of naphtha is the determination of total PIONA[22,23] (Paraffin, Isoparaffin, Olefin, Naphthene, Aromatic) compositions as well as PIONA based on individual carbon chain length. Gasoline is also a complex and blended product composed of C4 to C9

hydrocarbons. The most important properties of gasoline that should be tightly controlled for quality assurance are the research octane number (RON) and the Reid vapor pressure (RVP).

Especially for the analysis of naphtha, where a detailed compositional analysis is required using highly overlapping NIR spectra, performing a simple PLS calibration using the entire spectral range without any range optimization is the common practice in real NIR application fields. Use of the entire spectral range for PLS could be considered advantageous since all the information recorded can be utilized. However, this may not always be the case, considering instances when the spectral features relating to a given component are localized in a narrow spectral range and may even be overlapped with those from other components. It is possible that an optimal spectral range that yields improved PLS calibration may result. This possibility could be equally applied to the analysis of naphtha and gasoline, since the resulting spectral features are often broad, without any distinct spectral features. Several chemometric strategies, such as moving window PLS (MW-PLS)[24,25] and interval PLS (iPLS),[26] used to find the best calibration range have been reported. For example, the prediction performance of glucose in serum has been improved by an optimization of the spectral range.[27]

The aim of this publication is to demonstrate the importance of range optimization in PLS for the analyses of petroleum-driven products by using naphtha and gasoline as representative products. We used a similar methodology of MW-PLS for range optimization.

EXPERIMENTAL

Sample Preparation. Two hundred seventeen (217) naphtha samples were obtained over a seven-month period from a petrochemical company in Korea. Samples were collected cautiously so as to introduce more compositional variation into the datasets. The whole dataset was divided into calibration and validation sets as described in Table I. From a practical viewpoint, the most typical analyses of naphtha are group compositional analyses: total paraffin, total normal paraffin, total isoparaffin, total naphthene, and total aromatic content.[28] There were only trace olefin concentrations in all samples, so this was excluded. For detailed compositional analyses, C6 paraffin (*n*-hexane), C6 isoparaffin (2-methylpentane, 3-methylpentane, 2,2-dimethylbutane, 2,3-dimethylbutane), C6 naphthene (cyclohexane, methyl cyclopentane), C6 aromatic (benzene), and C7 aromatic (toluene) were selected. Additionally, the distillation temperature[29] at 10%, 50%, and 90% (D 10%, 50%, and 90%) was selected as a physical property. Table II shows the property variation ranges of each selected item for naphtha, and Fig. 1 provides the sample

distribution plots for the corresponding items. The concentration of each sample was measured by a Hewlett-Packard 5890 GC (Reformulyzer[TM]) equipped with three separation columns and a flame ionization detector (FID). For analysis, 0.2 mL of naphtha sample was injected using a syringe and helium was used as a carrier gas. Distillation temperature was determined using the ASTM D86 method.

For the gasoline dataset, 225 samples were collected over a period of 12 months for the calibration set. The main purpose of collecting the samples over a year was to include any compositional variations based on the four seasonal grade changes. Property variations of the research octane number (RON) and Reid vapor pressure (RVP) of the final gasoline product are usually small; therefore, we also acquired samples that were artificially varied by changing the blending ratios to values that were wider than the normal upper and lower limits of each component in order to increase the range of the blends and the chemical variations. By using this approach, we acquired 60 extra samples, which were added into the calibration set (total 285 samples). For the validation set, an extra 68 gasoline samples were collected over an additional 8 months. The RON and RVP are the most important properties for both quality control and assurance. The RON and RVP of all samples were determined using a conventional knock engine and an RVP analyzer, respectively. Overall descriptions of the dataset and corresponding property variations are also summarized in Tables I and II, respectively. Figure 1 shows the sample distribution plots of gasoline for RON and RVP.

