Determination of Research Octane Number using NIR Spectral Data and Ridge Regression

Hoeil Chung,* Hyeseon Lee,† and Chi-Hyuck Jun†

NIR Project Team, SK Corporation, 110 Kosa-Dong, Nam-Gu, Ulsan 680-130, Korea
†Department of Industrial Engineering Pohang University of Science and Technology, San31 Hyoja-Dong, Nam-Gu, Pohang 790-784, Korea
Received May 16, 2000

Ridge regression is compared with multiple linear regression (MLR) for determination of Research Octane Number (RON) when the baseline and signal-to-noise ratio are varied. MLR analysis of near-infrared (NIR) spectroscopic data usually encounters a collinearity problem, which adversely affects long-term prediction performance. The collinearity problem can be eliminated or greatly improved by using ridge regression, which is a biased estimation method. To evaluate the robustness of each calibration, the calibration models developed by both calibration methods were used to predict RONs of gasoline spectra in which the baseline and signal-to-noise ratio were varied. The prediction results of a ridge calibration model showed more stable prediction performance as compared to that of MLR, especially when the spectral baselines were varied. . In conclusion, ridge regression is shown to be a viable method for calibration of RON with the NIR data when only a few wavelengths are available such as hand-carry device using a few diodes.

Keywords: Near infrared spectroscopy, NIR, Multiple linear regression (MLR), Ridge regression, Collinearity, Gasoline, Research octane number (RON).

Introduction

Multiple linear regression (MLR)[1-5] is one of the most popular calibration methods in near-infrared (NIR) spectroscopy.[6-7] In comparison to other calibration methods, MLR is simple, easy to understand, and possible to clearly rationalize the relationship between the NIR spectral features and the calibration results. Especially, when miniaturization of instrumentation is necessary, such as a hand-held device, MLR can be successfully utilized in conjunction with such simple instrumentation as using a few diodes for a light source, as well as without a monochromator. In this case, conventional factor based analyses such as PCR (Principal Component Regression) and PLS (Partial Least Squares) regression can not be utilized since only 3 to 4 discrete wavelengths (variables) are available. Therefore, factor-based calibration methods can not be applicable to diode-based hand-held instrumentation even though their calibration performance is usually better than MLR in most NIR application fields. There are several methods of selecting wavelengths (variables) such as forward selection, backward elimination, and stepwise regression by examining the statistical parameters.[8,9]

In general, variables (absorbances at wavelengths) in a NIR spectrum are highly correlated each other (which is referred to as a multicollinearity). In the presence of multicollinearity, estimates of least square methods including MLR are unstable and tend to lead to poor prediction. It is known that biased estimation methods give considerably better prediction than ordinary least squares (OLS) when spectral data are noisy or the predictors are highly collinear. In this study, ridge regression[10] (a biased estimation method) has been evaluated and the prediction performance was compared with that of OLS-based MLR. To compare ridge regression and MLR, the determination of research octane number (RON) of gasoline [11,12] has been studied. Calibration models were initially developed using both methods, then each model was used to predict artificially modified spectra in which the baseline and signal-to-noise ratio were intentionally varied. The variations of baseline or signal-to-noise ratio are practically occurred in many actual fields. The results showed that the calibration model developed by ridge regression predicted RONs of gasoline samples with greater stability compared to that from MLR, especially when the baselines of spectra were changed.

*Corresponding Author. e-mail: hoeil@hotmail.com

Ridge Regression

Consider the MLR calibration model (with an intercept) having a single response variable Y and p explanatory variables $X_1,..., X_p$ with n samples. The estimator vector of regression coefficients by the OLS is given by

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'Y \quad (1)$$

where X is $n\times(p+1)$ matrix consisting of the sample data of variables $X_1,..., X_p$. If X_i's are highly correlated, the determinant $X'X$ of is near 0 and so $X'X$ becomes near-singular. Therefore, in the presence of multicollinearity, the OLS estimates could become very unstable due to the large variance of the estimates, which leads to poor prediction.

Ridge regression is one of several methods to overcome the multicollinearity problem by modifying the OLS to allow a

small bias *via* a constant *k* in the parameter estimates:

$$\hat{\beta}_R = (X'X + kI)^{-1}X'Y \quad (2)$$

where I is the identity matrix. When an estimator has only a small bias and is more precise than an unbiased estimator, it will be closer to the true parameter's value. If the variance of the ridge estimator $\hat{\beta}_R$ could be tremendously reduced, the mean squared error tends to be smaller than the OLS. In this situation the point estimate $\hat{\beta}_R$ becomes more stable, and the confidence interval of $\hat{\beta}_R$ is narrower.

