Comparison of calibrations based on partial least squares and multiple linear regression for near infrared prediction of composition and functionality in grains

Phil Williams

Canadian Grain Commission, Grain Research Laboratory, 1404-303 Main Street, Winnipeg, MB R3C 3G8, Canada. E-mail: pwilliams@cgc.ca.

Introduction

While neural network and genetic algorithm software patiently await their turn on the near infrared (NIR) stage, by far the majority of the calibrations in the thousands of NIR instruments employed in the working environment have been developed by multiple linear regression (MLR) or partial least squares (PLS) regression. Filter instruments still carry the heaviest burden of day-to-day grain analysis, although the rugged monochromator-based instruments are increasing in adoption as the older filter instruments are retired. Most filter instruments carry MLR-based calibrations, while the bench-top monochromator instruments in most common use are calibrated using PLS regression.

The use of PLS regression in NIR technology has been expertly described by Martens and Næs in 1987.[1] Since then a mild controversy has emerged, with one group favouring MLR and the other PLS. The optical data used by MLR in selection of wavelengths for calibration are highly correlated with one another and multi-collinearity is an ever-present hazard in deriving equations for prediction of composition or functionality by MLR. Partial Least Squares regression is based on the incorporation of principal components, which are derived from variance in spectral data and are orthogonal and reference data. As a result the possibility of erroneous conclusions drawn as a result of multi-collinearity is eliminated.

In theory PLS regression should mark an improvement over MLR. The speed of modern personal computers has facilitated comparison of MLR and PLS regressions derived from the same optical data. This publication presents seven such comparisons, using grain data. These range from "Difficult" studies wherein the NIR predictions are based on the influence of an infestation, or weather conditions on the texture of the grain, to the very "Simple" prediction of protein content in wheat flour.

Methods

Whole-kernel wheat samples were culled from railway carload deliveries at terminal elevators in Vancouver, Canada. Whole-kernel barley samples were culled from the western Canadian barley-breeding programme at Lacombe, Alberta, Canada. Whole-seed canola samples were selected from farmer's deliveries during the annual harvest surveys conducted by the Grain Research Laboratory of the Canadian Grain Commission. Flour samples were selected from plant breeders' annual trials tested at the Grain Research Laboratory. Comparisons 1–5 were carried out on whole seeds. Flour

Near Infrared Spectroscopy: Proceedings of the 9th International Conference

for comparisons 6 and 7 were milled at the Grain Research Laboratory, using the Allis–Chalmers laboratory flour mill. Analytical reference methods for oil, moisture, protein and ash contents are available on request, as are those for the determination of DON and Falling Number. True metabolisable energy (TME) was determined by the method of Zhang *et al*.[2]

Foss/NIRSystems NSAS software was used for the development of all calibrations, by MLR or PLS. The PLS calibration equations were derived using optical data transposed into the same mathematical format as that determined by optimisation, using MLR (forward stepwise regression). In all cases of PLS calibration, scatter correction methods were applied using the comprehensive WINISI software, but no improvement was gained by application of scatter correction.

Results and discussion

Table 1 summarises the statistics derived from the prediction studies by MLR and PLS. All of these data were the result of careful optimisation of mathematical treatment of the optical data. In all seven cases, differences in the values for *SEP* and RPD (4) were small and not statistically significant. In comparisons 3 and 7 PLS was slightly superior to MLR. The other five comparisons were essentially equal (for example, Nos. 2 and 5), or MLR was slightly superior.

Tables 2–8 provide details of the wavelengths selected by MLR, together with the areas of wavelength where display of the "weights" derived during development of PLS calibrations showed the degree to which variance in data was used in compiling the equations. This is illustrated further by Figures 1–4, which show the weights derived from the first two PLS factors for four of the comparisons. Several areas occurred where similar wavelengths were selected (by MLR) or employed in the development of the PLS equations, as indicated by the distribution of the weights. Consequently it is not surprising that the MLR and PLS approaches were essentially similar in their efficiency.

Columns 6 and 7 of Tables 2–8 illustrate the degree to which individual PLS factors "accounted for" the total variance upon which the equations were based. Four features were apparent:

1. The efficiency in prediction was not necessarily related to the degree to which total variance was accounted for (for example, FN and DON).

Table 1. Prediction statistics for all comparisons.

