PLS Modeling the Starch Contents of Corn Data Measured Through Different NIR Spectrometers

Tahir Mehmood
School of Natural Sciences, National University of Sciences and Technology, Islamabad, Pakistan.
Email: tahime@gmail.com

***Abstract*—A variety of filter wavelength region selection algorithm, including loading weight PLS (PLS-LW), regression coefficient PLS (PLS-RC), variable importance on PLS (PLS-VIP) and selectivity ratio PLS (PLS-SR) and significant multivariate correlation (PLS-SMC) are considered in modeling the starch contents of corn with corn spectral data. Corn samples were measured on three different NIR spectrometers known as M5, Mp5 and Mp6. Hence, the class of filter PLS methods were imposed on each data set obtained from different spectrometers. Filter PLS can select influential wavelength region of spectral data, through Leave-One-Out (LOO) cross validation procedure. The performance of each fitted PLS on each spectrometer data set was measured with root mean square error for prediction (RMSEP), which reveals the PLS-SR (p-value=0.001) and Mp6 (p-value=0.073) select the wavelength region which best explains the variation in starch corn contents.**

***Index Terms*—NIR; spectrometer; PLS; wavelength selection; corn data; starch**

I. Introduction

Chemometrics is associated with loosen up each particular and prognostic issues in preliminary normal sciences, particularly in science. In expressive applications, properties of engineered systems are sculpturesque with the reason for learning the essential associations and structure of the system. In prognostic applications, properties of manufactured systems are sculpturesque with the objective of anticipating new properties or direct of interest. For each circumstance, the datasets are practically nothing at any rate are as often as possible horrendously goliath and extremely tangled, including tons to a large number variables, and tons to an immense number of cases or recognitions.

Strategies are basically seriously utilized in symptomatic science and metabolomics, conjointly the} improvement of improved chemometric techniques for focus similarly continues impelling the bleeding edge in illustrative instrumentation and strategy. It's accomplice application-driven control, and in this way however the quality chemometric frameworks are terribly wide used currently, educational gatherings are focused on the procedure with headway of chemometric speculation, technique and application improvement.

In spectroscopy and chemometrics, It is common practice to extract the chemical information from near infrared (NIR) spectroscopic signals through multivariate modeling. Multivariate modeling unusually address the two main aspects, one is the presence of correlated variables (spectrum wavelengths) are often correlated while other is relatively large number of variables (spectrum wavelength) contrast to the number of samples. Ordinary regression based models are not suitable for this ill posed problem. An alternative to the ordinary regression in such situation is the multivariate approach called Partial Least Squares (PLS) which is widely used one in spectral multivariate calibration because of its simplicity [1].

Multivariate statistics is a subdivision of statistics consolidating the synchronous recognition and examination of more than one outcome variable. The use of multivariate statistics is multivariate examination.

Multivariate statistics concerns understanding the various focuses and establishment of all of the unmistakable sorts of multivariate examination, and how they relate to each other. The helpful utilization of multivariate statistics to a particular issue may incorporate a couple of sorts of univariate and multivariate examinations in order to grasp the associations among variables and their significance to the issue being thought about.

Similarly, multivariate statistics is stressed over multivariate probability movements, to the extent both how these can be used to address the scatterings of watched data; how they can be used as a part of genuine conclusion, particularly where a couple of one of a kind sums are essential to a comparative examination. Explicit sorts of issues including multivariate data, for example clear direct backslide and diverse backslide, are not typically seen as extraordinary cases of multivariate statistics in light of the fact that the examination is overseen by considering the (univariate) unforeseen scattering of a single outcome variable given distinctive components. PLS has focus on the response prediction instead of inevitably on probing or finding the hidden combination or relationship among the variables. It is challenging to make sense out of the high-dimensional data, particularly in the case of multicollinearity. Robust, efficient and scalable analysis of collinear data is the need of the hour as most fields including public health which deal with noisy and massive data [2]. Although, the main focus of PLSR is to determine the subspace of relevant

Manuscript received March 23, 2019; revised June 11, 2019.

doi: 10.18178/ijfe.5.2.132-135

predictor variables and it has no execution of selecting influential variables in its basic algorithm, but several variable selection methods in PLSR have been proposed so far. For improved understanding and interpretation of the model for the target response, variable selection is essential.

