Journal of Pharmaceutical and Biomedical Analysis 192 (2021) 113684

Contents lists available at ScienceDirect

Journal of Pharmaceutical and Biomedical Analysis

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

Improved prediction of tablet properties with near-infrared spectroscopy by a fusion of scatter correction techniques

Puneet Mishra [a,*], Alison Nordon [b], Jean-Michel Roger [c,d]

[a] Wageningen Food and Biobased Research, Bornse Weilanden 9, P.O. Box 17, 6700AA, Wageningen, The Netherlands
[b] WestCHEM, Department of Pure and Applied Chemistry and Centre for Process Analytics and Control Technology, University of Strathclyde, Glasgow, G1 1XL, United Kingdom
[c] ITAP, INRAE, Montpellier SupAgro, University Montpellier, Montpellier, France
[d] ChemHouse Research Group, Montpellier, France

ARTICLE INFO

Article history:
Received 22 August 2020
Received in revised form 6 October 2020
Accepted 7 October 2020
Available online 10 October 2020

Keywords:
Multiblock
Fusion
Spectroscopy
Pre-processing
Multivariate

ABSTRACT

Near-infrared (NIR) spectra of pharmaceutical tablets get affected by light scattering phenomena, which mask the underlying peaks related to chemical components. Often the best performing scatter correction technique is selected from a pool of pre-selected techniques. However, the data corrected with different techniques may carry complementary information, hence, use of a single scatter correction technique is sub-optimal. In this study, the aim is to prove that NIR models related to pharmaceuticals can directly benefit from the fusion of complementary information extracted from multiple scatter correction techniques. To perform the fusion, sequential and parallel pre-processing fusion approaches were used. Two different open source NIR data sets were used for the demonstration where the assay uniformity and active ingredient (AI) content prediction was the aim. As a baseline, the fusion approach was compared to partial least-squares regression (PLSR) performed on standard normal variate (SNV) corrected data, which is a commonly used scatter correction technique. The results suggest that multiple scatter correction techniques extract complementary information and their complementary fusion is essential to obtain high-performance predictive models. In this study, the prediction error and bias were reduced by up to 15 % and 57 % respectively, compared to PLSR performed on SNV corrected data.

1. Introduction

Near-infrared (NIR) spectroscopy is widely used for rapid and non-destructive analysis of pharmaceutical tablets [1,2]. However, the main challenge with NIR spectroscopy is related to data modelling, which can easily become sub-optimal if no proper investigation is performed during the pre-processing stage [3–5]. The interaction of NIR light with samples results in two major optical phenomena, i.e., absorption and scattering [6]. Absorption is the result of specific chemical molecules absorbing certain wavelengths of light, whereas scattering is a result of the complex interaction of light with the physical structure of samples [6,7]. NIR data are widely affected by light scattering, which can be identified as additive and multiplicative effects in spectra [6,8].

Traditional data analysis related to NIR data modelling involves exploration of different scatter correction techniques to remove the scattering effects from the spectra such that the absorption characteristics can be used for proper estimation of chemical components [9]. However, recently Mishra et al., (2020) demonstrated that the NIR modelling should not aim to select a single scatter correction technique but should utilise the information extracted by several scatter correction techniques [5,10,11]. This is because different scatter correction techniques remove the effects of light scattering differently and may reveal complementary information, which if modelled can lead to higher performance models [5]. Complementary information from data pre-processed with different scatter correction techniques can be fused using the recently developed sequential and parallel pre-processing fusion approaches [10,12].

In this study, the aim is to prove that NIR models of pharmaceutical products can directly benefit from the fusion of complementary information extracted from multiple scatter correction techniques. To perform the fusion, sequential and parallel pre-processing fusion approaches were used. Two different open source NIR data sets were used for the demonstration where the prediction of assay uniformity and active ingredient (AI) content was the aim. As a baseline, the fusion approach was compared to partial least-squares

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

https://doi.org/10.1016/j.jpba.2020.113684

regression (PLSR) performed on standard normal variate (SNV) corrected data, which is a commonly used scatter correction technique.