Spectra Collection and Data Processing. All of the NIR spectra for naphtha and gasoline were collected using an ABB Bomem MB-160 bench-top FT-NIR spectrometer (Quebec, Quebec City, Canada) equipped with a tungsten–halogen source and a DTGS detector. A 0.5 mm path length flow cell (transmission) incorporating CaF_2 windows was used to collect the spectra. Air was used as a background for all of the sample

spectra. Each NIR spectrum was obtained by accumulating 16 scans at a resolution of 4 cm^{-1}. NIR spectra were collected over a range of 4800–4000 cm^{-1} at 25 °C.

The processing of all spectra, including the moving window partial least squares (MW-PLS) regression, were accomplished using Matlab Version 7.0 (MathWorks Inc., Natick, MA).

RESULTS AND DISCUSSION

Near-Infrared Spectral Features. Figure 2 shows the NIR spectra of naphtha and gasoline (40 randomly selected samples for each case) in the 4800–4000 cm^{-1} range. The overall spectral features were similar to each other. The aromatic bands around 4600 cm^{-1} were slightly more distinct for gasoline due to a higher concentration of aromatic compounds. These bands arise from pure aromatic or olefin compounds, particularly from a combination of =C–H stretching at 3100–3000 cm^{-1} and C=C ring stretching at 1600–1450 cm^{-1} (e.g., 3050 cm^{-1} + 1550 cm^{-1} = 4600 cm^{-1}). The spectral variations were relatively small, even over the most sensitive range in the NIR region (closest to the mid-infrared range).

Figure 3 shows the NIR spectra (4800–4000 cm^{-1}) of *n*-hexane, 2,2-dimethylbutane, cyclohexane, benzene, and toluene. All of the spectra were horizontally offset for a clear comparison. In each plot, the spectra with a thick line correspond to the average spectrum of all naphtha samples in the calibration set. The spectral features of *n*-hexane and 2,2-dimethylbutane were most similar to those of the average naphtha spectrum in the 4500–4000 cm^{-1} range. For cyclohexane, its spectral features were considerably different from those of the average spectrum since there were only methylene (–CH_2–) vibrations in the cyclic ring structure. Distinct peaks were observed around 4270, 4180, and 4100 cm^{-1}. Benzene shows the most distinct spectral bands. A strong band and several smaller bands were observed at 4050 and over the 4700–4500 cm^{-1} range, respectively. Another benzene spectrum is presented inside the bottom plot to show the full spectral features. Toluene also shows distinct spectral features. The unique peak of the toluene ring can be observed in the 4700–4500 cm^{-1} range, as was the case for benzene. Several additional absorption bands were observed in the 4500–4000 cm^{-1} range.

Through a comparison with the average naphtha spectrum, the aromatic hydrocarbons were found to exhibit the most unique spectral features, especially benzene, followed by

TABLE II. The property variation ranges for each selected item of naphtha and gasoline.

