BIOSYSTEMS ENGINEERING 203 (2021) 93–97

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/issn/15375110

Research Note

Improved prediction of protein content in wheat kernels with a fusion of scatter correction methods in NIR data modelling

Puneet Mishra [a,*], *Santosh Lohumi* [b]

[a] *Wageningen Food and Biobased Research, Bornse Weilanden 9, P.O. Box 17, 6700AA, Wageningen, the Netherlands*
[b] *Department of Biosystems Machinery Engineering, College of Agricultural and Life Science, Chungnam National University, Yuseong-gu, Daejeon, 34134, South Korea*

ARTICLE INFO

Article history:
Received 25 August 2020
Received in revised form
19 December 2020
Accepted 7 January 2021
Published online 22 January 2021

Keywords:
Multiblock
Chemometrics
Fusion
Complementary

The study aims to test the hypothesis that modelling of near-infrared (NIR) spectroscopic data based on a single scatter correction technique is sub-optimal. Better predictive performance of the multivariate analysis method can be obtained when the information from differently scatter corrected data is jointly used. To demonstrate it, an open-source NIR spectroscopy data set related to protein prediction in wheat kernels was used. Two different pre-processing fusion approaches i.e., sequential and parallel fusion, were used for fusing the complementary information from four different scatter correction techniques, namely standard normal variate (SNV), variable sorting for normalisation (VSN), 2nd derivative, and multiplicative scatter correction (MSC). As a comparison, partial least-squares regression (PLSR) was performed on the SNV pre-processed data. The results showed that fusion of scatter correction can improve the predictive performance of NIR spectroscopic models. The results revealed that both sequential and parallel fusion approaches improved the predictive performance compared to the PLSR performed using a single scatter correction technique. The R^2_p was improved by up to 3% and the RMSEP was reduced by up to 13% compared to the results obtained with conventional PLSR model developed with a single scatter correction technique.

1. Introduction

Near-infrared (NIR) spectroscopy of wheat kernels is widely performed for rapid estimation of key quality attributes (Caporaso, Whitworth, & Fisk, 2018; Salgó & Gergely, 2012). Attributes such as protein, fat and moisture can be predicted with high accuracies as the electro-magnetic radiation in NIR region (780–2500 nm) is associated with overtones of the fundamental bond vibrations such as O–H, C–H, and N–H (Caporaso et al., 2018). NIR spectroscopy alone is of no use unless it is combined with chemometric analysis to develop predictive models.

A major challenge with the NIR spectroscopy is that the signal recorded by the instrument contains mixed

* *Corresponding author.*
E-mail address: puneet.mishra@wur.nl (P. Mishra).
https://doi.org/10.1016/j.biosystemseng.2021.01.003

information i.e., absorption and light scattering characteristics (Pasquini, 2018). The absorption information is related to the chemical components present in the samples, whereas the scatter information is related to the complex interaction of the light with the physical structure of the samples (Lu, Van Beers, Saeys, Li, & Cen, 2020). In NIR data modelling, to develop models for predicting chemical components, the aim is always to first remove the scatter information from the data (Isaksson & Næs, 1988). This can be achieved with application of several scatter correction techniques available in the chemometrics domain (Roger, Boulet, Zeaiter, & Rutledge, 2020). To select the best scatter correction technique, several models corresponding to different correction techniques are explored and the one with the best predictive performance is selected (Torniainen et al., 2020). However, this approach has two main drawbacks; first it is time-consuming and computationally expensive to explore all potential scatter correction techniques. The second is that selecting and using the single scatter correction technique is sub-optimal as data pre-processed with different scatter correction techniques carry complementary information (Mishra, Roger, Rutledge, & Woltering, 2020). Recently, Mishra et al., (2020) demonstrated that instead of finding the best pre-processing techniques, the user should explore the complementary information present in data pre-processed with different techniques as it can drastically improve the predictive performance of NIR models (Mishra, Biancolillo, Roger, Marini, & Rutledge, 2020; Mishra, Marini, Biancolillo, & Roger, 2020; Mishra, Nordon, & Roger, 2020; Mishra, Roger, Rutledge, et al., 2020).

