Sequential Preprocessing through ORThogonalization (SPORT) and its application to Near Infrared Spectroscopy

Jean-Michel Roger[1,2], Alessandra Biancolillo[3], Federico Marini[4]

[1]*ITAP, Irstea, Montpellier SupAgro, University of Montpellier, Montpellier, France*

[2]*ChemHouse Research Group, Montpellier, France*

[3]*University of L'Aquila, Department of Physical and Chemistry science, L'Aquila, Italy.*

[4]*University of Rome "La Sapienza", Department of Chemistry, Rome, Italy.*

Abstract

In spectroscopy, multivariate calibrations more than often include a pre-processing step to reduce the effect of unwanted (not Y-related) sources of variability. Because there are many types of background noise, there are many pre-treatment methods. It is therefore tedious to select and/or combine the best pre-treatments. This article proposes to combine several pre-treatments through the use of sequential and orthogonalized partial least squares (SO-PLS), thus leading to a boosting method. The performances and properties of this new method, called Sequential Preprocessing through ORThogonalization (SPORT), are compared to those of a previously published stacking method. SPORT demonstrates very good calibration performances, but also the ability to make significant pretreatment selections.

Keywords

Spectroscopy, analytical techniques, pretreatments, ensemble, boosting, multi-block analysis, sequential and orthogonalized partial least squares (SO-PLS), sequential preprocessing through orthogonalization (SPORT)

Introduction

Spectroscopy and, more generally, analytical chemistry techniques provide multivariate outcomes. These instrumental signals are often collected in order to evaluate one or more properties of a product. This evaluation is generally based on calibrating a chemometric model, e.g., by principal component regression (PCA) or partial least square regression (PLSR). However, in these techniques, model building is based on extracting components which account for a relevant share of the variance in the predictor space. Accordingly, whatever phenomenon (wanted or unwanted) has an impact on the data variance can be included in the predictive model, quite often, in the case of spurious variance, with a detrimental effect. For this reason, the calibration of chemometric models generally includes a pre-treatment step, to reduce the effect of these interfering phenomena. Such interferences are generally due to the variation of physical (e.g,. temperature) or chemical (e.g., humidity) influence factors. Variations of these factors induce different effects on the measured spectra: the particle size affects the baselines, the temperature affects the position and relative amplitude of the peaks, the geometric configuration of the measurement induces additive and multiplicative effects, etc. The wide variety of effects has led to the development of a large number of pre-treatment methods [1]. The removal of a baseline, whatever its shape, is generally done by calculating suitable approximations, either by a polynomial (Detrend [2]), a low-frequency filter (ALS, Asymmetric Least Squares [3]), and then subtracting it from the measured data. The reduction of multiplicative effects is achieved either by logarithmic transformation or by normalization (SNV, Standard Normal Variate [2]), possibly weighted (VSN, Variable Sorting for Normalization [4]) or in comparison with a reference spectrum (MSC, Multiplicative Scatter Correction [5]). High frequency noise reduction is achieved by low-pass filtering (SG, Savitsky and Golay [6]). The highlighting of spectral details is done by differentiation, usually through the SG algorithm, which allows the computation of derivatives, without noise magnification. The correction of

harmful spaces, which can be observed, for example, when transferring a calibration between two spectrometers, can be done by an orthogonal projection method (EPO, External Parameter Orthogonalization [7]; TOP, Transfer by Orthogonal Projection [8]).

This raises the problem of choosing a suitable pre-treatment. In [9], different methodologies are examined to select the appropriate pre-treatment. This article concludes that "... *how extremely difficult it can be to determine which method - of the vast number of available pre-processing methods - can successfully help*...". In addition, these pre-treatments are often combined. In many articles, it can be read that baseline reductions are associated with normalizations or even derivatives. These choices are often based on a trial and error procedure, where the combination of different pre-treatments is done sequentially, e.g. SNV followed by a derivative. In this regard, it should be noted that the order according to which pre-treatments are applied may be important [9].