| Sample	Item	Maximum	Minimum	Average	Standard deviation
Naphtha	t-Paraffin (wt %)	91.80	68.86	78.28	3.80
	t-*i*-Paraffin (wt %)	50.78	32.50	38.81	2.25
	t-*n*-Paraffin (wt %)	52.60	32.41	39.47	2.97
	t-Naphthene (wt %)	23.19	6.23	15.48	2.80
	t-Aromatic (wt %)	9.24	0.14	5.86	1.59
	C6-Paraffin (wt %)	18.94	8.27	10.20	1.30
	C6 *i*-Paraffin (wt %)	18.74	7.41	11.70	1.96
	C6 Naphthene (wt %)	7.37	2.74	4.91	1.05
	C6-Aromatic (wt %)	2.48	0.08	1.39	0.36
	C7-Aromatic (wt %)	4.89	0.06	2.22	0.91
	D 10% (°C)	65.60	37.20	47.72	5.09
	D 50% (°C)	100.50	50.20	72.90	11.01
	D 90% (°C)	142.40	76.40	121.38	11.29
Gasoline	RON	94.00	86.60	90.09	0.62
RVP (kPa)	86.50	60.50	77.37	5.78	
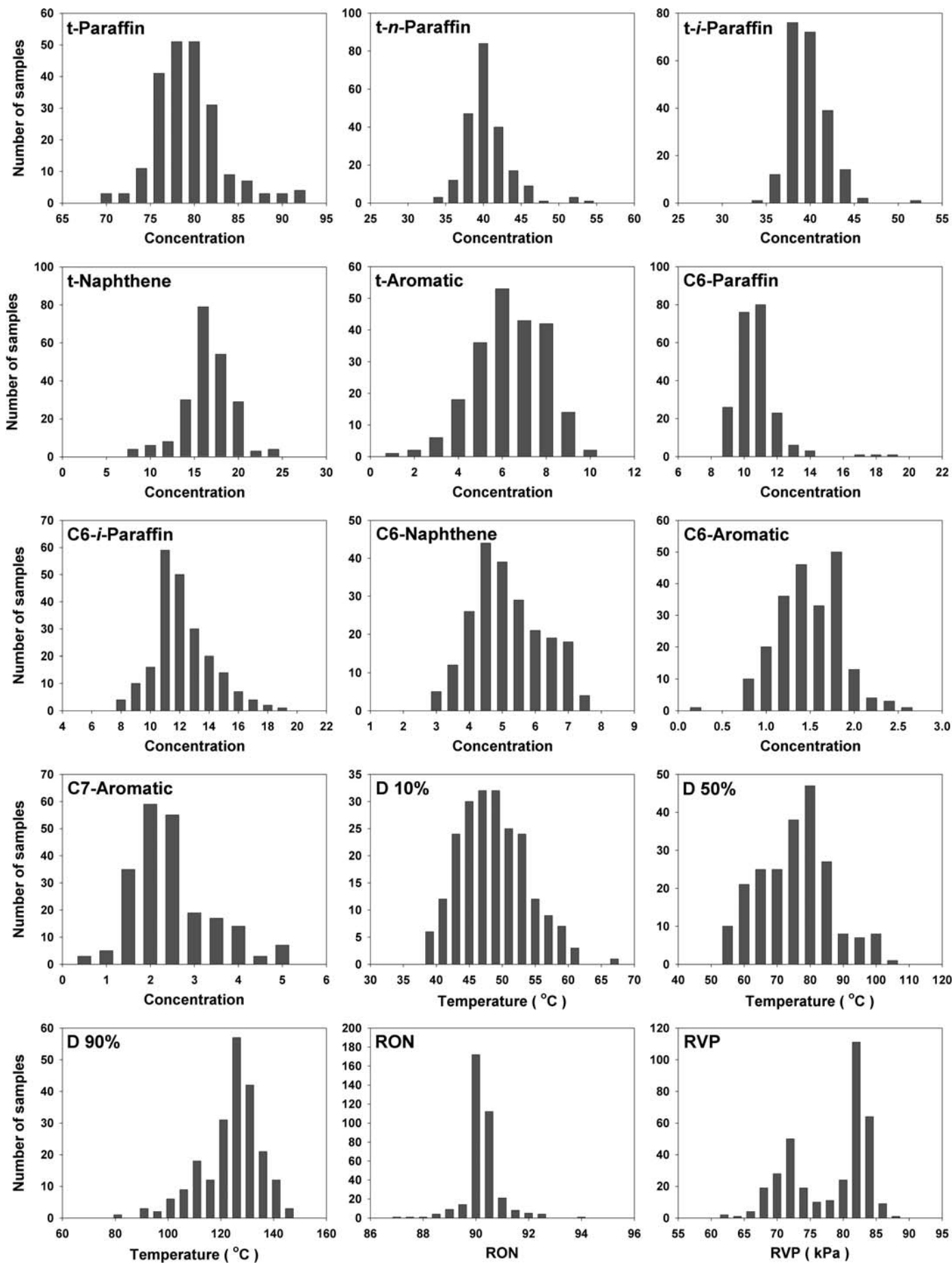
FIG. 1. Sample distribution plots for all calibration items of naphtha and gasoline.
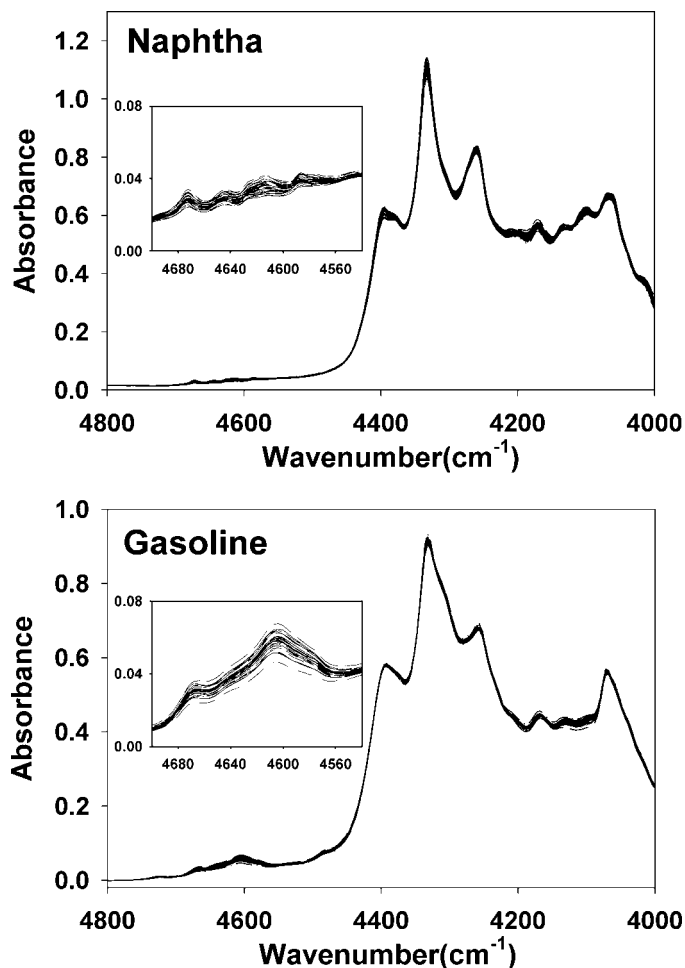
FIG. 2. NIR spectra of naphtha and gasoline (40 randomly selected samples for each case) in the 4800–4000 cm^{-1} range.