This study aims to demonstrate that modelling of NIR spectroscopic data based on the use of a single scatter correction technique is sub-optimal. Better predictive performance can be obtained when the information from differently scatter corrected data is jointly used. To demonstrate it, an open-source NIR spectroscopy data set related to protein prediction in wheat kernels was used. Two different pre-processing fusion approaches i.e., sequential and parallel, were used for fusing the complementary information from four different scatter correction techniques namely, standard normal variate (SNV), variable sorting for normalisation (VSN), 2nd derivative, and multiplicative scatter correction (MSC). To compare the performance of data fusion and conventional (single) scatter correction technique, partial least-squares regression (PLSR) was performed individually pre-processed data.

2. Materials and methods

2.1. Data set

The wheat kernel data set used in this study was obtained from the Mendeley repository of open data sets (Wenya, 2016). The data set can also be accessed at https://figshare.com/articles/wheat_kernel_dataset/4252217/1. The data set contains NIR spectra and reference protein concentration of 523 wheat kernels. The spectra were measured in the spectral range of 850–1050 nm with a total of 100 wavelengths (nm). In this analysis, the data set was divided into calibration (60%) and test set (40%) using the Kennard-Stone (KS) algorithm (Kennard & Stone, 1969).

2.2. Scatter correction techniques

This study uses four most common scatter correction techniques i.e., 2nd derivative (Roger, Boulet, et al., 2020), variable sorting for normalisation (VSN) (Rabatel, Marini, Walczak, & Roger, 2020), standard normal variate (SNV) (Barnes, Dhanoa, & Lister, 1989) and multiplicative scatter correction (MSC) (Isaksson & Næs, 1988), for pre-processing fusion. The second derivative estimation was performed using the Savitzky–Golay (2nd order polynomial + 21-point window). All the pre-processing techniques were implemented as discussed in (Roger, Boulet, et al., 2020) and executed using MATLAB 2018b (Natick, MA, USA).

2.3. Scatter correction fusion with sequential and parallel approaches

The sequential approach called SPORT (Roger, Biancolillo, & Marini, 2020) is based on the sequential orthogonalised partial least-squares regression and the PORTO (Mishra, Roger, Marini, Biancolillo, & Rutledge, 2020) approach on parallel orthogonalised partial least-squares regression. A schematic of the PORTO and SPORT approach is presented in Fig. 1. In the case of PORTO, a combination of PLS regression, generalised canonical analysis (GCA) and multiple orthogonalisation steps are performed to extract the common and distinct information presented in data pre-processed with differently scatter corrected data. The concept of PORTO is shown in Fig. 1a. The three circles represent three differently scatter corrected data and the letters D and C indicate the distinct and the common information. In SPORT, incremental learning is performed using a combination of PLSR and sequential orthogonalisation as presented in Fig. 1b. In PORTO, several local CVs were performed in sequence, as discussed in (Måge, Menichelli, & Næs, 2012). In the case of SPORT, all possible combinations of LVs were explored and the lowest cross-validation (CV) error was to choose the optimal combination of LVs from differently scatter corrected data. SPORT was implemented with the algorithm presented in (Roger, Biancolillo, et al., 2020). The PORTO was implemented using the multi-block data analysis codes from NOFIMA (https://nofima.no/en/) for the implementation of parallel orthogonalised partial least-square. All analysis was performed in MATLAB 2017b (The MathWorks, Natick, USA).

2.4. Partial least-squares regression

In comparison to the pre-processing fusion approaches, standard PLSR based on a single scatter correction technique was used. The PLSR was performed on the standard normal variate (SNV) estimated data. SNV is the most widely used scatter correction technique as it model-free and fast to implement (Roger, Boulet, et al., 2020). PLSR works by maximising the covariance of the NIR spectral data with the response variables to identify the subspaces of latent variables (LVs) on which high-dimensional data can be transformed into a low dimension information concentrated space.