5 filter-based variable selection methods were considered including Loading weights [3] regression coefficient [4] variable importance in projection [5], selectivity ratio [6] and Significance Multivariate Correlation [7].

II. Materials and Methods

A. Data Set

Eighty (80) samples of corn were measured from 1100 to 2498 nm at 2 nm intervals on three near-infrared (NIR) spectrometers designated M5, Mp5, and Mp6. Corn starch contents were considered as modeling response [8].

B. Partial Least Square Regression (PLSR)

The relationship between the interaction variable Y and the X network is generally illustrated by direct relapse display;

$$E(Y) = X\beta$$

where the regression coefficient β is a vector which describes the effect of spectrum wavelength variables on the response variable.

The algorithm starts by centering as

$$X_0 = X - 1\acute{\bar{x}}$$

and

$$y_0 = y - 1\bar{y}.$$

Accepting exactly that some A (with A≤p) equals the amount of important parts of the expectation, after Martens & Naes definition, and then for a = 1, 2, ..., A.

Record stacking weights by

$$w_a = \acute{X}_{a-1} y_{a-1}.$$

Normalization of the weight accumulation to be equivalent to 1 of

$$\frac{\acute{w}_a w_a}{w_a}.$$

The scores are denoted by t_a and are defined by

$$t_a = X_{a-1} w_a$$

The X-loading are denoted by p_a and is defined by regressing the variables in X_{a-1} on the score vector:

$$p_a = \acute{X}_{a-1} \frac{t_a}{\acute{t}_a t_a}$$

Same way the Y-loading denoted by q_a and is defined by

$$q_a = \acute{y}_{a-1} \frac{t_a}{\acute{t}_a t_a}.$$

Collapse X_{a-1} and $y_{a-1}$by detracting the contribution of t_a:

$$X_a = X_{a-1} - t_a \acute{p}_a$$

$$y_a = y_{a-1} - t_a q_a$$

Continues with the iteration/components if a<A return to 1.

Allow aggregation weights, scores and uploads recorded in each evolution of the network / vector calculation

$$W = [w_1, w_2, \ldots, w_A], T = [t_1, t_2, \ldots, t_A], P = [p_1, p_2, \ldots, p_A]$$

and

$$Q = [q_1, q_2, \ldots, q_A]$$

At this point, PLSR estimates of relapse coefficients of the direct model are found by:

$$\hat{\beta} = W(\acute{P}W)^{-1}q$$

and

$$\hat{\alpha} = \bar{y} - \bar{x}\hat{\beta}$$

C. Variable Selection in PLSR

5 filter-based wavelength i.e. Variable selection methods were considered including Loading weights (LW), regression coefficient (RC), selectivity ratio (SR), variable importance in projection (VIP) and Significance Multivariate Correlation (SMC). The LW is defined as

$$LW = \frac{w_a}{\max(w_a)}$$

While the RC defined as

$$RC = W(\acute{P}W)^{-1}q.$$

Large absolute value of LW and RC indicates the respective variable is influential in explaining the variation in response that is weight-for-age Z-scores, while a value approaches to zero indicates the respective variable is not influential. VIP defined by is the measure to assemble the importance of each variable based on loading weights. The VIP measure is

$$VIP = \sqrt{p \sum_{a=1}^{A} [SS_a (w_a/\|w_a\|)^2] / \sum_{a=1}^{A} SS_a}$$

where SS_a a denote the sum of squares explained by the ath component and the importance of respective variable is represented. Hence, VIP represents the contribution of respective variable based on variance explained by each component. If VIP is less than a defined threshold then respective variable can be excluded. The SR is the ratio between explained variance (EV) and residual variance (RV) for variable on y target-projected component i.e.