naphthene, isoparaffin, and then normal paraffin. As seen in Fig. 3, the spectral features related to the five detailed compositions were largely different from each other; therefore, the spectral range that provided the best PLS performance would be different and required optimization for each case.

Partial Least Squares Using an Entire Spectral Range versus an Optimized One (Naphtha). The main concept of moving window PLS (MW-PLS) is to perform PLS at all possible combinations of spectral ranges and to determine the spectral range that yields minimum error. For this purpose, the window size (number of spectral points) was varied from 10 to 800 cm^{-1} in 10 cm^{-1} increments. In a given window size, the window was moved from low to high wavenumber with a 10 cm^{-1} step to cover the entire spectral range (4800–4000 cm^{-1}). A total of 3081 spectral ranges with various window sizes as well as different spectral ranges were tested and the best range was identified.

To perform MW-PLS, each original calibration set was further divided into a new calibration set (143 spectra for naphtha and 230 spectra for gasoline) and validation set (34 spectra for naphtha and 55 spectra for gasoline). At each spectral range, PLS was performed using the new calibration set, and the standard error of prediction (SEP) resulting from the new validation set was evaluated to determine the optimal spectral range (corresponding to the lowest SEP). Then, PLS was performed using the original calibration set within the
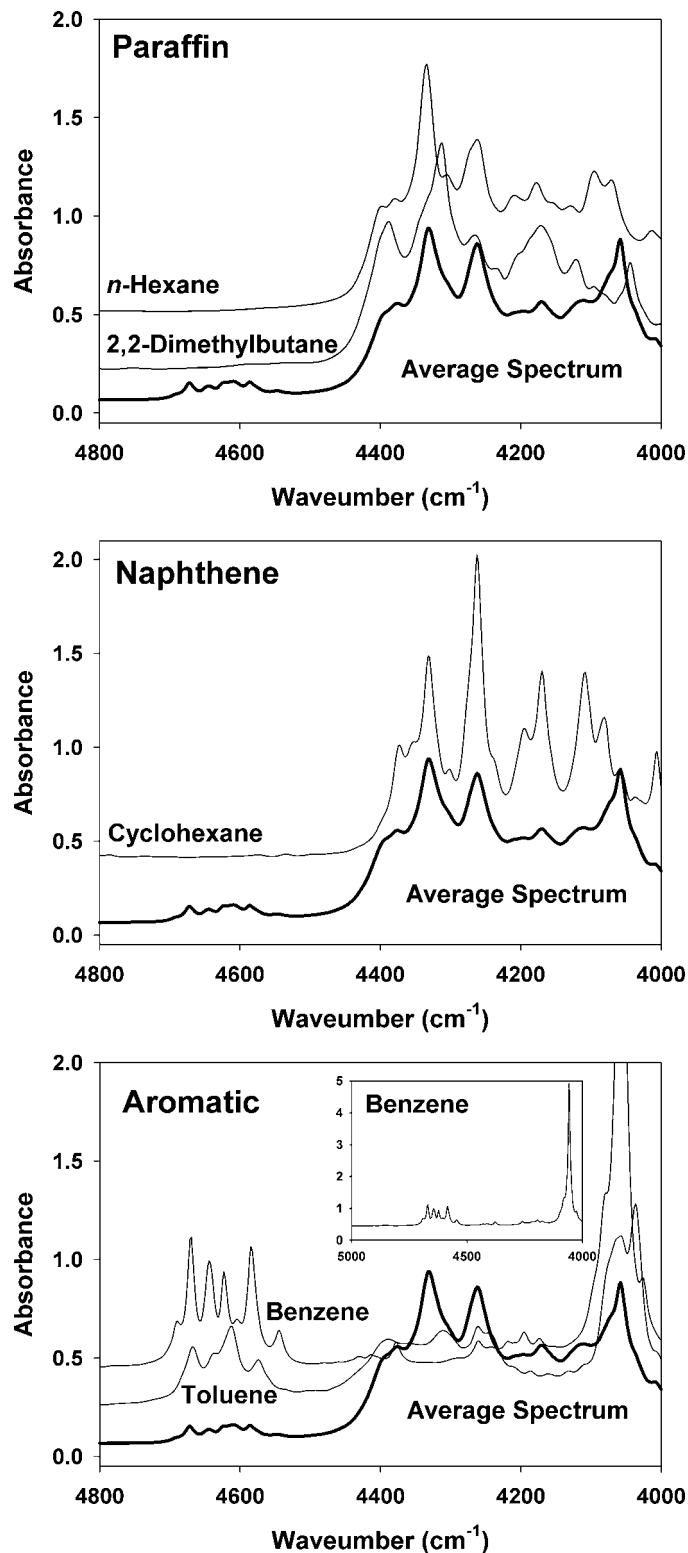

FIG. 3. NIR spectra (4800–4000 cm^{-1}) of *n*-hexane, 2,2-dimethylbutane, cyclohexane, benzene, and toluene. All of the spectra were horizontally offset for a clear comparison. In each plot, the spectra with a thick line correspond to the average spectrum of all naphtha samples in the calibration set.

optimal spectral range and the final SEP was calculated by predicting samples in the original validation set.

The optimum number of factors was identified by examining the pattern of decrease of the standard error of calibration (SEC) as a function of the number of PLS factors.[30] Figure 4
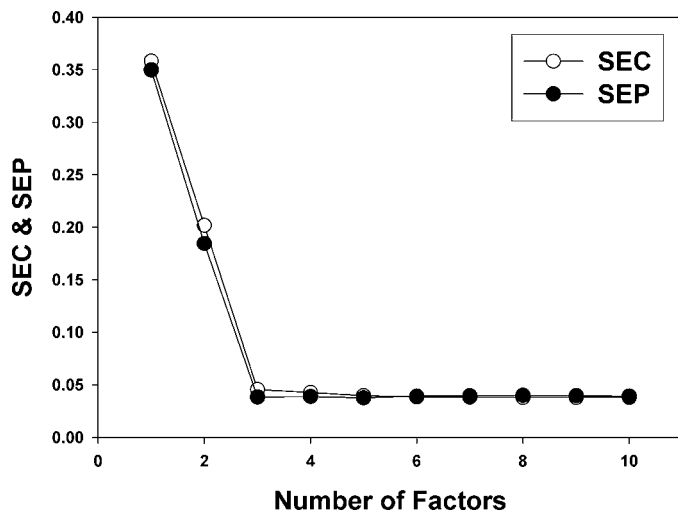
FIG. 4. SEC (open circles) and SEP (filled circles) plotted as functions of the number of PLS factors used for the determination of C6 aromatic using the optimized spectral range (4670–4620 cm^{-1}).