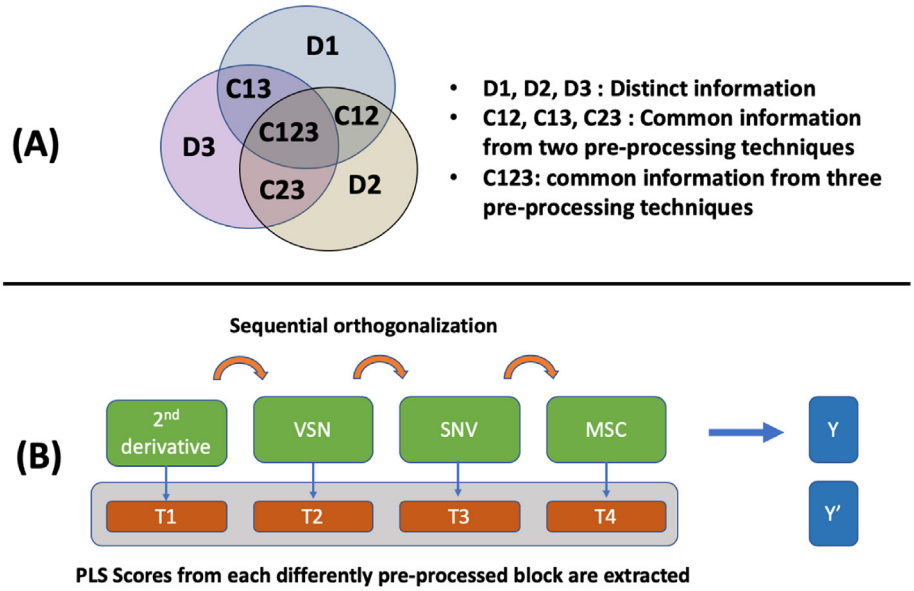

Fig. 1 – Schematic of parallel (A) and sequential (B) pre-processing fusion approaches. T1-T4 are the scores from sequential partial least square regression.

For more information regarding the PLSR, please refer to the (Geladi & Kowalski, 1986); here it is omitted for brevity. In this study, PLSR was implemented with the MATLAB's built-in function 'plsregress', combined with a 10-fold cross-validation procedure approach to select the optimal number of latent variables (LVs).

3. Results

3.1. *PLSR versus pre-processing fusion modelling*

The results of testing the PLSR model based on the SNV corrected data and from the fusion of four different scatter correction techniques are shown in Fig. 2. Both the sequential and parallel scatter correction fusion approaches improved the model performance compared to the standard PLSR performed using a single scatter correction technique (Table 1). With the sequential approach, the R^2_p was increased by 2% and the root means squared error of prediction (RMSEP) was decreased by 11%. Further, the sequential approach extracted the latent variables from two scatter correction techniques (2 LVs from 2nd derivative and 7 LVs from VSN), highlighting that two scatter

Table 1 – A summary of PLSR prediction results obtained with individual pre-processing techniques and with pre-processing fusion techniques. Standard normal variate (SNV), variable sorting for normalization (VSN), multiplicative scatter correction (MSC), 2nd derivative, sequential pre-processing through orthogonalization (SPORT) and parallel pre-processing through orthogonalization (PORTO).

Technique	Latent variables	R^2_P	RMSEP (%)	Prediction bias (%)
SNV	8	0.87	0.54	−0.01
VSN	8	0.85	0.52	0.04
MSC	8	0.84	0.55	−0.04
2nd derivative	5	0.81	0.59	0.05
SPORT	9	0.89	0.48	−0.01
PORTO	6	0.90	0.47	0.03

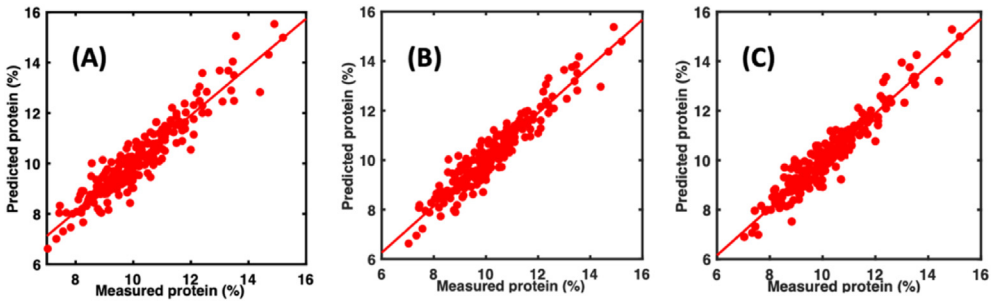

Fig. 2 – A summary of (A) partial least-squares regression modelling with 8 latent variables, (B) sequential pre-processing fusion with 2 latent variables from 2nd derivative and 7 latent variables from variable sorting for normalisation, and (C) parallel pre-processing fusion with 4 common components and 2 distinct components.