In ridge regression, X_i's are recommended to be transformed by the correlation form, which makes the diagonal element of $X'X$ equal 1 and the off-diagonal element represents the correlation coefficient of the two variables. Since the values of all elements are of the same order of magnitude, this would control round-off errors in inverting $X'X$ to obtain the ridge estimator $\hat{\beta}_R$.

Experimental Section

Instrument and Apparatus. NIR spectra were collected with a NIRSystems on-line model 5000 spectrometer (Foss NIRSystems Inc., Silver Spring, MD) equipped with a quartz halogen lamp, PbS detector, and a fiber optic interactance probe. The resolution of collected spectra was 10 nm with 2 nm data point intervals. The fiber optic probe consisted of concentric rings of illuminating fibers, receiving fibers, and a reflecting mirror. The size of fiber optic probe was 2.54 cm (outer diameter) and 15.2 cm (length). The distance between the optical fibers and the reflecting mirror was 1 cm, resulting with an actual pathlength of 2 cm.

Sample Preparation. Fifty-eight different unleaded gasoline samples were prepared by randomly blending 9 different gasoline feed stocks. Each feed stock has different chemical and physical properties. All gasoline feed stocks were obtained from SK Corporation at Ulsan, Korea. The gasoline samples ranged in research octane number (RON) from 90.5 to 98.4 (average: 94.8, standard deviation: 2.0). Samples were stored in a refrigerator at 4 °C to prevent evaporation of the hydrocarbons. RONs of samples were determined with a conventional knock engine.[13]

A total of 58 spectra were divided into 43 spectra for the calibration and 15 spectra for the validation data set. Spectra in the validation data set were randomly chosen.

Results and Discussion

Spectral Features. Gasoline, as generally known, is a highly complex mixture of various carbon chain length hydrocarbons and oxygenates such as methy *t*-butyl ether (MTBE). Therefore, the resulting NIR spectrum of gasoline is the summation and highly overlapped spectral features of each component in gasoline. All the NIR spectra of gasoline samples in the calibration set are shown in Figure 1. As shown in Figure 1, the spectral features of gasoline are dominated by overtone and combination bands of CH, CH_2, and CH_3 of various hydrocarbons. The most useful spectral information is located in the 1100 to 1650 nm and 1800 to 2100 nm spectral ranges. The absorption bands from 1100 to 1270 nm and 1800 to 2100 nm correspond to second overtone and combination bands, respectively. The 1650-1800 and 2100-2500 nm ranges contain no useful spectral information due to the strong absorption of the NIR radiation from the relatively long optical pathlength. Therefore, the spectral ranges of 1100-1650 and 1800-2100 nm were solely used for calibration in this study.

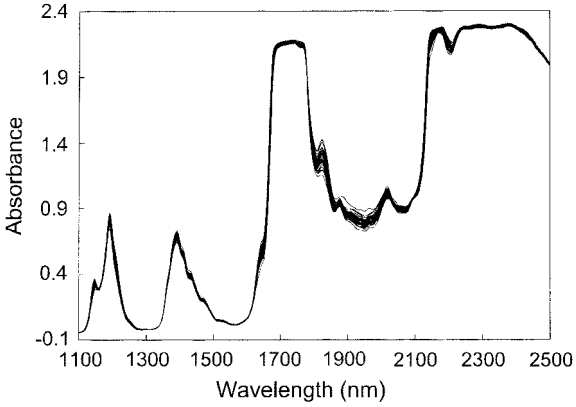

Figure 1. NIR (near-infrared) spectra of gasoline samples in the calibration set.

Table 1. MLR calibration results as increasing the number of variables in the calibration model

Unit: Research Octane Number (RON)

Number of wavelengths	Selected wavelengths (nm)	SEC
2	1190, 1812	0.37
3	1190, 1812, 2068	0.33
4	1190, 1812, 2068, 1824	0.29

MLR Calibration. For wavelength selection, stepwise regression was used. Stepwise regression is the most commonly used method of selecting a variable set, which is conduct as a step-by-step procedure by either adding or deleting one variable a time.[9] The number of wavelengths (variables) in the calibration model started with two and increased until the Standard Error of Calibration ($SEC = \sqrt{\sum_{i=1}^{n_c} (\hat{y}_i - y_i)/n_c}$) was similar to or slightly below 0.3, which is the ASTM reproducibility of the reference knock engine.[13] Table 1 shows the calibration results as increasing the number of variables (wavelengths) in the calibration model. With increasing number of variables, which corresponds to adding more spectral information, the resulting SEC is continuously decreased. Finally, the four-wavelength MLR calibration model was selected because it met the ASTM reproducibility.