Constitutent	Commodity	MLR[a]				PLS				
		N	λs	*r*	*SEP*	*RPD*	Factors	*r*	*SEP*	RPD
DON	Wheat	53	8	0.856	0.744 ppm	1.92	10/11[b]	0.853	0.750	1.91
FN	Wheat	174	9	0.772	44.1 seconds	1.57	14/1	0.771	44.2	1.57
TME	Barley	56	8	0.942	0.223 units	5.14	7/14	0.944	0.218	5.26
OIL	Canola	52	8	0.969	0.818 %	4.07	8/14	0.958	0.954	3.49
Water	Canola	52	2	0.992	0.458 %	8.76	5/13	0.991	0.500	8.02
Protein	Wheat flour	104	6	0.998	0.108 %	14.26	9/12	0.997	0.121	12.73
Ash	Wheat flour	95	8	0.904	0.022 %	2.34	12/13	0.917	0.021	2.50

[a]MLR = multiple linear regression; PLS = partial least squares regression; DON = deoxy-nivaleno; FN = falling number; TME = true metabolisable energy; *N* = number of samples in predicition sample set; λs = number of wavelength points used in calibration equation; *r* = coefficient of correlation; *SEP* = standard error of prediction; *RPD* = ration of standard error of prediction to standard deviation of reference data in predicition sample set.
[b]Number of factors used in predicition equation/theorectical number of factors as identified by PLS software

Table 2. Association between wavelengths selected by MLR and weights generated during development of PLS calibration equations: deoxy-nivalenol (*Fusarium* Head Blight/"Scab") in wheat.

MLR[a] D1OD 10/4	PLS weights		Proportion of total variance per PLS factor		
λ (nm)	F	Assoc.λ (nm)	Source	Factor	Proportion
1840	1 +	–	–	1	21.6/ 21.6[b]
2220	3 +	2214	1	2	26.2/ 4.6[b]
1200	7 +	1206/1204	1/2	3	64.0/ 37.8[b]
1280	2 +	1280	1	4	66.7/ 2.7[b]
1360	5 –	–	–	5	68.7/ 2.0[b]
2440	4 –	2442	3	6	71.6/ 2.9[b]
1520	8 –	–	–	7	76.2/ 4.6[b]
2280	6 +	2278/2276	2/3	15	89.7/ –[b]

[a]MLR = multiple linear regression; PLS = partial least squares regression; S = small; M = medium; L = large ("bands")
[b]running total variance accounted for/proportion of variance accounted for per factor

Table 3. Association between wavelengths selected by MLR and weights generated during development of PLS calibration equations: falling number in wheat.

MLR[a] D2OD 20/10		PLS weights		PLS proportion of total variance per PLS factor	
λ (nm)	F	Assoc.λ (nm)	Source	Factor	Proportion
1264	2 –	1262/1262	2 S/3 S	1	35.0/ 35.0[b]
1924	3 –	1930/1920	1 VL/2 VL	2	37.1/ 2.1[b]
1864	4 –	1860/1862	1 VL/2 VL	3	38.4/ 1.3[b]
2084	7 +	–	–	4	41.0/ 2.6[b]
2144	6 –	2150	2 VS	5	44.4/ 3.4[b]
1184	1 –	1194	2 VL	6	48.2/ 3.6[b]
1364	5 +	1374	1 M	7	49.8/ 1.6[b]
2404	9 +	2412	3 M	8	52.0/ 2.2[b]
1984	8 +	–	–	15	58.1/ –[b]

[a]MLR = multiple linear regression; PLS = partial least squares regression; S = small; M = medium; L = large; V = very ("bands")
[b]running total variance accounted for/proportion of variance accounted for per factor

2. All factors subsequent to the first factor used in development of the **F**alling **N**umber prediction equation were most likely related to system noise, rather than true variance in optical or reference data.

Table 4. Association between wavelengths selected by MLR and weights generated during development of PLS calibration equations: true metabolisable energy in feed barley.