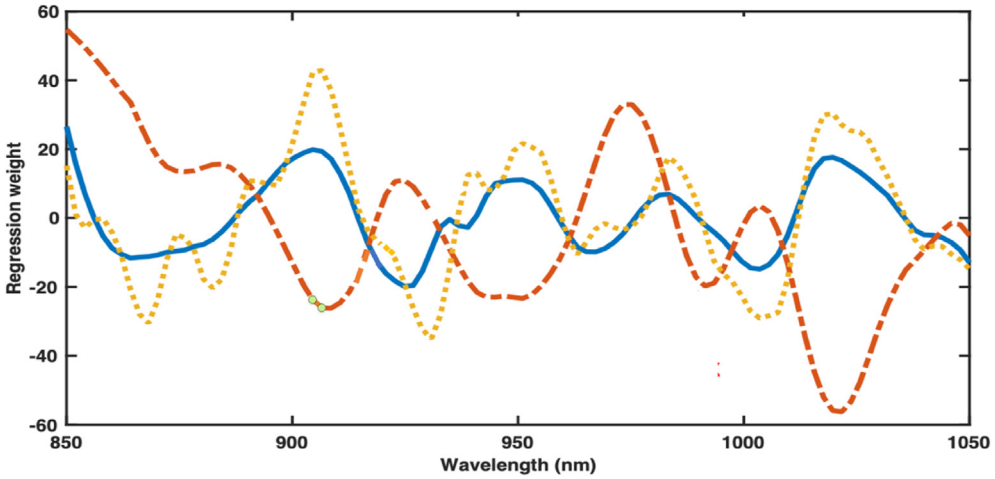

Fig. 3 – Regression vectors from standard partial least squares regression performed on standard normal variate corrected data (solid blue) and from the sequential pre-processing fusion (dashed red and dotted yellow). The sequential pre-processing fusion (SPORT) extracted complementary information from 2nd derivative (dashed red) and variable sorting for normalisation (dotted yellow), therefore, two regression vectors are presented for sequential approach (SPORT). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

correction techniques jointly improved the model performance. The parallel approach performed slightly better compared to the sequential approach. In comparison to the standard PLSR performed using single scatter correction technique i.e. SNV, the R^2_p was increased by 3% and the RMSEP was reduced by 13%. A further comparison of all four individual pre-processing techniques and the two-pre-processing fusion-based techniques is shown in Table 1. It can be noticed that pre-processing fusion techniques i.e. SPORT and PORTO, performed better than all individual techniques.

3.2. *An example of complementary information captured by pre-processing fusion modelling*

To demonstrate how the pre-processing fusion approach improved the model performance, an example related to the information modelled by sequential pre-processing fusion approach is shown in Fig. 3. It can be noticed that PLSR modelling performed on the data pre-processed with only SNV lacks several peaks (highlight with the pointed arrows in Fig. 3) related to C–H and O–H overtones which are abundantly present in the amino acids (Osborne, 2006). These peaks were better captured and resolved by a fusion of information from data pre-processed with multiple scatter correction techniques. In summary, multiple scatter correction techniques complement each other and allows capturing the information which is missed while modelling one scatter correction technique. These results agreed with several recent works, where fusion of different pre-processing with sequential (Mishra, Roger, Rutledge, et al., 2020) and parallel (Mishra, Roger, Marini, et al., 2020) approaches revealed underlying peaks which were missed by PLS regression analysis based on single scatter correction techniques.

4. Conclusion

The results from the fusion of information from data pre-processed with four scatter correction techniques showed improvement in the model performance in comparison to standard PLSR performed using only a single scatter correction technique. The R^2_p was increased by up to 3% and the RMSEP was decreased by up to 13% for the prediction of protein in individual wheat kernels. Such a decrease in prediction error will allow accurate and precise prediction of wheat kernel properties, thus, complementing the whole supply chain. The sequential and parallel pre-processing fusion approaches as presented in this work are not limited to NIR data but can be used to pre-processing any kind of spectral data which require a pre-processing selection step.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

Barnes, R. J., Dhanoa, M. S., & Lister, S. J. (1989). Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy*, 43(5), 772–777. https://doi.org/10.1366/0003702894202201

Caporaso, N., Whitworth, M. B., & Fisk, I. D. (2018). Near-Infrared spectroscopy and hyperspectral imaging for non-destructive quality assessment of cereal grains. *Applied Spectroscopy Reviews*, 53(8), 667–687. https://doi.org/10.1080/05704928.2018.1425214

Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: A tutorial. *Analytica Chimica Acta*, 185, 1–17. https://doi.org/10.1016/0003-2670(86)80028-9

Isaksson, T., & Næs, T. (1988). The effect of multiplicative scatter correction (MSC) and linearity improvement in NIR spectroscopy. *Applied Spectroscopy*, 42(7), 1273–1284.

