Infrared Physics & Technology 113 (2021) 103643

Contents lists available at ScienceDirect

Infrared Physics and Technology

journal homepage: www.elsevier.com/locate/infrared

Improved prediction of minced pork meat chemical properties with near-infrared spectroscopy by a fusion of scatter-correction techniques

Puneet Mishra [a,*], Theo Verkleij [a], Ronald Klont [b]

[a] Wageningen Food and Biobased Research, Bornse Weilanden 9, P.O. Box 17, 6700AA Wageningen, the Netherlands
[b] Vion Food, Boseind 15, 5280 AA Boxtel, the Netherlands

ARTICLE INFO

Keywords:
Multiblock
Fusion
Spectroscopy
Preprocessing
Multivariate

ABSTRACT

The modelling near-infrared (NIR) spectroscopy data requires removal of scattering effects from the data before applying advanced chemometrics methods. Often different scatter-correction techniques are explored, and the scatter-correction technique with the best performance is selected. However, the information highlighted by different scatter-correction techniques may be complementary and their fusion may result in better models for predicting characteristics, such as meat quality. To test this, sequential and parallel preprocessing fusion approaches will be used in this work to fuse information from different scatter-correction techniques to try to improve the predictive performance of NIR models. Three different chemical properties, i.e., moisture, fat and protein content, were predicted. For comparison, partial least-squares regression (PLSR) was performed on standard normal variate (SNV) corrected data, as this is a widely used scatter-correction technique. Compared to this commonly used procedure, the scattering fusion approaches reduced the error and bias by up to 52% and 84%, respectively. The results suggest that fusion of scatter-correction techniques is essential to achieve optimal NIR prediction models for predicting meat characteristics such as moisture, fat and protein content.

1. Introduction

In recent years, near-infrared (NIR) spectroscopy has emerged as a key non-destructive technique for rapid and cost-effective estimation of meat properties [1,2]. Applications of NIR spectroscopy for meat range from prediction of chemical [3,4], sensory and textural properties [5] to authenticity of meat products [6–8].

The interaction of NIR light with the meat product is highly complex and the signal recorded after the interaction has two main components i.e., absorption and scattering [1]. The absorption is mainly related to the chemical components present in the meat whereas the scattering may result from the physical structure of the meat [9]. Often for efficient prediction of chemical components, scattering effects are removed from the data [10] as they may mask the underlying spectral signal corresponding to the chemical components [11]. The masking can be observed as additive and multiplicative effects over the whole spectral range. Several scatter-correction methods are available to remove these effects [12], such as 2nd derivative which removes first-order additive effects (baseline shift) and also reveals underlying peaks [13]. Standard normal variate (SNV) removes additive and multiplicative effects by treating each spectrum by subtraction of its mean spectral intensity from each intensity response and then division by the standard deviation of the spectrum [14]. Multiplicative scatter-correction (MSC) assumes that each spectrum consists of a multiplicative, an additive and a residual part [15]. In summary, there are many pre-processing methods available in chemometrics to remove or reduce the scattering effects in NIR spectra [10].

In chemometrics, often the best scatter-correction technique is selected [16]. However, selecting and using a single scatter-correction may lead to sub-optimal modelling as the data preprocessed with different scatter-correction techniques may carry complementary information [10,17]. Recently, Mishra et al., (2020) [10] showed that a fusion of scatter-correction techniques is a better solution as the information highlighted by differently corrected data is complementary.

This study aims to demonstrate that a fusion of information from different scatter-correction techniques can improve NIR models for predicting meat properties, compared to the NIR models using a single scatter-correction technique. To perform the fusion, the sequential preprocessing through orthogonalization (SPORT) [18] and the parallel preprocessing through orthogonalization (PORTO) approaches were used. Three different chemical properties (moisture, fat and protein) were used for the predictive analysis. As a comparison, standard PLSR

[*] Corresponding author.
E-mail address: Puneet.mishra@wur.nl (P. Mishra).

https://doi.org/10.1016/j.infrared.2021.103643
Received 24 November 2020; Received in revised form 9 January 2021; Accepted 12 January 2021
Available online 20 January 2021

modelling was performed on SNV normalized data.