Another way to manage the choice and association of pre-processing is to use the so-called "ensemble learning" methods [10]. Ensemble learning is similar to that of "data augmentation", very common in deep learning [11]. Ensemble methods use several learning algorithms to achieve better performance than each algorithm used alone. Several approaches exist, as boosting, model averaging, model combination, buckets of models, stacking. All these methods differ roughly in the way the algorithms are aggregated. For instance, in boosting approaches weak learners are trained sequentially on different subset of samples (or, more generally, on different weighting scheme over the training individuals) and, in order for each successive model to be better than the predecessors, along the iterations higher weight (or higher probability of being selected) is given to the most difficult samples. Instead, model stacking, which is the form in which ensemble learning has mostly been used in chemometrics, consists in training a meta-model to output a prediction based on the outcomes of the various individual models, which, differently than in boosting can also be heterogeneous in nature; in

its simplest form, predictions from the weak learners can be gathered in a new data matrix, which is used to build a final model. In [12], it is proposed to divide the NIR spectra into intervals, then to stack the PLSR models performed on each interval by combining them by linear regression. In [13], it is proposed to stack different PLSR models using different pretreatments on the same data. Twenty different pre-processing operations, based on first and second derivatives, smoothing, SNV, MSC and their combinations were performed on NIR spectrum sets. A similar procedure, i.e., the exploitation of six PLS models (calculated on data preprocessed by diverse preprocessing approaches) as learners in an ensemble learning modelling algorithm was discussed in [14]. The output of this approach is computed by averaging the predicted values computed by its constituent learners.

In the present article, the concept that different preprocessing strategies applied to the same set of spectra could result in a multi-block data, and that, as such, the latter could be processed as a whole through dedicated multi-block strategies, is exploited in a boosting approach. In particular, due to its characteristics, the choice of the multi-block strategy to be adopted has fallen on sequential and orthogonalized partial least squares regression (SO-PLS, [15-16]), since it allows the possibility of including/excluding blocks, depending on their relevance, of evaluating the incremental contribution of the different matrices and, up to a certain extent, which blocks carry common and distinctive information. The resulting approach has been called Sequential Preprocessing through ORThogonalization (SPORT) and it will be described in detail in the following sections, together with examples of its application to real world NIR data sets.

Material and methods

Data sets

The proposed method was tested on three sets of real data:

Wheat grain data [17]: The NIR transmission spectra of wheat seeds were measured at 100 wavelengths and used to calibrate the protein content. The data set contained a calibration set of 415 samples and a test set of 108 samples.

Meat data [18]: The NIR transmission spectra of fine meat slices were measured at 100 wavelengths and used to calibrate the fat content. The data set contained a calibration set of 172 samples and a test set of 43 samples.

Tablet Data [19]: Near infrared spectra were collected on 310 tablets (spectral range between 7400 and 10507 cm-1) whose relative active substance contents (% w/w) were available. Data were divided by the Duplex algorithm [20] into a training and a test set of 210 and 100 samples, respectively.

Wheat and tablet data sets are freely available for download on the website of the Chemometrics and Analytical Technology group of the Copenhagen University (KU): www.models.life.ku.dk/datasets. The meat dataset is freely available for download on the website of the Carnegie Mellon University, at: http://lib.stat.cmu.edu/datasets/tecator.

The SPORT method

As anticipated in the Introduction, the proposed method is based on processing the different data matrices which result by preprocessing a spectral data set by different techniques through a multi-block approach called sequential and orthogonalized partial least squares regression (SO-PLS). As the name suggests, Sequential and Orthogonalized-Partial Least Squares (SO-PLS) is a multi-block regression method where the information is sequentially

extracted from the different predictor blocks. Very concisely, considering the simplest multi-block scenario, i.e., the case of two predictor blocks ($\mathbf{X}_1$ and $\mathbf{X}_2$) used to estimate a **y** response, the algorithm can be summarized by the following steps:

1. The **y** response is fitted to $\mathbf{X}_1$ by PLS.

$$\boldsymbol{y} = \boldsymbol{X_1} b_{X_1} + e_1 = T_{X_1} Q_{X_1}^T + e_1 \qquad (1)$$

2. $\mathbf{X}_2$ is orthogonalized with respect to the scores extracted from the first PLS regression.

$$X_2^{Orth} = \left[I - T_{X_1} \left(T_{X_1}^T T_{X_1} \right)^{-1} T_{X_1}^T \right] \boldsymbol{X_2} \qquad (2)$$

3. The orthogonalized $\mathbf{X}_2$ is used to predict the **y**-residuals obtained from step 1.

$$e_1 = \boldsymbol{X_2^{Orth}} b_{X_2^{Orth}} + e_2 = T_{X_2^{Orth}} Q_{X_2^{Orth}}^T + e_2 \qquad (3)$$

4. The full predictive model is calculated summing up results from step 1. and step 3.

$$\hat{y} = \boldsymbol{X_1} b_{X_1} + \boldsymbol{X_2^{Orth}} b_{X_2^{Orth}} = \boldsymbol{T}_{X_1} \boldsymbol{Q}_{X_1}^T + \boldsymbol{T}_{X_2^{Orth}} \boldsymbol{Q}_{X_2^{Orth}}^T \qquad (4).$$

In equations (1)-(4), b indicate regression coefficients, while $\boldsymbol{T}$ and $\boldsymbol{Q}$ the **X**-scores and **Y**-loadings, respectively.