2. Materials and methods

2.1. Data set

Two open source tablet data sets were used for this study. The first was the 'NIR shootout 2002' data set published by the International Council for Near-infrared Spectroscopy. The data were downloaded from the official website of Eigenvector Research Inc, USA in MATALAB readable format (http://www.eigenvector.com/data/tablets/index.html). The data in the downloaded. mat file contained spectra from two instruments. In this study, the spectra corresponding only to instrument 1 was used. In total, 654 tablets were measured with NIR (600–1898 nm) and reference analysis (assay uniformity). The second data set comprised NIR transmission data related to measurement of 310 tablets and the corresponding active ingredient (AI) content (%). The data were obtained from http://www.models.life.ku.dk/Tablets and were the same as presented in the original work [13]. The spectral range of this data set was 7398 – 10507 cm^{-1}. For both datasets, the samples were divided into calibration (60 %) and test (40 %) set using the Kennard-Stone algorithm [14].

2.2. Data analysis

2.2.1. Scatter correction methods

In the study, four of the most commonly used scatter correction techniques were selected [5]. The techniques were 2nd derivative [4], variable sorting for normalization (VSN) [8], standard normal variate (SNV) [15] and multiplicative scatter correction (MSC) [7]. The second derivative estimation was performed using the Savitzky-Golay method with a 2nd order polynomial and 21-point smoothing filter. All the pre-processing methods were implemented as discussed in [4] and in MATLAB, Natick, MA, USA.

2.2.2. Partial least-squares regression

PLSR is a common regression method used for NIR data modelling [16]. By maximising the covariance of the NIR data with the response(s), PLSR identifies the subspace of latent variables (LVs) on which the high-dimensional data can be transformed to give a low dimension information concentrated space. This transformation guarantees that the data are relevant for predicting the response variables. In the study, PLSR was performed using MATLAB's built-in function 'plsregress', with a 10-fold cross validation procedure approach used to select the optimal number of latent variables (LVs).

2.2.3. Sequential and parallel pre-processing through orthogonalisation

The sequential and parallel pre-processing through orthogonalisation (SPORT and PORTO) approaches are inspired from multiblock data analysis in chemometrics [12]. The sequential approach called SPORT is based on sequential orthogonalized partial least-squares regression and the PORTO approach on parallel orthogonalized partial least-squares regression. A schematic of the SPORT and PORTO approaches is presented in Fig. 1. In SPORT (see Fig. 1A), initially, a PLS regression model is fitted between **Y** and the **X** block after application of the first pre-processing method, and the scores for the first block (**T1**) are obtained. Then, the **X** block after application of the second pre-processing method is orthogonalized with respect to the scores (**T1**) of the first regression. Then the residuals of **Y** are fitted to the orthogonalized **X** block after application of the second pre-processing method and the scores (**T2**) are estimated. The procedure is continued for as many blocks (4 in this case) as there are pre-processing methods. In this work, the order of application of the scatter correction methods was 2nd derivative, VSN, SNV and MSC, making a total of 4 blocks of data. In the case of PORTO, a combination of PLS regression, generalized canonical analysis (GCA) and multiple orthogonalization steps are performed with the aim of extracting the common and distinct information presented in data pre-processed with different scatter correction methods. The concept of PORTO is to identify common and distinct information as shown in Fig. 1B. The three circles represent data pre-processed with three different scatter correction methods and the letters D and C indicate the distinct and the common information, respectively.

Optimising the number of LVs is highly important for both SPORT and PORTO. In the case of SPORT, all possible combinations of LVs were explored with the optimum number chosen based on the lowest RMSECV. In PORTO, several local cross-validations (CVs) were performed in sequence, as discussed in [17]. SPORT was implemented with the algorithm presented in [12] and with freely available multiblock data analysis toolbox [18]. PORTO was implemented using the multi-block data analysis codes from NOFIMA