The variation inflation factors (VIF) of this model were obtained to examine the degree of multicollinearity. The VIF for the *i*-th regression coefficient, VIF_i, is computed as follows:

$$VIF_i = \frac{1}{1 - R_i^2} \quad (3)$$

Table 2. The statistical results of four-wavelength MLR calibration model

Wavelength (nm)	Estimate	Standard Error	t-value (p-value)	VIF
1190	0.2229	0.0632	3.530 (0.0011)	7.20
1812	–2.2302	0.2345	–9.510 (0.0000)	99.34
1824	0.7408	0.2150	3.446 (0.0014)	83.50
2068	0.4478	0.1001	4.474 (0.0001)	18.09

Degree of Freedom: 39

where R_i^2 is the coefficient of determination (R-square) from the regression of X_i on the other independent variables. If R_i^2 is near 1.0, then VIF_i becomes large. The Eq. (4) shows that the variance of the *i*-th regression coefficient is inflated proportional to VIF_i:

$$Var(\hat{\beta}_i) = \frac{\sigma^2}{\sum_j x_{ij}^2}(VIF_i) \quad (4)$$

where x_{ij} is the *j*-th centered sample value of the *i*-th independent variable and σ^2 is the variance of error terms in the MLR model. As a rule of thumb, if VIF exceeds 30 (or 10 in other areas than chemometrics), it is an indication that the associated coefficients are poorly estimated because of multicollinearity.

The resulting estimate, standard error, *t*-value, and VIF of the four-wavelength MLR model are shown in Table 2. The estimates in Table 2 are the results of the correlation transformed data, which is $(\boldsymbol{X}-\bar{\boldsymbol{X}})/\boldsymbol{s}_{\boldsymbol{X}}\cdot\sqrt{n-1}$ and $(\boldsymbol{Y}-\bar{\boldsymbol{Y}})/\boldsymbol{s}_{\boldsymbol{Y}}\cdot\sqrt{n-1}\cdot$ to be compared with the ridge regression estimates in same magnitude. The VIF values at 1812 and 1824 nm are especially high. The results show that the calibration model with four variables apparently has multicollinearity problem that will adversely affect the prediction performance.

Ridge Calibration Model. One of the most important parameters in ridge regression is the bias constant (*k*). The bias constant in ridge regression is used to reduce the variance of estimates of regression coefficients that are due to multicollinearity. Several methods have been proposed for determining the optimal value of *k*.[14,15] A common strategy is to determine the smallest *k* that makes stable coefficients in the ridge trace with the lowest values of VIF. The ridge trace is the plot of VIFs versus *k* values. If *k* is too large, the resulting analytical performance will be degraded by applying a large bias, even though the collinearity problem can be solved. On the contrary, when the bias is too small, still the collinearity problem may exist in a calibration model. The set of variable in RR is selected by the stepwise method in MLR.

Figure 2 shows the relationship between *k* and VIF as increasing *k* in the four-variable model. The VIFs in the selected wavelengths are high when *k* is near to 0, but they decrease as the value of *k* increases. Most noticeably, VIFs are very high at 1812 and 1824 nm, and they drop steeply as the bias constant increases. The bias constant of 0.005 was

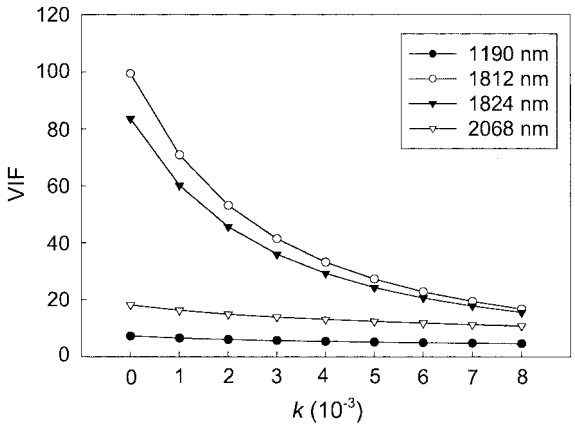

Figure 2. The relationship between *k* (bias constant) and VIF of each variable as increasing *k*.

