Investigation of Partial Least Squares (PLS) Calibration Performance based on Different Resolutions of Near Infrared Spectra

Hoeil Chung,* Seung-Yeol Choi,† Jaebum Choo,† and Youngil Lee‡

Department of Chemistry, College of Natural Sciences, Hanyang University, Seoul 133-791, Korea
†Department of Applied Chemistry, College of Science and Technology, Hanyang University, Ansan 425-791, Korea
‡Dongbu Advanced Research Institute, Chemical Analysis Team, 103-2 Munji-dong, Yusung-gu, Daejeon 305-708, Korea
Received January 12, 2004

Partial Least Squares (PLS) calibration performance has been systematically investigated by changing spectral resolutions of near-infrared (NIR) spectra. For this purpose, synthetic samples simulating naphtha were prepared to examine the calibration performance in complex chemical matrix. These samples were composed of C_6-C_9 normal paraffin, iso-paraffin, naphthene, and aromatic hydrocarbons. NIR spectra with four different resolutions of 4, 8, 16, and 32 cm^{-1} were collected and then PLS regression was performed. For PLS calibration, five different group compositions (such as total paraffin content) and six different pure components (such as benzene concentration) were selected. The overall results showed that at least 8 cm^{-1} resolution was required to resolve the complex chemical matrix such as naphtha. It was found that the influence of resolution on the PLS calibration was varied by the spectral features of a component.

Key Words : Near infrared spectroscopy, Partial least squares (PLS), Spectral resolution, Chemometrics

Introduction

NIR spectroscopy[1,2] is one of the most popular analytical methods, especially for many practical and industrial applications since it is fast, non-destructive without requiring chemical reagent, and easy to extent to on-line analysis in conjunction with optical fibers. It has been applied to widely different industrial areas such as agricultural, food, pharmaceutical, and petrochemical fields. Most frequently, PLS (partial least squares) regression[3-5] has been utilized for the calibration.

For actual NIR analyses, several types of NIR instruments are commercially available such as FT (Fourier Transform), dispersive, and AOTF (Acoustooptic Tunable Filter) that produce different spectral resolutions. Spectral resolution is one of the important parameters that should be considered for the optimal spectroscopic measurement. Since spectral bands in NIR region are broad due to the overtone and combination characteristics, it may not require a high spectral resolution for the quantitative calibration. However, there will be the possible influence of resolution on the calibration performance. Some NIR end-users decide their instrument types without considering the resolution in relation with the final analytical performance.

In this publication, the NIR analysis of petrochemical products is considered. For this purpose, the simulated naphtha samples were investigated. Naphtha is the complex mixture of C_6 to C_9 hydrocarbons.[6,7] The samples were composed of 25 pure hydrocarbons from C_6 to C_9 paraffin, iso-paraffin, naphthene (cycloalkane) and aromatic hydrocarbons. NIR spectra with four different resolutions of 4, 8, 16 and 32 cm^{-1} were measured. PLS models were built for the determination of five different group compositions and six different pure components. This was to evaluate the effect of resolution on general group compositional analysis as well as pure components with different spectral features. The overall results showed that at least 8 cm^{-1} resolution was required to resolve the complex chemical matrix such as naphtha.

*Author to whom correspondence should be sent. Tel: +82-2-2290-0937; Fax: +82-2-2299-0762; e-mail: hoeil@hanyang.ac.kr

Experimental Section

Reagents and data set preparation. Twenty five pure hydrocarbons used in this study were purchased from Sigma-Aldrich. The list of 25 components is shown in Table 1. The mixtures of different concentrations were prepared by mixing the appropriate amounts of each component, according to the random experimental design. This was to have random concentration distributions in the data set. The resulting concentration ranges, maxima, minima, and corresponding standard deviations are summarized in Table 1.

A total of 55 samples were prepared. All the samples were handled with a great care during sample preparation and spectral acquisition to minimize possible analytical errors due to volatility of components. Additionally, whenever a sample was prepared, the corresponding NIR spectrum was immediately collected.

NIR spectra and data processing. All the NIR spectra were collected using an ABB Bomem FT-NIR spectrometer (Quebec City, Quebec, Canada) equipped with a tungsten-halogen source and DTGS detector. Transmission cell (path-length: 0.5 mm) made of CaF_2 was used to collect spectra over 5000-4000 cm^{-1} range. Air was used as a background for the whole samples.