shows SEC (open circles) and SEP (filled circles) plotted as functions of the number of PLS factors used for the determination of C6 aromatic using the optimized spectral range (4670–4620 cm^{-1}). As expected, both the SEC and SEP decrease sharply with the initial factors as more analyte-dependent spectral variation is incorporated into the calibration model. As the number of factors increases further (i.e., the 3rd factor in the figure), the SEC continues to drop slightly while the SEP begins to increase. This type of response is common for factor-based multivariate calibrations. In this case, the optimum number of factors is three. This same procedure was used to determine the optimum number of factors for all other calibration items. The PLS calibration results using the entire spectral range and each optimized range using MW-PLS are summarized in Table III.

For group compositions, except for total aromatics, the best spectral ranges were determined to be within the range of 4550–4040 cm^{-1}, where spectral information from both aliphatic and aromatic hydrocarbons was present. In the case of total paraffin, the resulting SEP was improved using the

4400–4250 cm^{-1} range that incorporates three strong CH bands (as shown in Fig. 3). Additionally, the required number of factors was decreased by determining the range with fewer degrees of variation (or freedom). For the case of the total normal paraffin, total isoparaffin, and total naphthene, the improvements in the SEP through range optimization were marginal. This indirectly showed that the relevant information was spread over the entire spectral range rather than localized in a specific region; therefore, the required number of factors was almost the same, even when fairly narrow spectral ranges were used.

The best spectral range for total aromatic composition was determined to be over the 4700–4540 cm^{-1} range, where only the aromatic spectral features were present as shown in Fig. 3, and the resulting prediction performance was significantly improved based on a paired t-test ($p_{\text{calculated}} = 0.004 < 0.05 = p_{\text{95\% confidence interval}}$). Even though the absorbance of the aromatic bands in this range was considerably low, the high selectivity in the absence of overlapping from the aliphatic bands led to an improved calibration performance. The required number of factors was decreased by 2 due to the use of simpler and interference-free spectral variations related to the total aromatic content.

For all detailed compositional analyses for naphtha, each optimized spectral range yielded an improved prediction performance using fewer numbers of factors. This is clear evidence that the localized spectral range for a pure component could be determined even within broad spectral features from a highly complex matrix. For the C6 paraffin, C6 isoparaffin, and C6 napthene cases, narrow ranges were selected within the 4400–4100 cm^{-1} range. The average window size of the optimal ranges for the group compositions, except for total aromatic composition, was 315 cm^{-1}, while 173 cm^{-1} was used for C6 paraffin, isoparaffin, and naphthene. As described, the necessary information for a detailed compositional analysis was relatively localized.

A similar trend was observed for the C6 aromatic (benzene) and C7 aromatic (toluene), as fewer factors were used in the optimized spectral ranges (4670–4620 cm^{-1} range for C6 aromatic and 4640–4600 cm^{-1} range for C7 aromatic). Similar to the case of total aromatics, the use of only aromatic bands (without any influence from aliphatic features) helped to

TABLE III. The overall PLS calibration results for both naphtha and gasoline with entire and optimal spectral ranges.