MLR* D2OD 4/10		PLS		PLS Proportion of total variance per factor	
λ (nm)	F	Assoc. λ	Source	Factor	Proportion %
2420	4	2412	2 M	1	73.1/ 73.1[b]
2360	6 –	2360	2 S	2	78.1/ 5.0[b]
1560	1	1568/1558	2 VS/3 S	3	81.4/ 3.3[b]
2280	3 –	2284/2278	1 L/3L	4	82.6/ 1.2[b]
1720	5	1722/1720	1 M/2S	5	84.8/ 2.2[b]
2140	7	2144/2134	3 S/2 VS	6	85.9/ 1.1[b]
1260	8 P	1268	1 S	7	89.1/ 3.2[b]
2400	2 –	2394/2412	1 M/2 M	15	98.0/ –[b]

[a]MLR = multiple linear regression, PLS = partial least squares regression; P = poor; S = small; M = medium; L = Large (V = very)
[b]running total variance accounted for/proportion of variance accounted for per factor

Table 5. Association between wavelengths selected by MLR and weights generated during development of PLS calibration equations: oil content in canola seed.

MLR* D20D 4/4		PLS		PLS Proportion of total variance per factor	
λ (nm)	F	Assoc. λ	Source	Factor	Proportion %
1440	2 +	1436	2 M	1	73.8/ 73.8[b]
1860	3 +	1872	3 VL	2	87.4/ 13.6[b]
1380	4 –	1386/1386	1 M/3 VL	3	91.1/ 3.7[b]
1220	1 –	1218/1214	2 S/3 S	4	92.9/ 1.8[b]
2420	7 + P	–	–	5	93.7/ 0.8[b]
1500	6 +	1516	3 S	6	95.3/ 1.6[b]
2400	8 + VP	–	–	7	95.6/ 0.3[b]
1260	5 –	1254	2 VS	15	99.2 / –[b]

[a]MLR = multiple linear regression, PLS = partial least squares regression; P = poor; S = small; M = medium; L = Large (V = very)
[b]running total variance accounted for/proportion of variance accounted for per factor

3. The pattern of the degree to which variance was accounted for by sequential PLS factors differed. In some cases, such as canola water content and flour protein content, most of the variance was accounted for by the first two factors and over 97% of the total variance by as few as three factors. In

Table 6. Association between wavelengths selected by MLR and weights generated during development of PLS calibration equations: water in canola seed.

MLR[a] D2OD 4/10	PLS	PLS Proportion of total variance per factor			
λ (nm)	F	Assoc. λ	Source	Factor	Proportion %
1840	1 +	1866/1886	2 M/3 VL	1	94.2/ 94.2[b]
1400	2 –	1410	1 VL	2	96.2/ 2.0[b]
–	–	–	–	3	97.6/ 1.4[b]
–	–	–	–	4	98.4/ 0.8[b]
–	–	–	–	15	99.8 –[b]

[a]MLR = multiple linear regression, PLS = partial least squares regression; P = poor; S = small; M = medium; L = Large (V = very)
[b]running total variance accounted for/proportion of variance accounted for per factor

Table 7. Association between wavelengths selected by MLR and weights generated during development of PLS calibration equations: flour protein content.

MLR[a] D2OD 4/10		PLS		PLS Proportion of total variance per factor	
λ (nm)	F	Assoc. λ	Source	Factor	Proportion %
2020	1 -	2016	1 M	1	74.1/ 74.1[b]
2420	5 +	2420	1 VS	2	99.0/ 24.9[b]
1980	3 -	1980/1984	1 M/2M	3	99.2/ 0.2[b]
2080	2 +	2084/2086	1VS/2M	4	99.4/ 0.2[b]
1460	4 +	1476/1454	1 S/1/S	5	99.6/ 0.2[b]
2100	6 +	2086/2096	2 M/3 VS	15	99.9/ –[b]

[a]MLR = multiple linear regression, PLS = partial least squares regression; P = poor; S = small; M = medium; L = Large (V = very)
[b]running total variance accounted for/proportion of variance accounted for per factor

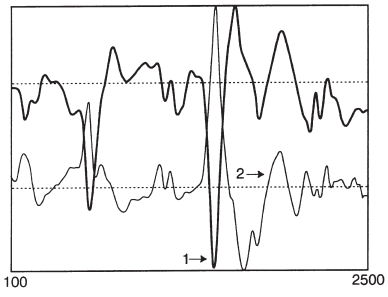

Figure 1. Distribution of weights consequent with development of PLS equations for the prediction of Vomitoxin (Deoxy-nivaleno) in wheat.