Table 3. The statistical results of four-wavelength ridge calibration model using a bias constant of 0.005

Wavelength (nm)	Estimate	Standard Error	t-value (p-value)	VIF
1190	0.3296	0.0600	5.493 (0.0000)	5.08
1812	–1.4648	0.1389	–10.546 (0.0000)	27.12
1824	0.0684	0.1311	0.522 (0.6046)	24.26
2068	0.2628	0.0934	2.814 (0.0076)	12.32

Degree of Freedom: 39

selected by examining the ridge trace. At this point, the estimated coefficients are stable and their VIFs become smaller. With a small bias of 0.005, the standard errors of wavelengths of 1812 and 1824 nm are reduced to almost half as compared to those in MLR and moderate VIF values are achieved. The results of ridge regression using a bias constant of 0.005 are summarized in Table 3. In comparison to MLR results, VIF values are greatly decreased and a more statistically stable model is achieved, although R^2 is slightly decreased.

Generation of Artificial Spectra. To compare the performance of MLR and RR for pertubed spectral data, the baseline and signal-to-noise ratio of the gasoline spectra in the validation set were artificially changed to simulate spectral variations as typically observed in an actual field environment. The baseline and the signal-to-noise ratio of a spectrum change due to instrumental drift, source intensity variation, or degradation of optical components. To generate spectra with more baseline variation, the average baseline of the spectra in the validation set were initially obtained. For this purpose, three wavelengths at 1100, 1300, and 1564 nm were used to calculate the straight baseline of each spectrum using least squares. As shown in Figure 1, absorbances at these three wavelengths are almost zero. After obtaining the straight baseline of each spectrum, an average and standard deviation of both slope and intercept were calculated. The calculated results are summarized in Table 4. As shown in this table, the intercept and the slope are changing by 2.5 to 3.0 %, relatively. To generate the spectra with more baseline variation, 15 random numbers of slope and intercept in the

Table 4. Calculated average and standard deviation of slope and intercept of baselines of spectra in the validation data set. Relative Standard Deviation (RSD) is the ratio in percentage of standard deviation to average

	Average	Standard Deviation	RSD (%)
Slope	–0.18450	0.00461	2.50
Intercept	0.0001260	0.0000037	2.96

level of 2 and 3 times higher than those of the original validation set were calculated. Finally, the modified slope and intercept points were re-applied to generate the synthetic spectra. These spectra will be referred to as 2-times baseline-amplified and 3-times baseline-amplified spectra, respectively. Figure 3 shows the resulting spectra with increased baseline variations. Compared to the original spectra, the larger baseline variations are clearly and visually observed.

To generate spectra with higher noise, it is required to know the signal-to-noise ratio of the original spectra. To calculate the signal-to-noise ratio, twenty NIR spectra of the same gasoline sample were continuously collected over a 10 hour period at 30 minute intervals over the same day. A background single beam spectrum of air was collected for each measurement immediately before the collection of the sample single beam spectra. Then, the subtracted spectrum was produced by subtracting each spectrum from the following spectrum continuously. This was conducted to examine only noise information by removing analytical information. A total of 19 subtracted spectra were produced and the corresponding spectra are shown in Figure 4. To evaluate noise level, RMS (root mean square) noise was calculated using the standard deviation of data points based on the straight baseline. To calculate RMS noise, 1800-2100 nm range was used because the noise level in this range was higher and baselines were straight, compared to 1100-1650 nm range. The calculated results are summarized in Table 5. Compared to the average absorption, the RMS noise level is very low.