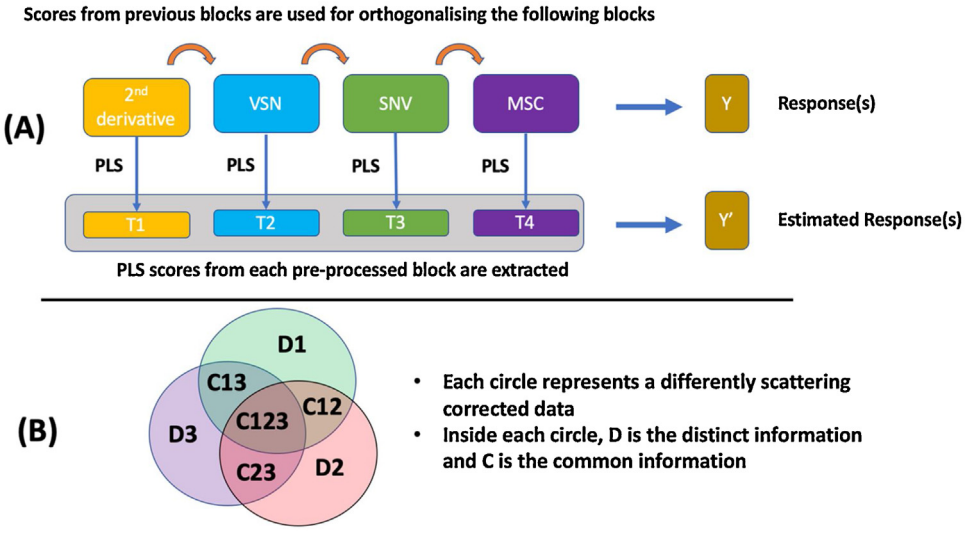

Fig. 1. A schematic of the sequential (SPORT) (A) and parallel (PORTO) (B) pre-processing fusion approaches. 1,2 and 3 in the PORTO approach are the data pre-processed with three different pre-processing techniques.