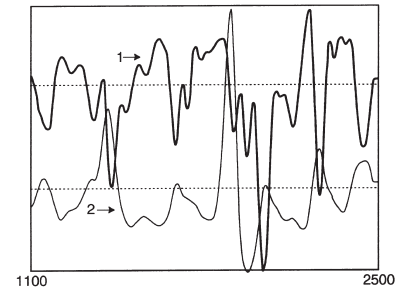

Figure 2. Distribution of weights consequent with development of PLS equations for the prediction of water in Canola seed.

Table 8. Association between wavelengths selected by MLR and weights generated during development of PLS calibration equations: flour ash content.

MLR* D1OD 10/20		PLS		PLS Proportion of total variance per factor	
λ (nm)	F	Assoc. λ	Source	Factor	Proportion %
1620	9 + VP	–		1	12.8/ 12.8[b]
1660	6 +	1660/1664	2/4	2	29.3/ 16.5[b]
1460	1 +	1464	2	3	31.5/ 2.2[b]
1740	5 +	1738/1746	1/4	4	50.1/ 18.6[b]
1960	4 –	1956	3	5	62.3/ 12.2[b]
2380	3 –	2368/2384	2/3	6	68.4/ 6.1[b]
1680	7 –	1678	4	7	71.8/ 3.4[b]
1160	8 – VP	–	–	8	80.2/ 8.4[b]
1700	2 +	–	–	15	92.6/ –[b]

[a]MLR = Multiple Linear Regression; PLS = Partial Least Squares Regression; λ = wavelength selected by MLR; Assoc. λ = wavelength identifiable from PLS "Weight", corresponding most closely to wavelength selected by MLR; Source = PLS "Weight'" for the respective factor; Proportion % + proportion of total variance accounted for by individual PLS factors; VP = Very Poor (low) F-value (less than 5)
[b]running total variance accounted for/proportion of variance accounted for per factor

the case of FN in wheat, TME in barley and water content in canola after the first factor, the remainder appeared to contribute little to the equations, but even in the water calibration, the best PLS predictions required at least five factors.

4. In comparisons 1 and 7 the progression in "accountability" of variance by individual factors was not regular in that Factor 3 was apparently significantly more important than Factors 1 or 2 in development of the equation. In comparison 7, the degree to which the first five factors apparently accounted for variance was quite different from that of all other comparisons.

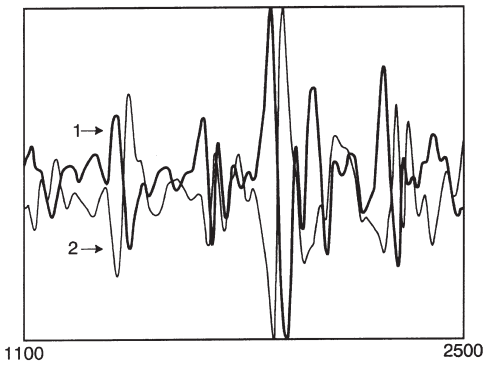

Figure 3. Distribution of weights consequent with development of PLS equations for the predictions of protein in wheat flour.

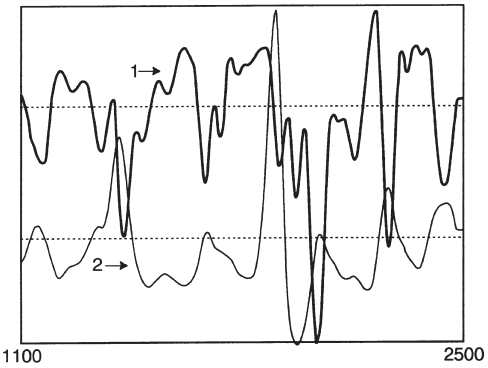

Figure 4. Distribution of weights consequent with development of PLS equations for the prediction of ash in wheat flour.

Conclusions

1. For a wide range of applications to grain analysis by NIR spectroscopy, the MLR and PLS methods were essentially equivalent in their efficiency in enabling prediction.
2. The most apparent reason for this is the fact that in computing the equations both approaches used optical data associated with similar wavelengths.

References

1. H. Martens and T. Næs, in *Near-infrared Technology in the Agriculture and Food Industries*, Ed by Phil Williams and Karl Norris. The American Association of Cereal Chemists, St Paul, MN, USA, pp. 55–87 (1987).
2. W.J. Zhang, L.D. Campbell and S.C. Stothers, *Can. J. Anim. Sci.* **74,** 355 (1994).