To minimize any possible influence of instrumental vari-

Table 1. Description of sample compositions

Carbon Number	Name	Type	Max. (wt. %)	Min. (wt. %)	Standard Deviation
6	*n*-Hexane	*n*-paraffin	12.69	3.66	2.14
	3-Methylpentane	*i*-paraffin	4.25	0.13	1.21
	2,2-Dimethylbutane	*i*-paraffin	4.17	0.18	1.05
	2,3-Dimethylbutane	*i*-paraffin	4.34	0.15	1.05
	Cyclohexane	naphthene	11.49	3.65	1.91
	Methylcyclopentane	naphthene	6.66	1.15	1.48
	Benzene	aromatic	11.69	3.99	1.95
7	*n*-Heptane	*n*-paraffin	11.85	3.77	2.32
	2,4-Dimethylpentane	*i*-paraffin	3.85	0.15	1.08
	2,2,3-Trymethylbutane	*i*-paraffin	3.72	0.20	1.05
	Cycloheptane	naphthene	12.21	3.70	2.13
	Methylcyclohexane	naphthene	3.74	0.16	1.10
	Toluene	aromatic	10.79	3.89	1.95
8	*n*-Octane	*n*-paraffin	12.03	3.77	2.16
	2,2,4-Trimethylpentane	*i*-paraffin	4.02	0.20	1.13
	1,2-Dimethylcyclohexane	naphthene	3.89	0.16	1.04
	1,3-Dimethylcyclohexane	naphthene	4.04	0.13	1.11
	Xylene	aromatic	12.20	3.48	2.22
9	*n*-Nonane	*n*-paraffin	12.25	3.71	2.33
	Indan	*i*-paraffin	3.77	0.16	1.05
	1,2,4-trimethylcyclohexane	naphthene	3.96	0.26	1.11
	Propylbenzene	aromatic	4.08	0.34	0.93
	1,2,3-trimethylbenzene	aromatic	3.83	0.19	1.06
	1,2,4-trimethylbenzene	aromatic	3.76	0.13	1.07
	Isopropylbenzene	aromatic	3.96	0.16	1.16

ations among data sets of four different resolutions (4, 8, 16, and 32 cm^{-1}), NIR spectra were continuously collected for a given sample by varying spectral resolutions without changing the sample. By this way, there were no significant instrumental variations among four spectra except spectral resolutions since the duration of collecting four spectra (one sample) was approximately 15 minutes.

PLS regression was accomplished using GRAMS/32 software with add-on PLS algorithm (Galactic Industries Corporation, Salem, NH, USA).

Results and Discussion

Spectral features of components. To rationalize the variation of PLS calibration performance by different spectral resolutions in a complex chemical mixture, it is initially necessary to investigate the spectral features of components as well as mixtures. The 5000 to 4000 cm^{-1} (combination bands) range is used for this study, since this range provides the most selective spectral information in the NIR region; therefore, the influence of resolution can be clearly and effectively investigated using this spectral range. As presented in the experimental section, the 0.5 mm optical pathlength was chosen to maximize spectral features in the 5000 to 4000 cm^{-1} range and maintain the maximum absorbance around 1.0.

Figure 1 shows NIR spectra (5000-4000 cm^{-1}) of selected normal paraffin (a) and iso-paraffin hydrocarbons (b). In addition, Figure 2 shows NIR spectra of selected naphthene (a) and aromatic hydrocarbons (b). All the spectra were horizontally offset for the clear comparison. In both figures, the spectrum with a thick line corresponds to the average spectrum of the whole data set. The degree of spectral (structural) difference of each component is an important parameter to determine quantitative calibration performance, since it is expected that the variation of spectral resolution would be more influential for the calibration of a component that has similar spectral features over those of whole data set.

In Figure 1(a), NIR spectra of *n*-hexane, *n*-heptane and *n*-octane are shown. As expected, the spectral features of these three components are similar each other because their molecular structures are similar except the number of methylene (-CH_2-) groups. In the average spectrum, the unique peaks of benzene ring are observed in the 4700-4500 cm^{-1} range; whereas, these bands are not observed in the spectra of *n*-hexane, *n*-heptane and *n*-octane. By comparing with the average spectrum, the spectral features of normal paraffins are similar with those of the average spectrum in the 4500-4000 cm^{-1} range. NIR spectra of 2,2-dimethylbutane and 2,4-dimethylpentane are shown in Figure 1(b). The spectral features of these two iso-paraffins are fairly different from those of the average spectrum. The branched molecular structure results in the distinguishable spectral features.