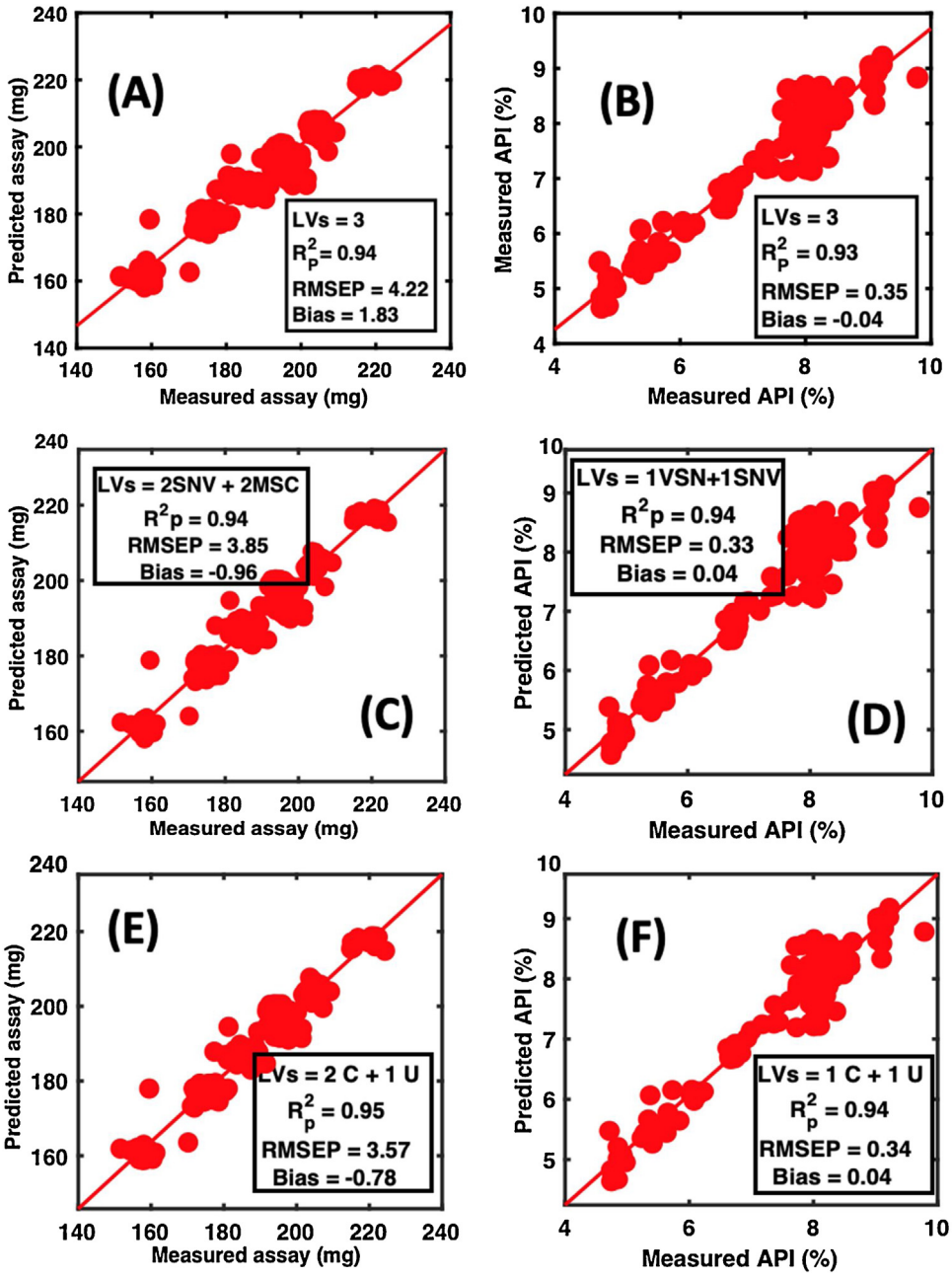

Fig. 2. Summary of PLSR applied to SNV pre-processed data, SPORT and PORTO models. (A) PLSR prediction for assay (mg), (B) PLSR prediction for API content (%), (C) SPORT prediction for assay (mg), (D) SPORT prediction for API content (%), (E) PORTO prediction for assay (mg), and (F) PORTO prediction for API content (%).

(https://nofima.no/en/) for the implementation of parallel orthogonalised partial least-squares (POPLS). All analysis was performed in MATLAB 2017b (The MathWorks, Natick, USA).

3. Results

3.1. PLSR modelling vs SPORT vs PORTO

The results from PLSR with SNV pre-processed data and the pre-processing fusion approaches (SPORT and PORTO) are shown in Fig. 2. It can be seen that fusion of information from different scatter correction methods improved the predictive performance of the models. In the case of prediction of assay uniformity with NIR data (Fig. 2A,C,E), the SPORT approach decreased the error and bias by 9% and 47 %, respectively, compared to PLSR with SNV pre-processed data. The PORTO approach further decreased the error and bias by 15 % and 57 %, respectively, compared to PLSR with SNV pre-processed data. For prediction of the AI content with NIR data (Fig. 2B,D,F), the main improvements were observed in error reduction with a 6% and 3% decrease using SPORT and PORTO approach, respectively. The improvement was attained with the same number of LVs as for the PLSR model constructed using SNV pre-processed data (3 LVs). Further, in all cases the best models were obtained using LVs from data corrected using multiple scatter correction methods, explaining that complementary information is being extracted and modelled by the pre-processing fusion approaches (PORTO and SPORT).

As an example to show the complementary and efficient information modelled by the SPORT and PORTO approaches, the loadings from prediction of the AI are shown in Fig. 3. It can be seen that com-