If more data blocks are involved, they are orthogonalized with respect to all the previous modelled predictors and then used to estimate the **y**-residuals. A wider discussion over the algorithm and the advantages/disadvantages of the method can be found in [15,21]. In SPORT, several pretreatments of the same data block are associated in an SO-PLS.

For each dataset, SPORT was applied with different pretreatments. A repeated cross-validation (2 random blocks by 10 repetitions) was performed for a number of latent variables varying from 0 to 15 for each pretreatment. The results of this cross-validation allowed us to choose

the appropriate number of latent variables to keep for each pretreatment, using the *global approach* described in [22]. Briefly, this consist in building cross-validated SO-PLS models with all the possible combinations of latent variables (within a fixed maximum). The optimal complexity is then defined inspecting the RMSECVs. Then, a SO-PLS model using these numbers was calibrated on the calibration set and applied on the test set. The root mean squared errors of calibration (RMSEC), cross-validation (RMSECV) and prediction (RMSEP) were calculated at each step of the process.

In order to compare the results of SPORT with the ones provided by the stacking approach, the same pre-treatments as in [13] were performed on the wheat and meat datasets, as reported in table 1.

*Table 1: list of the pretreatments applied on the datasets wheat and meat. *: SG-W-O-D means savitsky and Golay using a W points wide window, an O^{th} order polynomial and a D^{th} derivative*

Pretreatment
SG-9-3-0*
SG-9-4-0
SG-9-3-1
SG-9-4-1
SG-9-3-2
SG-9-4-2
SNV

The order of the blocks may affect the results of SO-PLS, especially as far as the selection of the blocks is concerned. The tablet dataset was used to test the order in which the blocks are

153 introduced in SPORT. Five pre-treatments, including “raw data” (i.e., just mean centering), first
154 and second derivatives, SNV and VSN [21] were selected and combined as reported in Table 2.
155 Not all the possible orders were tested, but only the ones that reversed the orders of raw,
156 differentiating and normalization processes.

157 *Table 2: List of boostings of pretreatments applied on the tablet dataset, with different orders.*

Boosting 1	**Boosting 2**	**Boosting 3**
raw data	SNV	SG-15-3-2
SG-15-2-1	raw data	SNV
SG-15-3-2	SG-15-3-2	raw data
SNV	VSN, tol 0.0067, Npar 2	VSN, tol 0.0067, Npar 2
VSN, tol 0.0067, Npar 2	SG-15-2-1	SG-15-2-1

158

159 All the calculations discussed in the present paper were run under Matlab (The Mathworks
160 Inc., MA). All the functions can be freely downloaded from:

161 https://www.chem.uniroma1.it/romechemometrics/research/algorithms/

162 Results and discussion

163 *Table 3: Results of the single and ensemble models on wheat and meat datasets. [a]: results*
164 *taken from [13].*

	Wheat			Meat		
Pre-treatment	LVs	RMSEC	RMSEP	LVs	RMSEC	RMSEP
SG-9-3-0	11	0.53	0.71	6	2.97	2.80
SG-9-4-0	10	0.55	0.78	6	2.97	2.80
SG-9-3-1	8	0.55	0.66	11	2.11	2.09

SG-9-4-1	9	0.53	0.72	14	1.89	2.00
SG-9-3-2	6	0.54	0.52	10	1.97	2.08
SG-9-4-2	8	0.52	0.55	8	1.88	2.13
SNV	10	0.54	0.68	4	2.09	2.01
stacked[a]	-	*0.50*	*0.57*	-	*1.55*	*1.82*
boosted	0,0,4,0,0,0,11	0.47	0.47	0,0,0,0,0,7,7	1.50	1.65