Figure 2(a) shows NIR spectra of cyclohexane and 1,2-dimethylcyclohexane. For cyclohexane, its spectral features

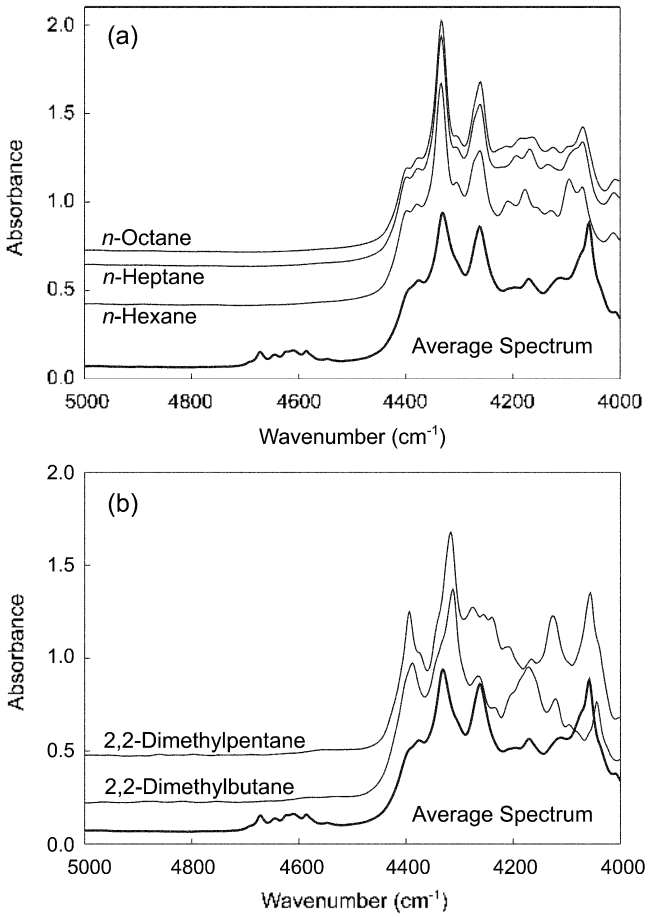

Figure 1. NIR spectra (5000-4000 cm^{-1}) of selected normal paraffin (a) and iso-paraffin hydrocarbons (b). The spectrum with thick line corresponds to the average spectrum of the whole data set.

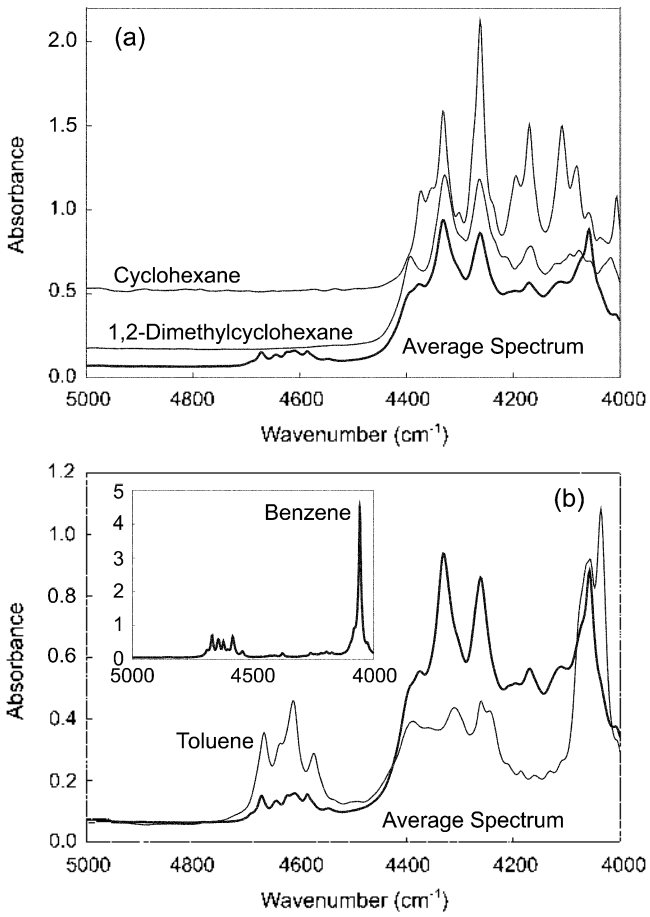

Figure 2. NIR spectra (5000-4000 cm^{-1}) of selected naphthene (a) and aromatic hydrocarbons (b). The spectrum with thick line corresponds to the average spectrum of the whole data set.