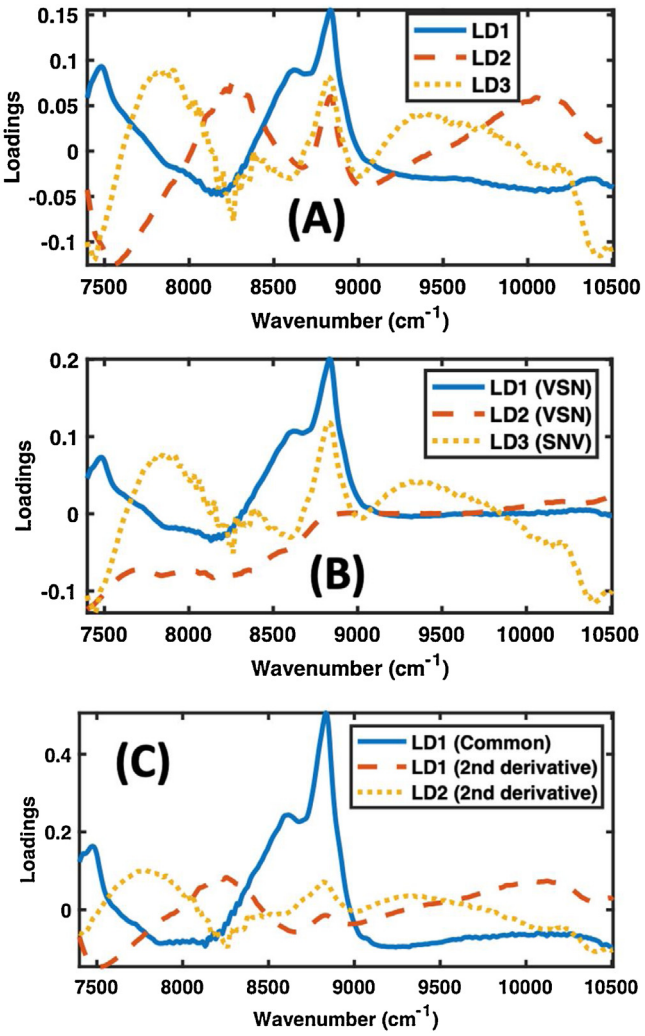

Fig. 3. Loadings corresponding to PLSR applied to SNV pre-processed data (A), SPORT (B) and PORTO (C) modelling approaches. (A) Three loadings were extracted by PLSR for optimal modelling, (B) three loadings were extracted by SPORT approach (2 from VSN pre-processed data and 1 from SNV pre-processed data), and (C) three components were extracted by the PORTO approach (1 common component for all pre-processing methods and 2 distinct components from 2nd derivative).

pared to the loadings of standard PLSR, the loadings of SPORT and PORTO are less noisy, and especially in the case of PORTO the loading weights in the spectral region around 10500 cm^{-1} are close to 0 with the region <9000 cm^{-1}, which arises from the C≡N overtones from the active ingredient, having the highest loadings [13]. Such efficient modelling can be understood as the reason for the improvement in the model performance.

4. Conclusions

This study has proved that NIR models of pharmaceuticals tablets can be improved with a fusion of complementary information present in data pre-processed using different scatter correction methods. Both sequential and parallel approaches to pre-processing can be used for the fusion. In this study, the parallel approach performed slightly better compared to the sequential approach. The error and bias decreased by up to 15 % and 57 %, respectively. Based on the results obtained, it is recommended that fusion of multiple scatter correction techniques should be performed rather than spending time on exploring and identifying the best scatter correction technique.

CRediT authorship contribution statement

Puneet Mishra: Conceptualization, Data curation, Investigation. **Alison Nordon:** Formal analysis, Writing - review & editing. **Jean-Michel Roger:** Formal analysis, Writing - review & editing.

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

[1] L.M. Kandpal, J. Tewari, N. Gopinathan, J. Stolee, R. Strong, P. Boulas, B.-K. Cho, Quality assessment of pharmaceutical tablet samples using Fourier transform near infrared spectroscopy and multivariate analysis, Infrared Phys. Technol. 85 (2017) 300–306.
[2] M. Alcalà, J. León, J. Ropero, M. Blanco, R.J. Romañach, Analysis of low content drug tablets by transmission near infrared spectroscopy: selection of calibration ranges according to multivariate detection and quantitation limits of PLS models, J. Pharm. Sci. 97 (2008) 5318–5327.
[3] Å. Rinnan, Fvd. Berg, S.B. Engelsen, Review of the most common pre-processing techniques for near-infrared spectra, Trac Trends Anal. Chem. 28 (2009) 1201–1222.
[4] 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.
[5] 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. (2020), 116045.
[6] C. Pasquini, Near infrared spectroscopy: a mature analytical technique with new perspectives – a review, Anal. Chim. Acta 1026 (2018) 8–36.
[7] T. Isaksson, T. Næs, The effect of multiplicative scatter correction (MSC) and linearity improvement in NIR spectroscopy, Appl. Spectrosc. 42 (1988) 1273–1284.
[8] G. Rabatel, F. Marini, B. Walczak, J.-M. Roger, VSN: Variable sorting for normalization, J. Chemom. 34 (2020) e3164.
[9] 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.
[10] 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.
[11] P. Mishra, F. Marini, A. Biancolillo, J.-M. Roger, Improved prediction of fuel properties with near-infrared spectroscopy using a complementary sequential fusion of scatter correction techniques, Talanta (2020), 121693.
[12] J.-M. Roger, A. Biancolillo, F. Marini, Sequential preprocessing through ORThogonalization (SPORT) and its application to near infrared spectroscopy, Chemom. Intell. Lab. Syst. 199 (2020), 103975.
[13] M. Dyrby, S.B. Engelsen, L. Nørgaard, M. Bruhn, L. Lundsberg-Nielsen, Chemometric quantitation of the active substance (Containing C≡N) in a pharmaceutical tablet using near-infrared (NIR) transmittance and NIR FT-Raman spectra, Appl. Spectrosc. 56 (2002) 579–585.
[14] R.W. Kennard, L.A. Stone, Computer aided design of experiments, Technometrics 11 (1969) 137–148.
[15] 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.
[16] 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. 158 (2019).
[17] 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.
[18] 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, Chemom. Intell. Lab. Syst. (2020), 104139.