2. Materials and methods

2.1. Data set

The NIR absorbance data set acquired on meat (minced pork) used in the study is the open-source data set available on the website of the Carnegie Mellon University http://lib.stat.cmu.edu/datasets/tecator. This data set was chosen as it is freely available and covers most common meat properties such as moisture, protein and fat content. The spectra from 240 samples covered the range of 850–1050 nm with a total of 100 channels. The NIR data were recorded in transmission mode on a Tecator Infratec Food and Feed Analyzer and converted to absorbance by calculating the -log10 of the transmittance. Each sample contains finely chopped pure meat with different moisture, fat and protein contents. The moisture, fat and protein, measured in per cent, were determined by oven drying, the Soxhlet and the Kjeldahl methods [19]. The samples were further portioned into calibration (60%) and test (40%) set using the Kennard-Stone algorithm [20].

2.2. Data analysis

2.2.1. Scatter-correction methods

In the present work, four of the most commonly used scatter-correction techniques were selected. Two techniques were model-based (variable sorting for normalization and multiplicative scatter-correction) and two were model-free (standard normal variate and 2nd derivative). Multiplicative scatter-correction (MSC) models each spectrum as a mixture of scattering and absorbance [15]. The MSC was implemented using the mean as the reference. In standard normal variate (SNV) [14] for each spectrum, the offset correction is done by subtracting the mean intensity while the multiplicative effect is then reduced by dividing each intensity by the standard deviation of the centered spectrum. Variables sorting for normalization (VSN) [21] assumes that not all the bands are equally altered by the unwanted effects and, consequently, assigns to each variable a weight in the range [0,1] corresponding to its probability of it being affected only by scattering. VSN estimates these weights based on the random consensus (RANSAC) algorithm which estimates the extent to which a wavelength is affected by size effects (additive and multiplicative offsets) or by shape effects (chemical related features). In this way, variables that are strongly related to chemical components have a low weight and negligible role in the calculation of the size effect. The main benefit of the VSN approach in comparison to MSC is that it does not require a reference spectrum to perform the weight estimation. In the present work, the weights estimated by VSN were integrated into SNV leading to a weighted SNV. The fourth method was the calculation of 2nd derivative, which is commonly used to remove both additive and multiplicative effects [13]. Numerical differentiation, i.e., calculation of the second derivative, was performed using the Savitzky-Golay approach (2nd order polynomial with a 21-point window). All the pre-processing methods were implemented using the MBA-GUI [22] under MATLAB 2018b (The Mathworks, Natick, MA, USA). All the models were evaluated based on the coefficient of determination ($R^2{}_P$), root mean squared error of prediction (RMSEP) and prediction bias.

2.2.2. Partial least-squares regression

Partial least-squares regression (PLSR) is a common chemometric method [23] widely used for NIR data modelling [24]. PLSR projects the NIR data onto a subspace of latent variables (LVs) which have maximum covariance with the response(s). The transformed data are relevant for predicting the response variables. In this study, PLSR was calculated by means of the MATLAB's built-in function 'plsregress', combined with a 10-fold cross-validation procedure to select the optimal number of latent variables (LVs).

2.2.3. Sequential preprocessing through orthogonalization

SPORT is a two steps process involving a PLSR followed by orthogonalization [18]. A schematic of SPORT approach is presented in Fig. 1a. The SPORT algorithm for two pre-processing blocks ($\mathbf{X_1}$ and $\mathbf{X_2}$) is as follows:

1. The **Y** responses are fitted to the $\mathbf{X_1}$ by the PLS regression
2. $\mathbf{X_2}$ is orthogonalized with respect to the scores obtained from the first regression
3. The orthogonalized $\mathbf{X_2}$ is used to predict the **Y** residuals
4. The overall predictive model is obtained by combining the sub-models calculated in steps 1 and 3

The procedure is continued for as many blocks (4 in this case) as there are pretreatments. In this work, the scatter-correction order was 2nd derivative, VSN, SNV and MSC, making a total of four blocks of data. The number of LVs is optimized exploring all possible combinations of LVs and the optimal one is the model resulting in the lowest RMSECV.