$$SR = EV/RV$$

The variable with SR measure greater than the threshold is included in the model. The SR provides the numerical contribution of each variable included in the model. The higher the value of SR, the more important the variable is for prediction purpose. Lowest SR allows eliminating the corresponding variables without affecting the performance. The basic concept of SMC is to minimize the influence of irrelevant variables in X-structure and enhance the importance of variables which have high contribution related to response variable and is defined as

$$SMC = MS_{Regression} / MS_{residual}$$

where $MS_{Regression}$ is the mean square regression and $MS_{residual}$ is the mean square residual.

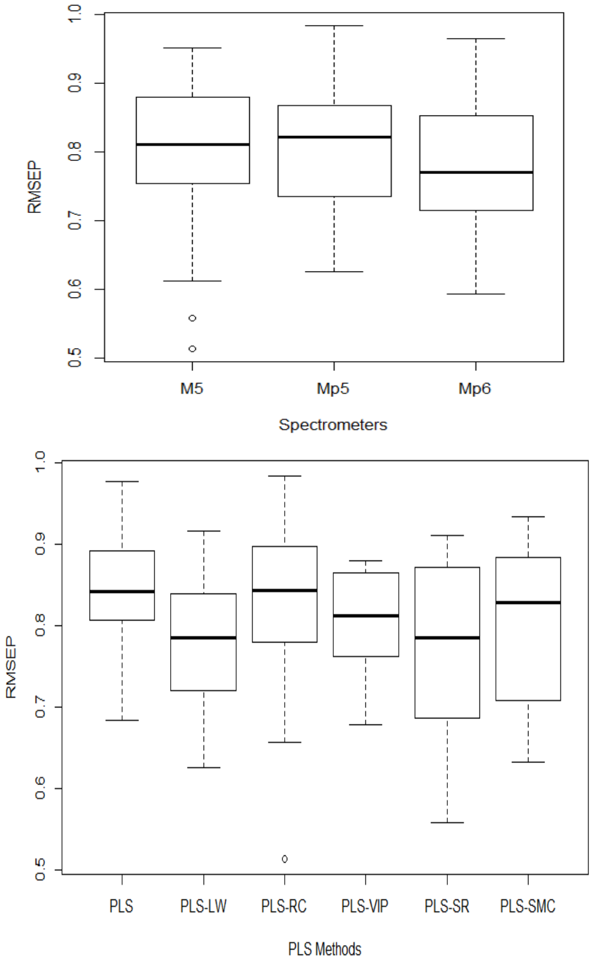

Figure 1. The distribution of RMSEP over the range of spectrometers is presented in upper panel, while the distribution of RMSEP over the range of PLS methods is presented in lower panel.

III. Results and Discussions

In this article corn's starch contents were considered as response variable. The variation in corn's starch contents can be explained through the spectrum obtained from near-infrared (NIR) spectrometers. Some of the wavelength region can be considered as noise which does not explain the variation in corn's starch while some wavelength region can be considered as influential. Hence the influential wavelength selection i.e. variable selection multivariate approaches are of interest.

Since three spectrometers were used, hence corn's starch can be modeled with three spectrum data sets, moreover we have considered 6 PLS based modeling algorithm called partial least squares (PLS), PLS variable selection based on loading weights (PLS-LW), PLS variable selection based on regression coefficients (PLS-RC), PLS variable selection based on variable importance on projection (PLS-VIP), PLS variable selection based on selectivity ratio (PLS-SR) and PLS variable selection based on significant multivariate correlation (PLS-SMC). Except PLS all used version of PLS do variable selection. The use of PLS in its standard form is used as reference method.

TABLE I. ANOVA Results Presenting the Significance of Spectrometers and PLS methods in Defining the Variation in Corn's starch RMSEP.

