ELSEVIER

Chemometrics and Intelligent Laboratory Systems 37 (1997) 255–259

Chemometrics and intelligent laboratory systems

Two data sets of near infrared spectra

John H. Kalivas *

Department of Chemistry, Idaho State University, Pocatello, ID 83209, USA

Received 18 December 1996; revised 4 March 1997; accepted 8 March 1997

Abstract

Described in this paper are two data sets of near infrared (NIR) spectra that are now available for general use. One data set consists of NIR spectra of 100 wheat samples with known protein and moisture content. The second set contains NIR spectra of 60 gasoline samples with known octane numbers. Results from a recent wavelength selection study using the two data sets are summarized. Other results based on the new regression approach of cyclic subspace regression (CSR) are briefly described. Included in CSR are principal component regression, partial least squares and least squares. An explicit description of calibration and validation samples used in both investigations is provided. © 1997 Elsevier Science B.V.

Keywords: Near infrared spectroscopy; Wheat; Protein; Moisture; Octane number; Wavelength selection; Cyclic subspace regression; Principal component regression; Partial least squares

1. Introduction

The importance of having reference data sets accessible for general use was recently noted [1]. Without reference data sets, it becomes difficult to evaluate strengths and weaknesses of a newly developed methodology compared to other approaches. Having reference data sets available permits intercomparison and careful analysis of newly developed methods.

Our laboratory was recently provided with two near infrared (NIR) spectral data sets [2]. One set contends with wheat samples while the other set deals with gasoline samples. The data was used to investigate the possibility of improving quantitation using partial least squares (PLS) in conjunction with wavelength selection compared to using full spectra [3,4]. Two approaches were implemented for wavelength selection. One strategy consisted of selecting unique wavelengths and required an optimization algorithm such as simulated annealing or genetic algorithm. The other form of wavelength selection made use of spectral properties to exclude unfit spectral regions. In other words, window selection was studied where the window can be one or more consecutive wavelengths. Refs. [3,4] both describe simulated annealing results using the NIR spectra identified in this paper and Ref. [4] compares simulated annealing results with a window selection approach. Because Ref. [4] details an in-depth study of the data, only its results are summarized in this paper. Presented here is a description of which samples were used to form calibration and validation sets. If the same distribution is used by other investigators, results will be directly comparable to those presented in Ref. [4].

The two data sets were also used in studying a new

* Tel.: +1-208-2362165; fax: +1-208-2364373.

PII S0169-7439(97)00038-5

regression approach known as cyclic subspace regression (CSR) [5–7]. Ref. [6] contains the explicit mathematical relationship between principal component regression (PCR), PLS and least squares (LS) and the NIR spectral data sets were used to demonstrate the relationships. The same distribution of calibration and validation samples used in the wavelength selection study was used in the CSR investigation.

Both data sets are available as ASCII files at an ftp site. The URL for the site is sun.mcs.clarkson.edu in the directory pub/hopkepk/data. The readme.wheat and readme.gas files lists appropriate file names containing the data for wheat and gasoline samples, respectively.

2. NIR spectra of wheat

This data set contains 100 wheat samples with specified protein and moisture content. Samples were measured using diffuse reflectance as $\log(1/R)$ from 1100 to 2500 nm in 2 nm intervals. Of the 100 spectra, 87 were utilized in the studies described in Refs. [4,6]. Spectra were reduced to contain only 141 response by using every fifth response and then mean centered. The calibration set contains 50 spectra (WHT_C) and two validation sets were created encompassing 20 spectra each (WHT_{V1} and WHT_{V2}). Table 1 identifies which of the 100 samples were used to form calibration and validation sets. Note that three samples are duplicated in the calibration set. Samples were assigned to calibration and validation sets with the constraint that the protein content for validation sets be embedded in the calibration range.

Ref. [4] contains plots of loading vectors from PLS and the singular value decomposition. Also plotted are PLS regression vectors corresponding to different number of factors and wavelength correlation plots for protein and moisture. While the figures have not been duplicated here, Tables 2 and 3 contain error criteria values for PLS with wavelength selection. Note that depending on the wavelength region, error criteria increase or decrease. The reader should consult Ref. [4] for further discussion on this.