2.2.4. Parallel pre-processing through orthogonalization

PORTO is a combination of PLSR, generalized canonical analysis (GCA) and multiple orthogonalization steps. PORTO aims to extract common and distinct information within the differently scatter-corrected data to improve data modelling. The concept of PORTO to identify common and distinct information is shown in Fig. 1b. The three circles represent three differently scatter-corrected data and the letters D and C indicate the distinct and the common information. The algorithm for PORTO is similar to parallel partial least-squares regression and is as follows:

1. Standard PLS models are calculated between the **Y** and each of the differently pre-processed blocks $\mathbf{X_p}$ ($p = 1,\ldots,P$), leading to as many scores matrices $\mathbf{T_p}$ from each model.
2. GCA is performed on all possible subsets of blocks to identify global and local common components ($\mathbf{T_{Ck}}$) as those linear combinations of the block scores with high correlation.
3. The scores $\mathbf{T_p}$ are orthogonalized with respect to $\mathbf{T_{Ck}}$ to obtain $\mathbf{T_{po}}$
4. Step 2 and 3 are repeated for all relevant blocks by using the $\mathbf{T_{po}}$ as the input to the GCA in step 2
5. For each block, a PLS regression model is calculated between **Y** and the orthogonalized scores of each of the blocks, leading to distinct scores for each differently pre-processed block, i.e., $\mathbf{T_{Up}}$
6. The final model is built by running an ordinary least square regression between the concatenated scores matrix and the **Y**, leading to regression coefficients, **Beta.**

To optimize the number of LVs for each block, several local cross-validations (CV) are performed in sequence, as discussed in [25]. The PORTO was implemented in MATLAB 2017b using the multi-block data analysis codes from NOFIMA (https://nofima.no/en/) for the implementation of parallel orthogonalized partial lest-squares.

3. Results

3.1. PLSR modelling vs SPORT vs PORTO

The results from PLSR, SPORT and PORTO modelling are shown in Fig. 2. For all the properties, the preprocessing fusion approaches outperformed the traditional PLSR modelling done using a single scatter-correction technique. A summary of the improvements in RMSEP attained with the SPORT and PORTO approach compared to the PLS regression analysis are shown in Table 1. It can be noted that for all three properties the SPORT and PORTO outperformed the PLS regression analysis performed on signal scatter correction technique. Further, the improvements were much better with the PORTO compared to the SPORT. For example, for protein prediction, the PORTO approach

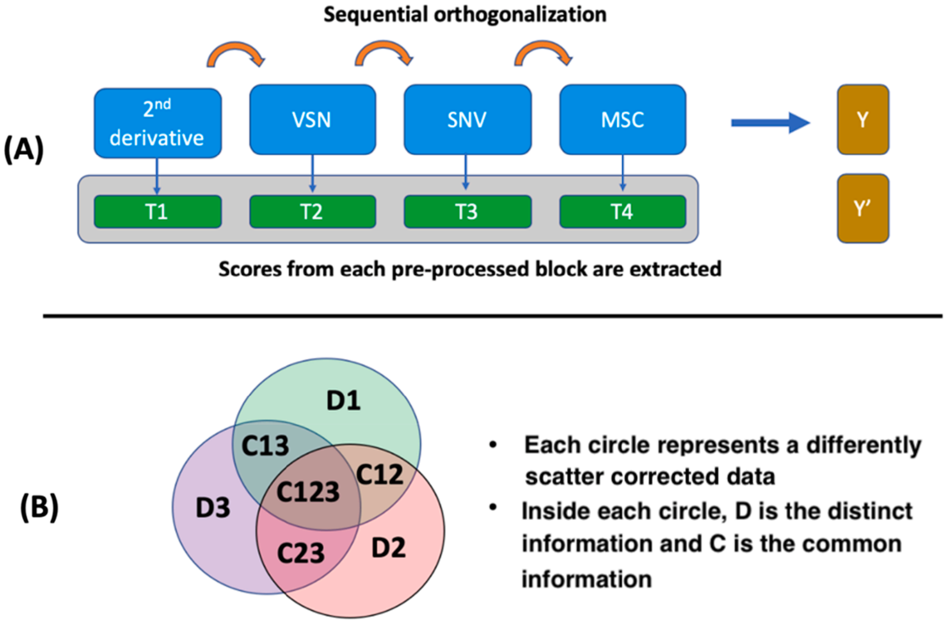
Sequential orthogonalization
(A)
2nd derivative
VSN
SNV
MSC
Y
Y'
T1
T2
T3
T4
Scores from each pre-processed block are extracted
(B)
D1
D2
D3
C13
C12
C23
C123
• Each circle represents a differently scatter corrected data
• Inside each circle, D is the distinct information and C is the common information

Fig. 1. A schematic of the sequential (SPORT) (A) and parallel (PORTO) (B) preprocessing fusion approaches.