| | Item | Entire range (4800–4000 cm^{-1}) | | | | Optimal range | | | | |
		SEC	Number of PLS factors	SEP	Window size (cm^{-1})	Range (cm^{-1})	SEC	Number of PLS factors	SEP	Window size (cm^{-1})
Naptha	t-Paraffin (wt%)	0.71	6	0.85	800	4400–4250	0.64	5	0.59	150
	t-i-Paraffin (wt%)	0.62	9	0.65	800	4550–4080	0.59	9	0.62	470
	t-n-Paraffin (wt%)	0.26	7	0.25	800	4430–4040	0.28	6	0.26	390
	t-Naphthene (wt%)	0.65	6	0.68	800	4540–4290	0.70	6	0.63	250
	t-Aromatic (wt%)	0.16	6	0.19	800	4700–4540	0.11	4	0.09	160
	C6-Paraffin (wt%)	0.25	9	0.26	800	4240–4100	0.17	7	0.18	140
	C6 i-Paraffin (wt%)	0.40	8	0.49	800	4310–4170	0.29	7	0.29	140
	C6 Naphthene (wt%)	0.20	6	0.16	800	4400–4160	0.15	6	0.12	240
	C6-Aromatic (wt%)	0.06	6	0.06	800	4670–4620	0.04	3	0.03	50
	C7-Aromatic (wt%)	0.19	6	0.16	800	4640–4600	0.06	3	0.06	40
	D 10% (°C)	1.18	8	1.13	800	4150–4020	1.27	8	0.99	130
	D 50% (°C)	1.40	7	0.95	800	4460–4020	1.41	7	0.91	440
	D 90% (°C)	4.40	8	5.10	800	4220–4100	4.41	8	4.47	120
Gasoline	RON	0.20	5	0.26	800	4410–4010	0.19	5	0.23	400
	RVP (kPa)	1.21	7	1.59	800	4290–4150	1.14	7	1.33	140

describe the variation of a fewer number of factors. The window size of the optimal range for the total aromatic composition was 160 cm^{-1}, while 50 cm^{-1} and 40 cm^{-1} were best for the C6 and C7 aromatics, respectively. A very narrow spectral range was selected as an optimum for the determination of the C6 and C7 aromatic content.

Quite different results were achieved for the determination of the distillation temperatures (D 10%, 50%, and 90%). Since this is a physical property that results from contributions of all components rather than just a specific component in the naphtha, it was expected that the necessary spectral information would be present over the entire spectral range. For the D 10, 50, and 90% cases, a narrow range was selected within the 4460–4020 cm^{-1} range. The aromatic spectral range (4700–4540 cm^{-1}) could provide relevant information; however, it was excluded due to its low absorbance. The resulting SEPs were slightly improved using each optimized spectral range, while the number of factors used was the same for both cases (both the entire and optimal spectral range). This result showed that while the degree of variation was similar over the entire spectral range, an optimal range existed that could provide a better correlation with the desired distillation property.

Partial Least Squares Using the Entire Spectral Range versus an Optimized One (Gasoline). Similar results were achieved for gasoline as in the distillation properties of naphtha, since the RON and RVP are physical properties governed by all of the components in gasoline. For the RON and RVP cases, the optimal range was selected to be within 4410–4010 cm^{-1}, where both aliphatic and aromatic spectral information was available. Even though the absorbance of the aromatic spectral range (4700–4540 cm^{-1}) in the gasoline spectra was slightly higher compared to that of the naphtha spectra, it was still excluded. The resulting SEPs for the RON and RVP were improved (by 0.03 and 0.26 kPa, respectively) by a range optimization that afforded better correlation, while the number of factors used was the same for both the entire and optimal spectral ranges. This result also showed that the degree of variation was similar over the entire range. While a 0.03 improvement in the RON prediction could be considered small from the result of the paired t-test ($p_{calculated} = 0.051$ that is similar to $0.05 = p_{95\% \text{ confidence interval}}$), just a slight improvement in RON accuracy can yield substantial economic savings when combined with a process optimization strategy.

CONCLUSION

The overall PLS prediction performances of naphtha and gasoline have been improved through range optimization. For the detailed compositional analysis of naphtha, range optimization especially helped to determine localized information with a smaller number of factors. For the determination of physical properties such as distillation temperature, RON, and RVP, optimized narrow ranges were found; however, the number of PLS factors used in the entire spectral range and the optimized range remained the same. Even though the overall study was accomplished using two important petroleum-driven products, the reasoning and logic obtained from this research could be similarly applied to the analysis of other products such as diesel, kerosene, lube base oil, reformate, and so on.