are considerably different from those of average spectrum since there are only methylene (-CH_2-) vibrations in cyclic ring structure. The distinct peaks are observed around 4270, 4180, and 4100 cm^{-1}. Cyclic ring structure provides the characteristic spectral features. Figure 2(b) shows the NIR spectral features of benzene and toluene. Benzene spectrum is separately shown on the same plot because its spectral features are so peculiar. The strong band and several smaller bands are observed at 4050 and 4700-4500 cm^{-1} range, respectively. Toluene shows also distinct spectral features. The unique peak of benzene ring is observed in the 4700-4500 cm^{-1} range, as in benzene. Additionally several absorption bands are observed in the 4500-4000 cm^{-1} range.

In overall, aromatic hydrocarbons show the most unique spectral features, especially benzene, then followed by naphthenes, iso-paraffins, and normal paraffins. It is expected that the variation of resolution is relatively less influential on the components with unique spectral features, while the more influence on the components with the less unique spectral features.

Figure 3 shows NIR spectra of *n*-hexane with four different resolutions (4, 8, 16, and 32 cm^{-1}). From the top, it corresponds to 4, 8, 16, and 32 cm^{-1} resolution. There are no

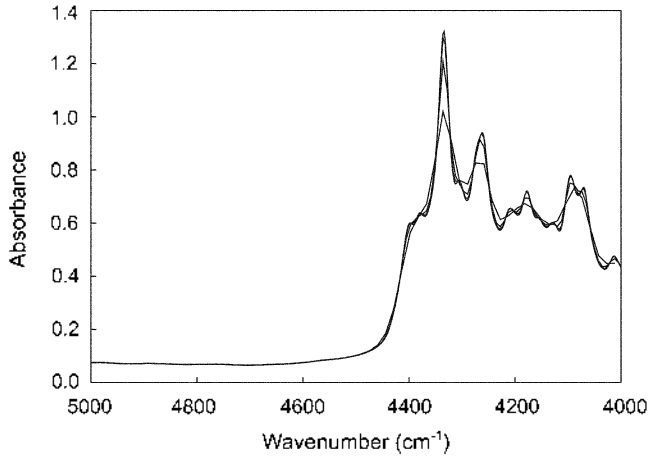

Figure 3. NIR spectra of *n*-hexane with four different resolutions (4, 8, 16, and 32 cm^{-1}). From the top, it corresponds to 4, 8, 16 and 32 cm^{-1} resolution.

significant differences between the spectra with 4 and 8 cm^{-1}. From the 16 cm^{-1} resolution, it is clear that absorption bands start to loose its features.

Calibration items. The mixtures used in this study were to mimic the NIR analysis of naphtha. In a practical view-

Table 2. The concentration maxima, minima, averages, and standard deviations of selected 11 PLS calibration items

Component	Maximum	Minimum	Average	Standard Deviation
Total Paraffin	49.04	33.68	41.94	3.78
Total normal-paraffin	35.93	20.32	30.15	3.09
Total iso-paraffin	17.07	6.60	11.79	2.35
Total naphthene	31.92	16.14	26.03	3.27
Total Aromatic	41.93	23.65	32.03	3.89
n-Hexane	12.69	3.66	7.31	2.15
n-Heptane	11.85	3.77	7.38	2.36
2,2-Dimethylbutane	4.17	0.18	1.90	1.04
Cyclohexane	10.51	3.65	6.96	1.87
Benzene	11.69	3.99	7.76	1.96
Toluene	10.79	3.89	6.98	1.99

*All units are in weight percentage

Table 3. The calibration results for each calibration item with different spectral resolutions