Factors	Levels	Estimate	Std. Error	P-value
Spectro-meter	M5	Reference		
	Mp5	0.003	0.016	0.861
	Mp6	-0.03	0.017	0.073
Method	PLS	Reference		
	PLS-LW	-0.056	0.024	0.018
	PLS-RC	-0.028	0.024	0.235
	PLS-VIP	-0.032	0.023	0.169
	PLS-SR	-0.075	0.023	0.001
	PLS-SMC	-0.056	0.024	0.019

In all PLS related methods two parameters are used in model fitting, one is the number of PLS components and other is the threshold on respective variable selection index for influential wavelength selection. A range of possible values of these parameters are considered for instance A (number of components) = (1,2,3,…,10) , RC= LW= (0.1, 0.2,…,1), VIP= (0.7, 0.8, …, 1.3) ,SR= (0.5, 1, 1.5,…, 5) and SMC= (1,2,3,…, 10). The optimum value of these threshold was chosen through a double leave one out cross validation, where in fist loop of double cross validation optimum number of PLS components were estimated while in second loop of double cross validation optimum threshold was derived.

In order to have reliable comparison of fitted models over the NIR spectrum data, Monte-Carlo simulation with 10 runs was used and in each run root mean square error on prediction was (RMSEP) together with number of influential variables were recorded. The distribution of RMSEP over the range of spectrometers is presented in upper panel, while the distribution of RMSEP over the range of PLS methods is presented in lower panel of Fig. 1. More over the analysis of variance was conducted to attach the statistical significance with our findings and ANOVA results are presented in Table I.

Since there spectrometers are used M5, Mp5 and Mp6, we have considered M5 as reference spectrometer, it appears from Fig. 1 and Table I the prediction capability of corn's starch of Mp5 is not much different from M5 spectrometer, while the prediction capability of Mp6 is significantly better than M5 spectrometer (p-value= 0.073). Among the prediction or modeling methods standard PLS was considered as reference method. It appears all variable selection methods has improved prediction capability compared to standard PLS. PLS-SR has the best prediction capability in terms of RMSEP (p-value= 0.001). Similarly PLS-SMC (p-value= 0.019) and PLS-LW (p-value= 0.018) are also appeared significant.

It appears from Fig. 2 and Table II the number of influential variables i.e. influential wavelength region obtained from the Mp5 is significantly smaller than from M5 spectrometer (p-value= 0.042), while number of influential variables i.e. influential wavelength region obtained from the Mp6 is similar to M5 spectrometer. It appears all PLS variable selection methods has significantly smaller number of variables being selected compared to standard PLS (p-value=< 0.001).

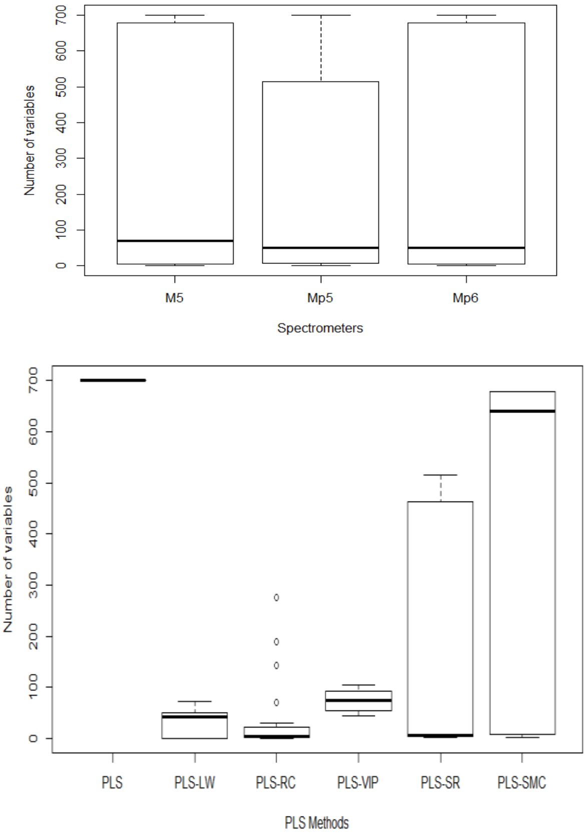

Figure 2. The distribution of number of variables over the range of spectrometers is presented in upper panel, while the distribution of number of variables over the range of PLS methods are presented in lower panel.