Work described in Refs. [5–7] show mathematically the connection between PCR, PLS and LS through CSR. Using the wheat data, it is shown that CSR generates solutions with the low factor advantage of PLS and at the same time, the low noise advantage of PCR [6]. Without going into the details, a solution matrix **S** is generated by CSR where **S** is a $k \times k$ upper triangular matrix and k is the mathematical rank of the $m \times w$ matrix of independent variables in which $k \leq \min(m, w)$. For these data sets, m

Table 2
Prediction error values using PLS for protein in wheat

Wavelength indices	d [a]	PRESS	SEV_{V1}	SEV_{V2}	SEC
All	12	5.28	0.46	0.40	0.18
1–25	9	4.76	0.29	0.35	0.21
1–10, 12–17, 19–23	8	4.44	0.31	0.36	0.22
40–80	9	5.44	0.50	0.47	0.25
90–120	10	9.58	0.76	0.38	0.25
80–100	9	13.44	0.64	0.53	0.35

[a] Number of factors used (prediction rank).

Table 3
Prediction errors using PLS for moisture in wheat

Wavelength indices	d [a]	PRESS	SEV_{V1}	SEV_{V2}	SEC
All	4	2.11	0.24	0.30	0.10
9–13, 81–90	6	2.86	0.22	0.30	0.28
81–91	4	5.87	0.29	0.30	0.29
41–75, 99–131	5	2.93	0.32	0.30	0.19

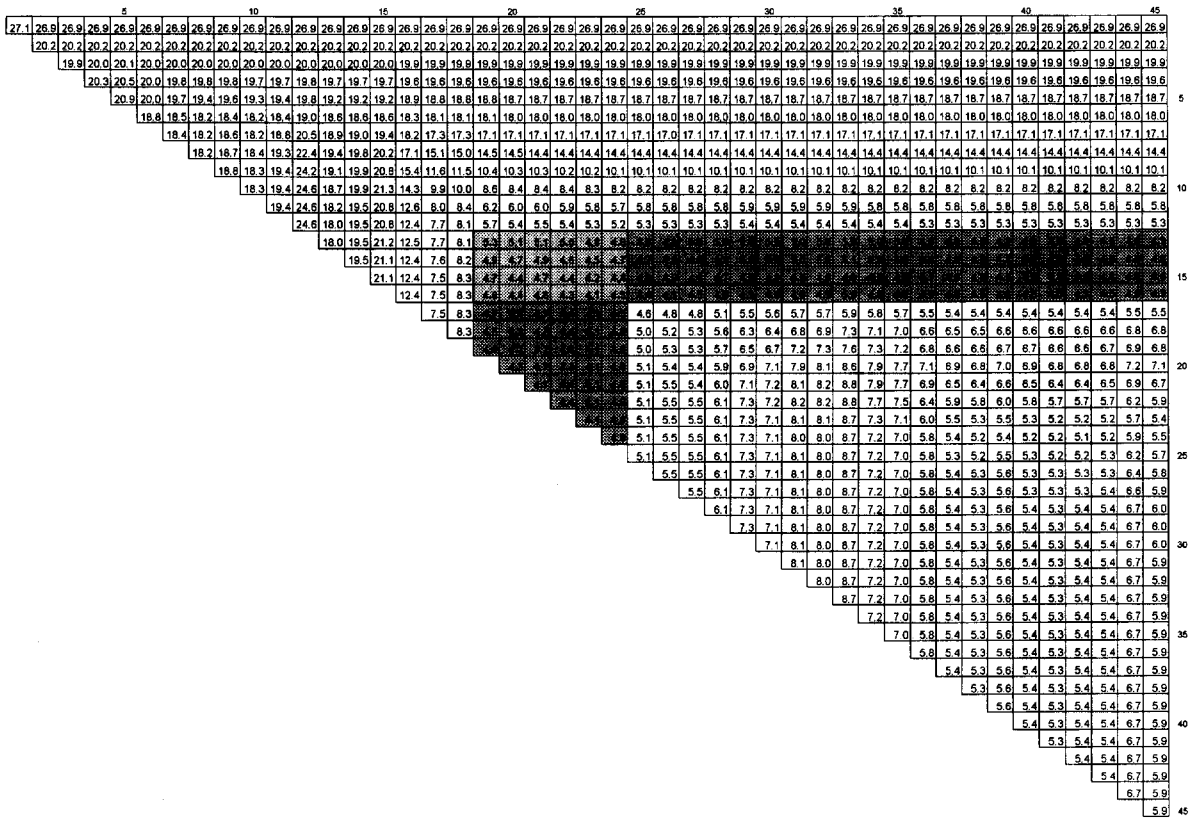
[a] Number of factors used (prediction rank).