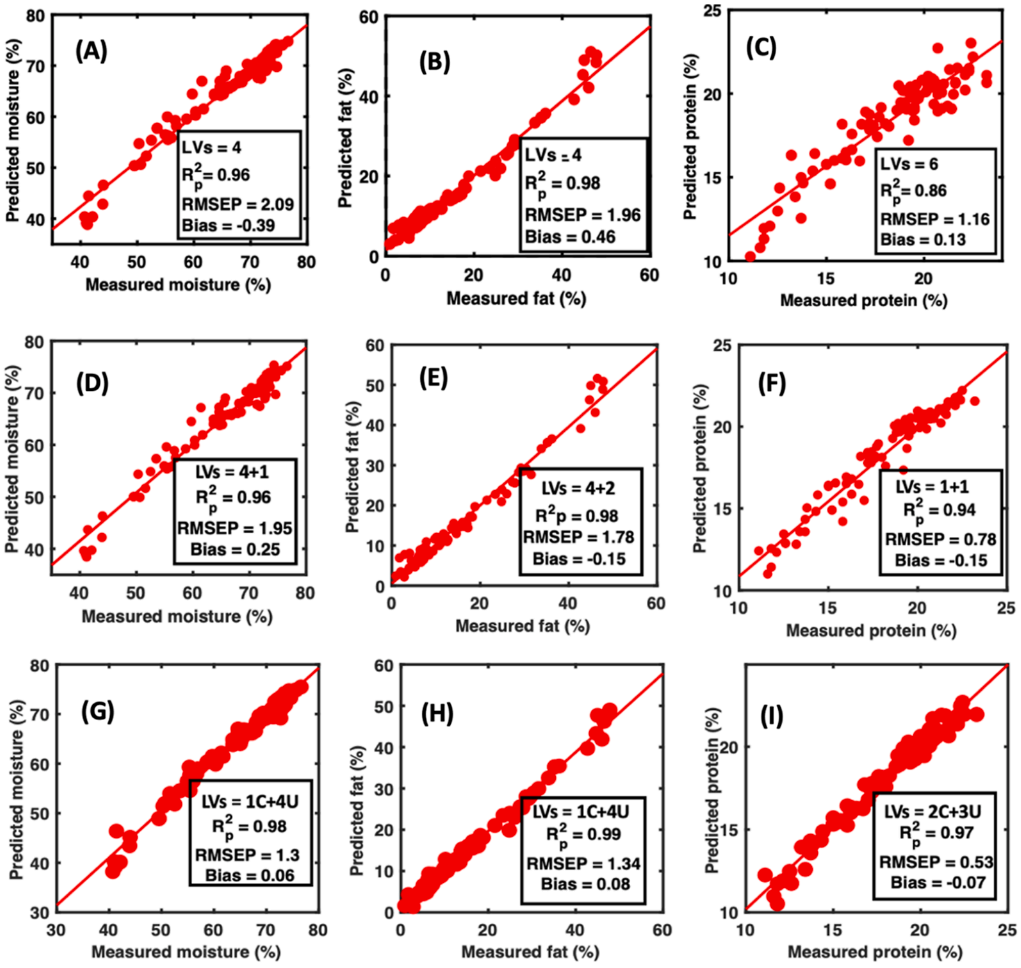

Fig. 2. Summary of the partial least-squares regression (PLSR), sequential preprocessing through orthogonalization (SPORT) and parallel preprocessing through orthogonalization (PORTO) models. Model details are expressed in latent variables (LVs), coefficient of determination (R^2_P), prediction bias (Bias) and root mean squared error of prediction (RMSEP). The PLSR prediction for (A) moisture content (%), (B) fat content (%), and (C) protein content (%). SPORT prediction for (D) moisture content (%) (4 LVs + 1 MSC), (E) fat content (%) (4 SNV + 2 MSC), and (F) protein content (%) (1 from 2nd derivative + 1 MSC). PORTO prediction for (G) moisture content (%) (1 common component + 4 unique components), (H) fat content (%) (1 common component + 4 unique components), and (I) protein content (%) (2 common components + 3 unique components).