Kennard, R. W., & Stone, L. A. (1969). Computer aided design of experiments. *Technometrics*, 11(1), 137–148. https://doi.org/10.1080/00401706.1969.10490666

Lu, R. F., Van Beers, R., Saeys, W., Li, C. Y., & Cen, H. Y. (2020). Measurement of optical properties of fruits and vegetables: A review. *Postharvest Biology and Technology*, 159. https://doi.org/10.1016/j.postharvbio.2019.111003

Måge, I., Menichelli, E., & Næs, T. (2012). Preference mapping by PO-PLS: Separating common and unique information in several data blocks. *Food Quality and Preference*, 24(1), 8–16. https://doi.org/10.1016/j.foodqual.2011.08.003

Mishra, P., Biancolillo, A., Roger, J. M., Marini, F., & Rutledge, D. N. (2020a). New data preprocessing trends based on ensemble of multiple preprocessing techniques. *TRAC Trends in Analytical Chemistry*, 116045. https://doi.org/10.1016/j.trac.2020.116045

Mishra, P., Marini, F., Biancolillo, A., & Roger, J.-M. (2020b). Improved prediction of fuel properties with near-infrared spectroscopy using a complementary sequential fusion of scatter correction techniques. *Talanta*, 121693. https://doi.org/10.1016/j.talanta.2020.121693

Mishra, P., Nordon, A., & Roger, J.-M. (2020c). Improved prediction of tablet properties with near-infrared spectroscopy by a fusion of scatter correction techniques. *Journal of Pharmaceutical and Biomedical Analysis*, 113684. https://doi.org/10.1016/j.jpba.2020.113684

Mishra, P., Roger, J. M., Marini, F., Biancolillo, A., & Rutledge, D. N. (2020d). Parallel pre-processing through orthogonalization (PORTO) and its application to near-infrared spectroscopy. *Chemometrics and Intelligent Laboratory Systems*, 104190. https://doi.org/10.1016/j.chemolab.2020.104190

Mishra, P., Roger, J. M., Rutledge, D. N., & Woltering, E. (2020e). SPORT pre-processing can improve near-infrared quality prediction models for fresh fruits and agro-materials. *Postharvest Biology and Technology*, 168, 111271. https://doi.org/10.1016/j.postharvbio.2020.111271

Osborne, B. G. (2006). Near-infrared spectroscopy in food analysis. In *Encyclopedia of analytical chemistry*.

Pasquini, C. (2018). Near infrared spectroscopy: A mature analytical technique with new perspectives – a review. *Analytica Chimica Acta*, 1026, 8–36. https://doi.org/10.1016/j.aca.2018.04.004

Rabatel, G., Marini, F., Walczak, B., & Roger, J.-M. (2020). VSN: Variable sorting for normalization. *Journal of Chemometrics*, 34(2), Article e3164. https://doi.org/10.1002/cem.3164

Roger, J.-M., Biancolillo, A., & Marini, F. (2020). Sequential preprocessing through ORThogonalization (SPORT) and its application to near infrared spectroscopy. *Chemometrics and Intelligent Laboratory Systems*, 199, 103975. https://doi.org/10.1016/j.chemolab.2020.103975

Roger, J.-M., Boulet, J.-C., Zeaiter, M., & Rutledge, D. N. (2020a). Pre-processing Methods☆. In *Reference module in chemistry, molecular sciences and chemical engineering*. Elsevier.

Salgó, A., & Gergely, S. (2012). Analysis of wheat grain development using NIR spectroscopy. *Journal of Cereal Science*, 56(1), 31–38. https://doi.org/10.1016/j.jcs.2012.04.011

Torniainen, J., Afara, I. O., Prakash, M., Sarin, J. K., Stenroth, L., & Toyras, J. (2020). Open-source python module for automated preprocessing of near infrared spectroscopic data. *Analytica Chimica Acta*, 1108, 1–9. https://doi.org/10.1016/j.aca.2020.02.030

Wenya, L. (2016). *Wheat kernel dataset: Figshare*. Retrieved from https://figshare.com/articles/wheat_kernel_dataset/4252217/1.