TABLE II. ANOVA RESULTS PRESENTING THE SIGNIFICANCE OF SPECTROMETERS AND PLS METHODS IN DEFINING THE VARIATION IN CORN'S STARCH NUMBER OF VARIABLES

Factors	Levels	Estimate	Std. Error	P-value
Spectro-meter	M5	Reference		
	Mp5	-60.017	29.274	0.042
	Mp6	-44.417	29.274	0.131
Method	PLS	Reference		
	PLS-LW	-668.767	41.4	<0.001
	PLS-RC	-671.6	41.4	<0.001
	PLS-VIP	-626.2	41.4	<0.001
	PLS-SR	-545.9	41.4	<0.001
	PLS-SMC	-312	41.4	<0.001

IV. CONCLUSION

Based on two model performance which are number of wavelength being selected i.e. number of selected variables and prediction capability on test data set which was measured through the predicted root mean square error. The meta analysis reveals that among the filter wavelength region selection algorithm, including PLS-LW, PLS-SR and PLS-SMC are significantly modeling the starch contents of corn with corn spectral data.

Moreover compared to M5 Mp6 best explain the variations in corn's starch contents and Mp5 results in least number of influential wavelength selection. Hence the study provide the guideline not only for wavelength selection in modeling the corn's starch but also provide the guideline over the choice of NIR spectrometer.

ACKNOWLEDGMENT

The author wish to thank National University for Sciences and Technology (NUST), Islamabad, Pakistan, for ICFEB 2019 participation, presentation and publication in this journal.

REFERENCES

[1] S. Wold, M. Sjöström, and L. Eriksson, "Partial least squares projections to latent structures (PLS) in chemistry," *Encyclopedia of Computational Chemistry*, 2002.

[2] T. Mehmood, H. Martens, S. Sæbø, J. Warringer, and L. Snipen, "A Partial Least Squares based algorithm for parsimonious variable selection," *Algorithms for Molecular Biology*, vol. 6, no. 1, p. 27, 2011.

[3] I. S. Helland, "Some theoretical aspects of partial least squares regression," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 2, pp. 97-107, 2001.

[4] A. Höskuldsson, "Variable and subset selection in PLS regression," *Chemometrics and Intelligent Laboratory Systems*, vol. 55, no. 1-2, pp. 23-38, 2001.

[5] M. Farrés, S. Platikanov, S. Tsakovski, and R. Tauler, "Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation," *Journal of Chemometrics*, vol. 29, no. 10, pp. 528-536, 2015.

[6] T. Rajalahti, R. Arneberg, F. S. Berven, K. M. Myhr, R. J. Ulvik, and O. M. Kvalheim, "Biomarker discovery in mass spectral profiles by means of selectivity ratio plot," *Chemometrics and Intelligent Laboratory Systems*, vol. 95, no. 1, pp. 35-48, 2009.

[7] T. N. Tran, N. L. Afanador, L. M. Buydens, and L. Blanchet, "Interpretation of variable importance in partial least squares with significance multivariate correlation (sMC)," *Chemometrics and Intelligent Laboratory Systems*, vol. 138, pp. 153-160, 2014.

[8] Y. Liu, W. Cai, and X. Shao, "Standardization of near infrared spectra measured on multi-instrument," *Analytica Chimica Acta*, vol. 836, pp. 18-23, 2014.

Tahir Mehmood did his PhD in Statistics from Norwegian University for Life Science (NMBU) Norway in 2012. Currently working as Associate Professor in School of Natural Science (SNS), National University for Sciences and Technology (NUST), Islamabad, Pakistan. His interest lies in Multivariate Statistics, Chemometrics, Enviormetrics and in related areas.