1. D. A. Burns and E. W. Ciurczak, *Handbook of Near-Infrared Analysis* (Marcel Dekker, New York, 1992).
2. D. L. Wetzel, Anal. Chem. **55**, 1165 (1983).
3. H. M. S. Lababidi, S. Kotob, and B. Yousuf, Comput. Chem. Eng. **26**, 1303 (2002).
4. J. R. Whiteley, J. Hazard. Mater. **130**, 42 (2006).
5. H. Martens and T. M. Naes, *Multivariate Calibration* (John Wiley and Sons, New York, 1989).
6. K. R. Beebe, R. J. Pell, and M. B. Seasholtz, *Chemometrics: A Pratical Guide* (John Wiley and Sons, New York, 1998).
7. E. Stark, K. Luchter, and M. Margoshes, Appl. Spectrosc. Rev. **22**, 335 (1986).
8. H. Chung, S. Choi, J. Choo, and Y. Lee, Bull. Korean Chem. Soc. **25**, 647 (2004).
9. H. Chung, J. S. Lee, and M. S. Ku, Am. Lab. **31**, 24 (1999).
10. M. S. Ku and H. Chung, Appl. Spectrosc. **53**, 557 (1999).
11. M. S. Ku, H. Chung, and J. S. Lee, Bull. Korean Chem. Soc. **19**, 1189 (1998).
12. G. R. Flaten and A. D. Walmsley, Chemom. Intell. Lab. Syst. **73**, 55 (2004).
13. S. Macho and M. S. Larrechi, Trac-Trends Anal. Chem. **21**, 799 (2002).
14. H. Chung, H. Lee, and C. H. Jun, Bull. Korean Chem. Soc. **22**, 37 (2001).
15. G. Bohacs, Z. Ovadi, and A. Salgo, J. Near-Infrared Spectrosc. **6**, 341 (1998).
16. N. M. Faber, D. L. Duewer, S. J. Choquette, T. L. Green, and S. N. Chesler, Anal. Chem. **70**, 2972 (1998).
17. G. Buttner, Proc. Control Qual. **9**, 197 (1997).
18. I. Litani-Barzilai, I. Sela, V. Bulatov, I. Zilberman, and I. Schechter, Anal. Chim. Acta **339**, 193 (1997).
19. J. J. Kelly, C. H. Barlow, T. M. Jinguji, and J. B. Callis, Anal. Chem. **61**, 313 (1989).
20. R. N. Watkins, *Petroleum Refinery Distillation* (Gulf Publishing Company, Houston, 1973).
21. P. Wiseman, *An Introduction to Industrial Organic Chemistry* (Applied Science Publishers Ltd, London, 1976).
22. C. Vendeuvre, F. Bertoncini, D. Espinat, D. Thiebaut, and M. Hennion, J. Chromatogr., A **1090**, 116 (2005).
23. D. Sanpui, M. K. Singh, and A. Khanna, Comput. Chem. Eng. **28**, 1529 (2004).
24. J. H. Jiang, R. J. Berry, H. W. Siesler, and Y. Ozaki, Anal. Chem. **74**, 3555 (2002).
25. S. Sasic, Y. Katsumoto, H. Sato, and Y. Ozaki, Anal. Chem. **75**, 4010 (2003).
26. R. Leardi and L. Norgaard, J. Chemom. **18**, 486 (2005).
27. A. K. Amerov, J. Chen, G. W. Small, and M. A. Arnold, Anal. Chem. **77**, 4587 (2005).
28. N. Kosal, A. Bhairi, and M. Ashraf Ali, Fuel **69**, 1012 (1990).
29. C. Pasquini and S. H. F. Scafi, Anal. Chem. **75**, 2270 (2003).
30. H. Chung, J. S. Lee, and M. S. Ku, Appl. Spectrosc. **52**, 885 (1998).