Table 1
Calibration and validation samples for the wheat data

Sample No.	Function [a]
10, 12, 15, 16, 17, 22, 24, 25, 36–39, 41, 43–45, 49, 54–58, 60, 63–65, 69, 70, 72, 73, 77, 79, 82, 83, 85–94, 97–99	WHT_C [b]
1, 6, 13, 18, 21, 26, 34, 35, 46, 50, 51, 59, 62, 67, 74, 76, 80, 81, 84, 96	WHT_{V1}
14, 19, 23, 29–33, 42, 47, 48, 52, 53, 61, 66, 68, 71, 75, 78, 95	WHT_{V2}

[a] WHT_C = calibration sample, WHT_{V1} = validation sample in set 1 and WHT_{V2} = validation sample in set 2.
[b] Samples 10, 43 and 97 are entered twice in WHT_C.

Table 4
CSR cross-validation PRESS values for prediction of protein. For presentation clarity, values are displayed to one decimal place



represents the number of calibration spectra measured at w wavelengths. The PCR results are on the diagonal of S, PLS values are located in the last column (kth column) and the LS result is located at the $s_{k,k}$ position i.e. the intersection of the last column and the diagonal. Table 4 contains PRESS values for prediction of protein. Highlighted is a corridor of low PRESS values. Note that the better values are in the interior of S where the vertical and horizontal parts of the corridor intersect. For example, at $s_{16,23}$ the lowest PRESS value is obtained by CSR. This result is generated using 23 eigenvectors from the singular

Table 5
Calibration and validation samples for the gasoline data

Sample No.	Function [a]
1–3, 5–7, 10, 12–14, 16–20, 22, 25, 26, 28–31, 33, 36, 38–40, 42, 45, 47–49, 51, 53, 55–60	GAS_C
4, 8, 9, 11, 21, 23, 24, 27, 34, 35, 41, 43, 46, 50, 52	GAS_V

[a] GAS_C = calibration sample and GAS_V = validation sample.

Table 6
Prediction error values using PLS for octane number in gasoline

Wavelength indices	d [a]	PRESS	SEV	SEC
All	5	2.52	0.24	0.18
118–123, 155–175, 229–238, 255–270, 367–370	5	1.78	0.21	0.17
155–175, 229–238, 255–270, 367, 370	5	1.80	0.23	0.17
118–175, 229–270, 367–370	4	1.17	0.21	0.18
118–175, 229–270, 367–401	4	2.21	0.27	0.22
155–175, 381–401	8	4.54	0.35	0.21
155–175	4	3.93	0.36	0.26
381–401	6	46.33	1.26	0.82

[a] Number of factors used (prediction rank).