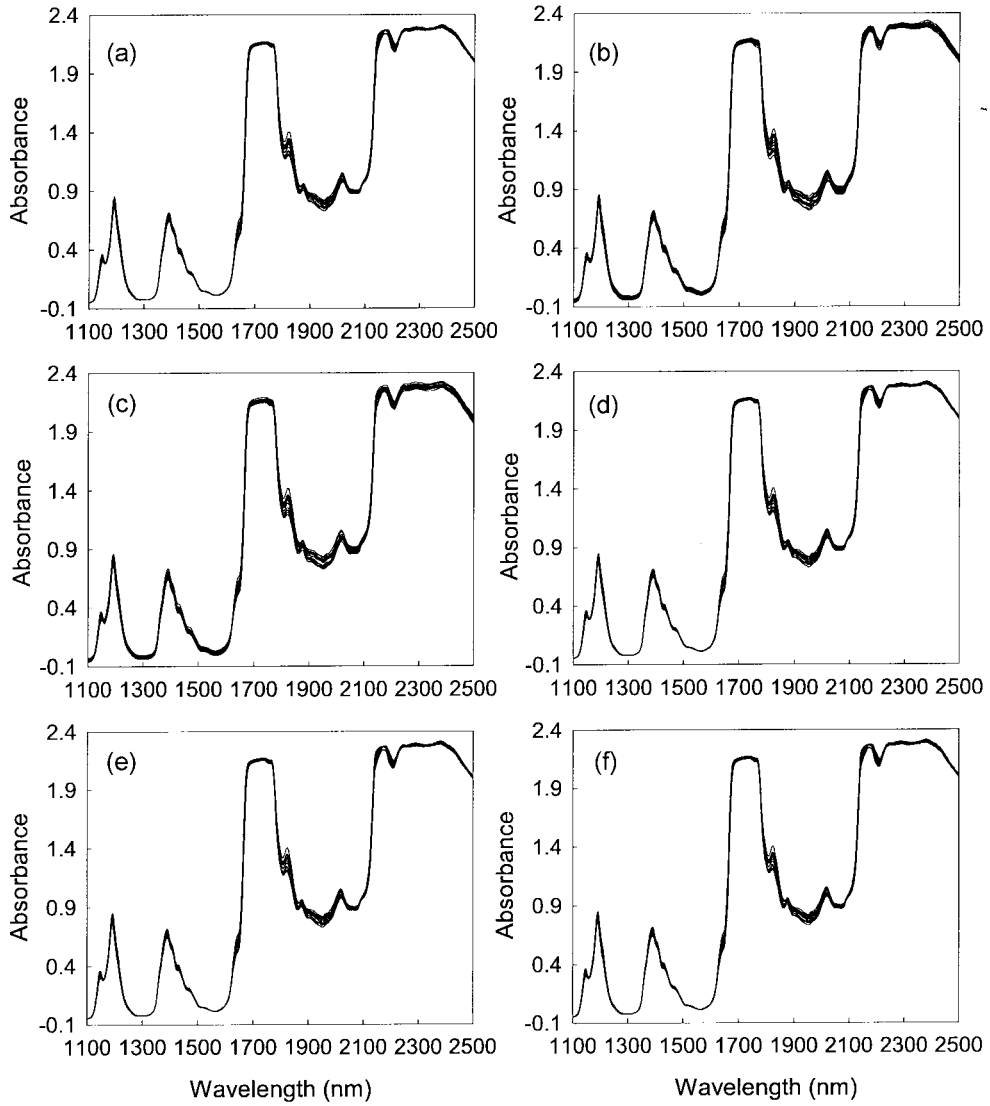

Figure 3. Original NIR spectra (a), 2-times baseline-amplified spectra (b), 3-times baseline-amplified spectra (c), 2-times noise-amplified spectra (d), 3-times noise-amplified spectra (e), and 10-times noise-amplified spectra (f).

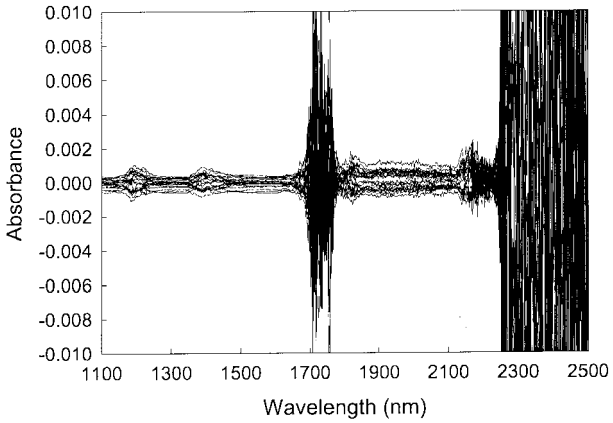

Figure 4. Subtracted spectra produced by subtracting each spectrum from following spectrum out of twenty spectra of the same gasoline sample.