Component	SECV (4 cm^{-1})	SECV (8 cm^{-1})	SECV (16 cm^{-1})	SECV (32 cm^{-1})
Total Paraffin	0.54 (12)	0.51 (12)	0.52 (13)	1.25 (12)
Total normal-paraffin	0.44 (15)	0.37 (15)	0.43 (16)	0.76 (22)
Total iso-paraffin	0.36 (15)	0.37 (15)	0.40 (15)	0.96 (21)
Total naphthene	0.78 (14)	0.81 (14)	0.89 (14)	
Total Aromatic	0.81 (12)	0.81 (12)	0.84 (12)	1.60 (10)
n-Hexane	0.29 (16)	0.29 (17)	0.30 (17)	1.64 (17)
n-Heptane	0.29 (20)	0.28 (21)	0.42 (22)	
2,2-Dimethylbutane	0.24 (17)	0.23 (16)	0.23 (17)	0.78 (22)
Cyclohexane	0.16 (16)	0.17 (17)	0.17 (19)	0.37 (24)
Benzene	0.17 (6)	0.17 (6)	0.17 (9)	0.23 (20)
Toluene	0.21 (16)	0.21 (17)	0.26 (16)	0.32 (19)

*The numbers in parenthesis correspond to the number of PLS factors used.

point, the most typical analyses are the group compositional analyses: total paraffin, total normal paraffin, total iso-paraffin, total naphthene, and total aromatic.[8,9] Therefore, primarily these five PLS calibration items were selected. The conventional method of analyzing group compositions is a gas chromatography (GC).[10,11]

For the calibration items of pure components, *n*-hexane and *n*-heptane were selected in normal paraffins. As observed in Figure 1(a), these two components show the similar features, but slightly broader spectral features for *n*-heptane. Presumably there are some differences in calibration results between two components when spectral resolution is varied. As iso-paraffin and naphthene, 2,2-dimethylbutane and cyclohexane were selected, respectively. Benzene and toluene were chosen as aromatic hydrocarbons. As discussed above, benzene shows the most distinct spectral features, so it could be expected that the calibration is relatively less susceptible on the variation of spectral resolutions. Overall 11 PLS calibration items were selected.

Table 2 shows the concentration maxima, minima, averages, and standard deviations of selected 11 PLS calibration items. All units are in weight percentage. Out of 11 items, 5 items of hydrocarbon group compositions are in fairly high concentration ranges. The concentration ranges of other 6 items (pure components) are in lower than those of 5 group compositions.

PLS calibrations. PLS (Partial Least Squares) regression was used to build calibration models for each calibration item with different spectral resolutions. In the application of the PLS algorithm, it is generally known that the spectral range and the number of PLS factors are critical parameters.[12] The spectral range determines the location and quality of spectral information, and the number of PLS factors should be optimally selected to avoid an overfitting. For the spectral range, only the 5000-4000 cm^{-1} range including the whole spectral information was used.

The optimum number of factors was identified as the one that gave a minimum SECV (standard error of cross validation). The cross validation was used as a validation method by dividing the data set into 11 segments. The decrease pattern of SECV as a function of number of PLS factors was examined. Usually, the SECV was decreased sharply for the first several factors and gradually decreased for the following factors. At a certain factor, it started to increase. This trend is fairly typical in multivariate calibrations. The factor before the increase of SECV was chosen as the optimum number of factors.

The calibration results with different spectral resolutions are summarized in Table 3. Numbers in parenthesis correspond to the number of PLS factors used. Empty spaces mean that no acceptable results are achieved due to poor calibration. For group compositions, resulting SECVs are similar both at the 4 and 8 cm^{-1} resolution. However, SECVs start to be slightly worse at the 16 cm^{-1} and worst at the 32 cm^{-1} resolution. For total aromatic content, relatively the less number of factors was required due to non-overlapping aromatic features at 4700-4500 cm^{-1} range. For total aromatic content, 12 factors are required to achieve the optimal calibration; whereas, the more factors are necessary for total normal and iso-paraffin contents. It is required to use more factors (spectral description) for the more detailed discrimination between normal and iso-paraffin features.