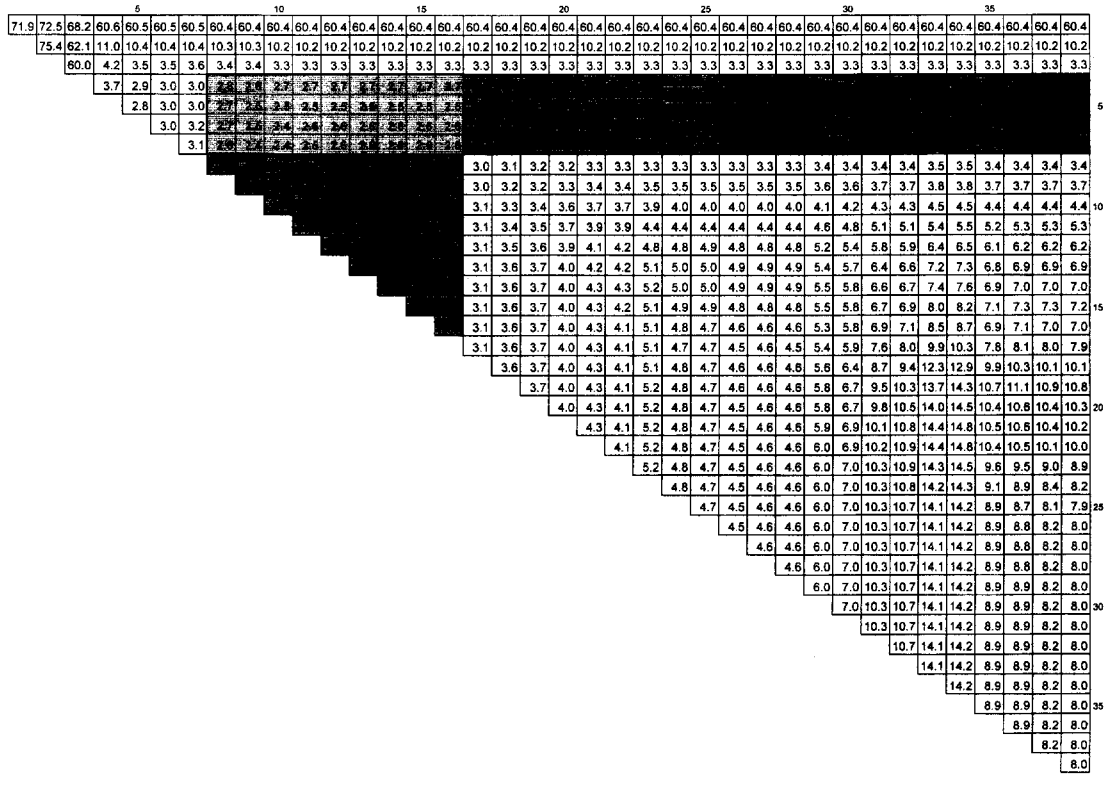
value decomposition of the calibration spectra and 16 factors. Further discussion on the interpretation of **S** is available in Ref. [6].

3. NIR spectra of gasoline

This data set contains 60 gasoline samples with specified octane numbers. Samples were measured using diffuse reflectance as $\log(1/R)$ from 900 to 1700 nm in 2 nm intervals. Of the 60 spectra, 55 were utilized in studies described in Refs. [4,6] and were mean centered prior to all computations. The calibration set contains 40 spectra (GAS_C) and a validation set consisting of 15 spectra (GAS_V) was used. Table 5 catalogs which of the 60 samples were used to form calibration and validation sets.

As for the wheat data set, Ref. [4] contains figures with loading vectors from PLS and the singular value decomposition for the gasoline data. Also available are PLS regression vector plots corresponding to different number of factors and wavelength correlation plots for octane number. The figures have not been

Table 7
CSR cross-validation PRESS values for prediction of octane number. For presentation clarity, values are displayed to one decimal place



duplicated here, but Table 6 contains the error criteria values for PLS with wavelength selection. Analogous to the wheat data, depending on the wavelength region, error criteria increase or decrease. The reader should consult Ref. [4] for further discussion on this.

Table 7 shows the CSR results for the gasoline data. Again, a corridor of low values are observed with the lowest values in the vicinity of where the vertical and horizontal parts of the corridor intersect. Ref. [6] should be consulted for further analysis.

Acknowledgements

The author is grateful to Dr. Prüfer of Bran + Luebbe for providing the described data sets and granting permission for release of the spectra. Dr. Hörchner is appreciated for his efforts in obtaining the data sets. Mr. Jason Brenchley is thanked for preparing the data sets for distribution.

References

[1] P.K. Hopke, D.L. Massart, Chemom. Intell. Lab. Syst. 19 (1993) 35–41.
[2] Dr. Heinrich Prüfer, Bran + Luebbe, Norderstedt, Germany.
[3] U. Hörchner, J.H. Kalivas, in: J.H. Kalivas (Ed.), Adaption of Simulated Annealing to Chemical Optimization Problems, Elsevier, Amsterdam, 1995, ch. 2.
[4] J.M. Brenchley, U. Hörchner, J.H. Kalivas, Appl. Spectrosc., in press.
[5] P.M. Lang, Technometrics, submitted.
[6] P.M. Lang, J.M. Brenchley, R.G. Nieves, J.H. Kalivas, Anal. Chem., submitted.
[7] J.M. Brenchley, P.M. Lang, R.G. Nieves, J.H. Kalivas, Chemom. Intell. Lab. Syst., submitted.