Table 5. The results of RMS noise and average intensity using 1800-2100 nm range Unit: Absorbance Unit (A.U.)

RMS noise	Average Intensity
0.0016	0.8764

To generate more noisy spectra, 15 random numbers in the level of 2, 3, and 10 times higher than the calculated RMS noise were chosen. Finally, randomly generated noise factors were applied to generate the synthetic spectra. These spectra will be referred to as 2-times noise-amplified, 3-times noise-amplified, and 10-times noise-amplified, respectively. Figure 3 shows the resulting spectra with the noise amplification. As discussed earlier, the noise level in the original spectra is very low and no significant visual differences are observed even with 10-times noise-amplified spectra.

Prediction with Artificially Altered Spectra. To evaluate the calibration robustness using ridge regression, artificially altered spectra were predicted. The original and artificially altered spectra were used to predict RONs using both OLS-based MLR and ridge calibration models, and the resulting SEPs (Standard Error of Prediction, SEP= $\sqrt{\sum_{i=1}^{n_p}(\hat{y}_i - y_i)/n_p}$) are compared. The results are summarized in Table 6. The prediction of original spectra using the MLR calibration model shows slightly better results compared to that of the ridge model, which is due to an inten-

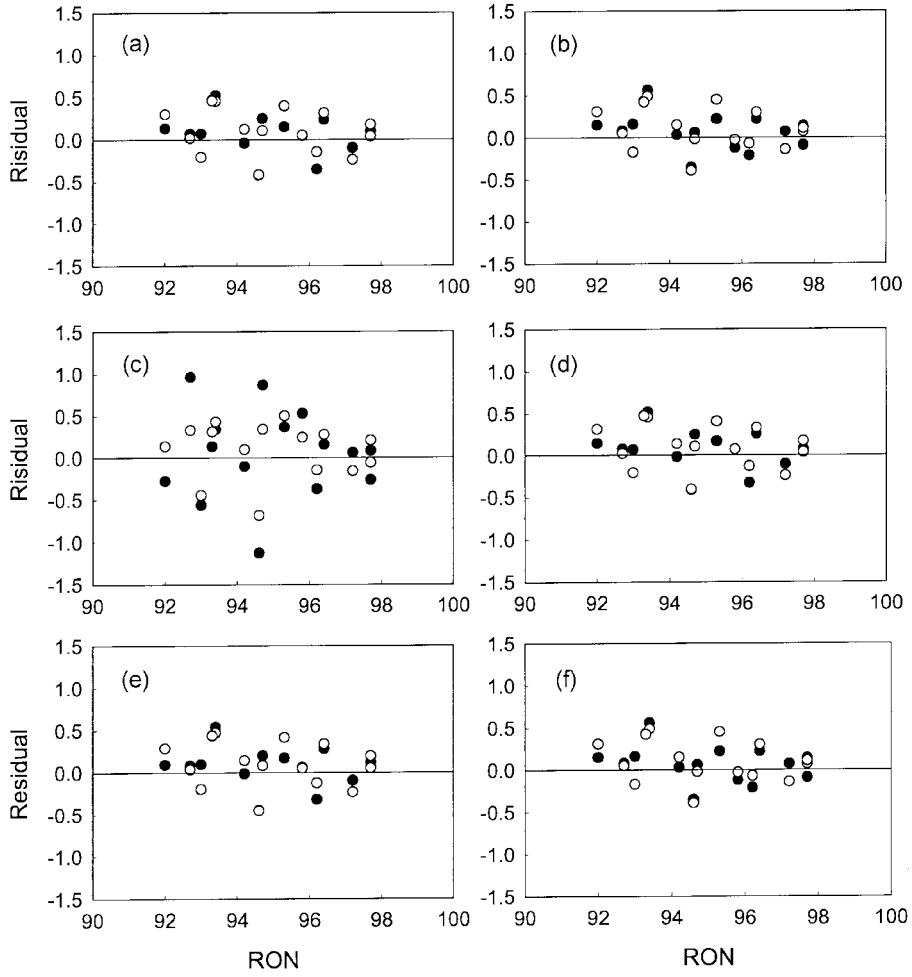

Figure 5. The prediction residual plots using MLR (filled circles) and ridge (open circles) calibration models for original NIR spectra (a), 2-times baseline-amplified spectra (b), 3-times baseline-amplified spectra (c), 2-times noise- amplified spectra (d), 3-times noise-amplified spectra (e), and 10-times noise-amplified spectra (f).