Table 1
A summary of improvement in RMSEP for moisture, fat and protein with SPORT and PORTO approaches compared to PLS regression.

Properties	% improvement with SPORT compared to PLS	% improvement with PORTO compared to PLS	% improvement with PORTO compared to SPORT
Moisture	7	38	33
Fat	9	32	25
Protein	33	54	32

reduced the RMSEP by 54% compared to 33% reduction by the SPORT. On comparing the improvement by PORTO over the SPORT, it can be noted that the PORTO reduced the RMSEP by up to 33%. A better performance of PORTO can be accounted to its ability to perform a parallel fusion of information from several pre-processing, which extracts the information much more efficiently than the SPORT [26].

In the case of SPORT, the improvements were possible due to the ability of SPORT to model extra latent variables from a differently preprocessed data. For example, compared to the 4 LVs used by PLSR (based on SNV) for the prediction of moisture (Fig. 2A), the SPORT approach identified 1 extra LV from the MSC preprocessed block (Fig. 2D). Similarly, for the fat content, the SPORT approach identified 2 extra LVs (MSC) (Fig. 2E) which are not identified by the standard PLSR modelling performed utilizing only SNV (Fig. 2B). In the case of protein, SPORT only identified 2 LVs (1from 2nd derivative +1 from MSC) compared to the 6 LVs identified by standard PLSR performed using SNV. The identification of LVs from multiple scatter-correction techniques by SPORT suggests that complementary information was present in differently scatter-corrected data and that the fusion led to better performing models. Similarly, for PORTO modelling, there were both common and unique latent variables involved in optimal models indicating a synergistic use of complementary information present in different pre-processings.

4. Conclusions

Selection of a single scatter-correction technique in NIR data modelling does not allow proper utilization of complementary information highlighted by different scatter-correction techniques. This study shows that the performance of NIR models for predicting moisture, fat and protein content in minced pork can be improved with the fusion of information from different scatter-correction techniques. Both the sequential (SPORT) or parallel (PORTO) fusion approaches allowed modelling of complementary information highlighted by different scatter-correction techniques. PORTO performed slightly better than SPORT. This could be because PORTO is based on identifying more detailed information (common and distinct) in differently scatter-corrected data, whereas the performance of SPORT may be affected by the order in which the scatter-correction techniques are arranged. These fusion approaches have the further benefit that they take the user out of the loop of identifying the best technique, thus, saving time and resources. Based on the results of this study, it is highly recommended that scientific community explore these preprocessing fusion approaches to improve the predictive performance of NIR models.

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.

Acknowledgments

The authors would like to acknowledge the funding received from Top Sector Agri & Food foundation, The Netherlands under the project Pork Quality and Safety Assessment Tools (PorQSAT). Project number: LWV19176.

Prof. Douglas N. Rutledge is acknowledged for proof reading and correcting the English language for this manuscript.