For both *n*-hexane and *n*-heptane, the resulting SECVs are similar each other up to the 8 cm^{-1} resolution, and begin to increase from the 16 cm^{-1} resolution. However, for *n*-heptane, the increase of SECV at the 16 cm^{-1} resolution is greater than that of *n*-hexane. At the 32 cm^{-1} resolution, no calibration result was achieved for *n*-heptane since it was too poor. By comparing the spectra of *n*-hexane and *n*-heptane in Figure 1(a), the spectral features of *n*-heptane are slightly less characteristic compared to those of *n*-hexane. It was previously studied that the spectral features were broaden when the chain length of normal paraffin increased.[13] The degradation of spectral resolution is more influential for *n*-heptane since it requires to resolve the broader spectral features. Additionally, the less characteristic spectral features of *n*-heptane require the more numbers of PLS factors

compared to *n*-hexane.

As shown in Figure 1(b). 2.2-dimethylbutane (iso-paraffin) shows the more distinct spectral features than normal paraffins due to the branched structure. With the better spectral features. there are no significant differences in SECV and the number of PLS factors up to the 16 cm^{-1} resolution. The SECV is increased significantly and the more number of PLS factors is required at the 32 cm^{-1} resolution. The similar results are observed for cyclohexane (naphthene). In comparison with the results from normal paraffin, it is apparent that the components with distinct spectral features (such as cyclohexane) are relatively less dependent on the variation of spectral resolutions.

The spectral features of benzene are so unique that it is easy to distinguish by simple visual comparison. For benzene. the resulting SECVs are similar up to the 16 cm^{-1} resolution as other components: however, the fewer number of PLS factors (only six factors) is used to describe the spectral variation. For toluene. the similar results are obtained as the case of cyclohexane.

Conclusions

The overall results show the spectral resolution is an important factor to determine the PLS calibration performance, especially in the complex chemical matrix. For the analysis of petroleum products such as naphtha. the minimum of 8 cm^{-1} resolution is required to achieve reasonable calibration performances. When NIR measurement is considered to measure a specific component of petroleum products. both its spectral features and spectral resolution should be considered simultaneously. Therefore. the end-users in petroleum and petrochemical industries should consider the spectral resolution as one of important criteria when they have to decide the type of NIR instrument for a given application.

Future researches will be directed to perform the analogous experiments using the first overtone band (6300-5200 cm^{-1}) region that shows the broader spectral features than those in combination band (5000-4000 cm^{-1}) region. The overtone band region is more frequently used in combination with optical fiber for practical remote analysis and easy spectral collection; therefore. it would be the great interest for many practical NIR users.

Acknowledgements. This work is supported by the Korea Science and Engineering Foundation (Grant number: R14-2002-004-01000-0).

References

1. Burns. D. A.; Ciurczak, E. W. *Handbook of Near-Infrared Analysis*; Marcel Dekker: New York, U. S. A., 1992.
2. Wetzel. D. L. *Anal. Chem.* **1983**. *55*. 1165A.
3. Martens. H.: Naes. T. M. *Multivariate Calibration*. John Wiley and Sons: New York. U. S. A.. 1989.
4. Beebe. K. R.: Pell, R. J.; Seasholtz. M. B. *Chemometrics: A Practical Guide*; John Wiley and Sons: New York. U. S. A.. 1998.
5. Stark. E.; Luchter. K.; Margoshes, M. *Appl. Spec. Rev.* **1986**. *22*(4). 335.
6. Watkins. R. N. *Petroleum Refinery Distillation*. Gulf Publishing Company: Houston. 1973.
7. Wiseman. P. *An Introduction to Industrial Organic Chemistry*; Applied Science Publishers Ltd: London. 1976: Chapter 2.
8. Kosal. N.; Bhairi. A.; Ashraf Ali. M. *Fuel* **1990**, *69*(8). 1012.
9. Ku. M. S.: Chung. H.: Lee. J. S. *Bull. Korean Chem. Soc.* **1998**. *19*. 1189.
10. Leveque. R. E. *Anal. Chem.* **1967**. *39*(14). 1811.
11. Petrakls, L.; Allen. D. T.: Gavalas. G. R.; Gates, B. C. *Anal. Chem.* **1983**, *55*(9). 1557.
12. Lee. J. S.: Chung. H. *Vibrational Spectrosc.* **1998**. *17*. 193.
13. Chung. H.: Lee. J. S.: Ku. M. S. *Appl. Spectrosc.* **1998**. *52*. 885.