Table 6. The prediction results of original and altered spectra using MLR and ridge calibration models

Unit: Research Octane Number (RON)

Spectra Used	SEP (MLR)	SEP (Ridge Regression)
Original	0.26	0.27
2-times baseline-amplified	0.43	0.26
3-times baseline-amplified	0.53	0.33
2-times noise-amplified	0.26	0.28
3-times noise-amplified	0.26	0.28
10-times noise-amplified	0.25	0.27

tional bias is associated in the ridge calibration model. However, when the baseline-varied spectra were predicted, the ridge calibration model exhibits much more stable prediction performance. As baseline variation is increased 2 and 3 times, SEPs from MLR calibration models are significantly increased, while SEPs from ridge regression are relatively stable even with baseline variations. This result clearly shows that, by decreasing multicollinearity by ridge regression, the prediction result can be more robust and less sensitive to instrumental variation. When both models are used to predict with RONs of the noise-amplified spectra, no significant changes in prediction results are observed as increasing noise in the spectra. As discussed, the noise level in these NIR spectra is very low, therefore, it does not directly influence the prediction performance of each calibration model even with 10-times noise-amplified spectra.

Figure 5 shows the prediction residual plots corresponding to the results in Table 6. Filled and open circles correspond to MLR and ridge regression results, respectively. As expected, no significant differences are observed between the original and the noise-amplified spectra. However, in the prediction residual plot for 3-times baseline-amplified spectra, the residuals from the MLR model are much more scattered, while those from ridge regression are less scattered (which shows more stable prediction performance).

Conclusion

This comparative study clearly presents that the calibration model built from ridge regression is more stable and robust, especially in the situation of spectral variation. The spectral variations due to instrumental changes are commonly observed in a field environment. Ridge regression provides more robust estimates than MLR estimates for pertubations such as baseline change and noise in the data. The ridge estimators are stable in the sense that they are not affected by slight variation and tend to give more accurate predicted value. Using factor based calibration methods, such as PLS (Partial Least Squares) and PCR (Principal Component Regression), the collinearity problem can be eliminated. However, these methods are effective and useful when many variables (full or fairly wide range of spectrum) are available. In developing a small hand-held analytical device using a few diodes, only a few variables (wavelengths) are available, which ridge regression will be the better choice compared to MLR for long-term stable analytical performance by removing or decreasing the collinearity problem.

References

1. Burns, D. A.; Ciurczak, E. W. *Handbook of Near-Infrared Analysis*; Marcel Dekker Inc.: New York, 1992; p 107.
2. Martens, H.; Naes, T. M. *Multivariate Calibration*; John Wiley and Sons: New York, U. S. A., 1989; p 64.
3. Brown, C. W.; Lynch, P. F.; Obremski, R. J.; Lavery, D. S. *Anal. Chem.* **1982**, *54*, 1472.
4. Norris, K. H.; Williams, P. C. *Cereal Chem.* **1984**, *61*, 158.
5. Mark, H. *Anal. Chem.* **1985**, *58*, 2814.
6. Wetzel, D. L. *Anal. Chem.* **1983**, *55*, 1165A.
7. Stark, E.; Luchter, K.; Margoshes, M. *Appl. Spec. Rev.* **1986**, *22*(4), 335.
8. Neter, J.; Kutner, M. H.; Nachtsheim, C. J.; Wasserman, W. *Applied Linear Statistical Models*, 4th ed.; Richard, D. Ed.; Irwin: Chicago, 1996; p 347.
9. Rawling, J. O. *Applied Regression Analysis*; Wordsworth & Brooks/Cole: California, 1988; p 175.
10. Rawling, J. O. *Applied Regression Analysis*; Wordsworth & Brooks/Cole: California, 1988; p 338.
11. Kelly, J. J.; Barlow, C. H.; Jinguji, T. M.; Callis, J. B. *Anal. Chem.* **1989**, *61*, 313.
12. Maggard, S. M. *U. S. Patent* **1990**, *506*, 391.
13. *American Society for Testing and Materials*; Annual Book of ASTM Standards, Method D2699, Philadelphia, 1997.
14. Lee, T.; Campbell, D. *Comm. in Stat. Theor. Meth.* **1985**, *14*, 1589.
15. Hoerl, A.; Kennard, R.; Baldwin, *K. Comm. in Stat.* **1975**, *4*, 105.