References

[1] N. Prieto, O. Pawluczyk, M.E.R. Dugan, J.L. Aalhus, A review of the principles and applications of near-infrared spectroscopy to characterize meat, fat, and meat products, Appl. Spectrosc. 71 (2017) 1403–1426.
[2] C. Kapper, R.E. Klont, J.M.A.J. Verdonk, P.C. Williams, H.A.P. Urlings, Prediction of pork quality with near infrared spectroscopy (NIRS) 2. Feasibility and robustness of NIRS measurements under production plant conditions, Meat Sci. 91 (2012) 300–305.
[3] D. Andueza, A. Listrat, D. Durand, J. Normand, B.P. Mourot, D. Gruffat, Prediction of beef meat fatty acid composition by visible-near-infrared spectroscopy was improved by preliminary freeze-drying, Meat Sci. 158 (2019), 107910.
[4] Z. Nogalski, P. Pogorzelska-Przybyłek, I. Białobrzewski, M. Modzelewska-Kapituła, M. Sobczuk-Szul, C. Purwin, Estimation of the intramuscular fat content of m. longissimus thoracis in crossbred beef cattle based on live animal measurements, Meat Sci. 125 (2017) 121–127.
[5] J. Cafferky, T. Sweeney, P. Allen, A. Sahar, G. Downey, A.R. Cromie, R.M. Hamill, Investigating the use of visible and near infrared spectroscopy to predict sensory and texture attributes of beef M. longissimus thoracis et lumborum, Meat Sci. 159 (2020), 107915.
[6] X. Zheng, Y. Li, W. Wei, Y. Peng, Detection of adulteration with duck meat in minced lamb meat by using visible near-infrared hyperspectral imaging, Meat Sci. 149 (2019) 55–62.
[7] F. Qu, D. Ren, Y. He, P. Nie, L. Lin, C. Cai, T. Dong, Predicting pork freshness using multi-index statistical information fusion method based on near infrared spectroscopy, Meat Sci. 146 (2018) 59–67.
[8] L. Moran, S. Andres, P. Allen, A.P. Moloney, Visible and near infrared spectroscopy as an authentication tool: Preliminary investigation of the prediction of the ageing time of beef steaks, Meat Sci. 142 (2018) 52–58.
[9] C. Pasquini, Near infrared spectroscopy: a mature analytical technique with new perspectives – A review, Anal. Chim. Acta 1026 (2018) 8–36.
[10] P. Mishra, A. Biancolillo, J.M. Roger, F. Marini, D.N. Rutledge, New data preprocessing trends based on ensemble of multiple preprocessing techniques, TrAC, Trends Anal. Chem. 116045 (2020).
[11] H. Martens, J.P. Nielsen, S.B. Engelsen, Light scattering and light absorbance separated by extended multiplicative signal correction. Application to near-infrared transmission analysis of powder mixtures, Anal. Chem. 75 (2003) 394–404.
[12] J.-M. Roger, J.-C. Boulet, M. Zeaiter, D.N. Rutledge, Pre-processing Methods☆, Reference Module in Chemistry, Molecular Sciences and Chemical Engineering, Elsevier, 2020.
[13] A. Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem. 36 (1964) 1627–1639.
[14] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra, Appl. Spectrosc. 43 (1989) 772–777.
[15] T. Isaksson, T. Næs, The effect of multiplicative scatter correction (MSC) and linearity improvement in NIR spectroscopy, Appl. Spectrosc. 42 (1988) 1273–1284.
[16] J. Gerretzen, E. Szymańska, J.J. Jansen, J. Bart, H.-J. van Manen, E.R. van den Heuvel, L.M.C. Buydens, Simple and effective way for data preprocessing selection based on design of experiments, Anal. Chem. 87 (2015) 12096–12103.
[17] P. Mishra, J.M. Roger, D.N. Rutledge, E. Woltering, SPORT pre-processing can improve near-infrared quality prediction models for fresh fruits and agro-materials, Postharvest Biol. Technol. 168 (2020), 111271.
[18] J.-M. Roger, A. Biancolillo, F. Marini, Sequential preprocessing through ORThogonalization (SPORT) and its application to near infrared spectroscopy, Chemomet. Intell. Lab. Syst. 199 (2020), 103975.
[19] C. Borggaard, H.H. Thodberg, Optimal minimal neural interpretation of spectra, Anal. Chem. 64 (1992) 545–551.
[20] R.W. Kennard, L.A. Stone, Computer aided design of experiments, Technometrics 11 (1969) 137–148.
[21] G. Rabatel, F. Marini, B. Walczak, J.-M. Roger, VSN: Variable sorting for normalization, J. Chemom. 34 (2020), e3164.
[22] P. Mishra, J.M. Roger, D.N. Rutledge, A. Biancolillo, F. Marini, A. Nordon, D. Jouan-Rimbaud-Bouveresse, MBA-GUI: A chemometric graphical user interface for multi-block data visualisation, regression, classification, variable selection and automated pre-processing, Chemomet. Intell. Lab. Syst. 104139 (2020).
[23] P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial, Anal. Chim. Acta 185 (1986) 1–17.
[24] W. Saeys, N.N. Do Trong, R. Van Beers, B.M. Nicolai, Multivariate calibration of spectroscopic sensors for postharvest quality evaluation: a review, Postharvest Biol. Technol. (2019) 158.
[25] I. Måge, E. Menichelli, T. Næs, Preference mapping by PO-PLS: separating common and unique information in several data blocks, Food Qual. Prefer. 24 (2012) 8–16.
[26] P. Mishra, J.M. Roger, F. Marini, A. Biancolillo, D.N. Rutledge, Parallel pre-processing through orthogonalization (PORTO) and its application to near-infrared spectroscopy, Chemomet. Intell. Lab. Syst. 104190 (2020).