Table 3 reports the results of the single block models, of the stacked approach and of the SPORT approach, for both wheat and meat datasets. It can be noted that the models built on raw data (after smoothing and mean centering) have poor performances and, in the case of wheat, also require a large number of latent variables. For the wheat dataset, the best pre-treatment is second derivative. It allows us to decrease the RMSEP from 0.70 to 0.51, i.e., a gain of 50% in the error variance (from 0.49 to 0.25). Moreover, the number of latent variables decreases from 11 to 6. When looking at the application of individual preprocessing strategies, for the meat dataset, the gain is less spectacular. Indeed, the best first-derivative and SNV approximately give the same RMSEP of about 2.00, in comparison to 2.01 for the model built on raw data (after smoothing and mean centering). However, SNV allows us to decrease dramatically the number of latent variables (from 14 to 4). As reported in [13], the stacked model is slightly outperformed by the best single block pretreatments for the wheat dataset, whilst it gives better performances for the meat dataset (RMSEP decreases from 2.01 to 1.82, i.e. the error variance decreases by 20%). The SPORT models give the best performances on both datasets. On the wheat data set, the resulting RMSEP is 0.47, representing a 55% gain in the error variance compared to the raw model. On the meat data set, the RMSEP is 1.65, representing 32% less error variance than the raw model. Compared to the stacked models, the RMSEP decreases from 0.57 to 0.47 and from 1.82 to 1.65 for the wheat and meat datasets, respectively, resulting in a corresponding improvement in error variance of 32 and

185 18%.

186 For both datasets, it can be noted that SPORT led to a very parsimonious selection of blocks,
187 since only two out of 7 blocks were used in the multi-block process. In both cases, a linear
188 pretreatment (SG) was associated with a non-linear pretreatment (SNV). For the wheat
189 dataset, SO-PLS has associated a first derivative (block 3) with SNV (block 7). This may seem
190 surprising at first sight, as the second derivative blocks (5 and 6) give much better performance
191 than the first derivative blocks (3 and 4). This illustrates well the ability of SO-PLS to identify
192 the blocks for additional (non-redundant) information, thanks to the orthogonalization steps
193 that take place between the addition of each new block. For meat data, the same
194 phenomenon can be observed. SO-PLS has associated a second derivative (block 6) with the
195 SNV (block 7), while the first derivative blocks (3 and 4) were the ones giving better
196 performances, when taken individually. In [16], Fig. 4 shows that stacking has also selected a
197 second derivative with SNV. This shows a good agreement between the two methods, with
198 regard to the selection of pre-treatments. Therefore, the reason of the difference in
199 performances observed should be ascribed to something else. It can be hypothesized that
200 stacking aggregation, which consists of a linear combination of the predictions of single block
201 PLS, is less effective than the iterative extraction of complementary information by SO-PLS. In
202 addition to the results obtained on our examples, this hypothesis is based on the fact that in
203 the case where the blocks are made up of different pretreatments of the same initial data set,
204 they contain a large redundancy of information, which is a situation that has to be handled
205 with precaution in a linear regression.

206 *Table 4: Results of different boosting orders on the tablet dataset.*

block number	Boosting 1	Boosting 2	Boosting 3
1	raw data	SNV	SG-15-3-2

2	SG-15-2-1	raw data	SNV
3	SG-15-3-2	SG-15-3-2	raw data
4	SNV	VSN, tol 0.0067, Npar 2	VSN, tol 0.0067, Npar 2
5	VSN, tol 0.0067, Npar 2	SG-15-2-1	SG-15-2-1
#LV	0,3,0,0,4	0,5,0,2,0	0,0,5,2,0
RMSEC	0.27	0.28	0.28
RMSEP	0.33	0.34	0.34

Table 4 shows the influence of the order in which the blocks are introduced into SPORT on the tablet data. First of all, it can be noted that the performances obtained, both in RMSEC and RMSEP, are not very much influenced by the order of the blocks. The total number of latent variables used is the same in any order (7), but is distributed differently over the selected blocks. SO-PLS systematically selected pairs of pre-treatments. It can be noted that the second derivative is never selected. The SNV is not selected either, to the benefit of VSN, which is on the contrary, always selected. This is in line with the fact that, on the one hand, VSN is a generalization of SNV and is therefore supposed to give better results [21] and that, on the other hand, at least one standardization method is selected because the tablet spectra contain a very strong multiplicative effect [1,18]. The "raw data" + VSN pair is selected twice out of three (boostings 2 and 3). In both cases, the block of the first derivative (SG1) was arranged after these two blocks. On the other hand, if block SG1 is interposed between the raw data block and the VSN block, it is selected instead of the raw data block.

Conclusion

This paper explores a new way to associate and select pre-treatments of spectra before their use in a calibration model. This involves applying the different pre-treatments to the same set of spectra, then combining the resulting blocks of data through a multi-block approach (SO-

PLS). As it proceeds by a sequential inclusion of the blocks that carry non-redundant information, the proposed method, called sequential preprocessing through orthogonalization (SPORT), belongs to boosting family of ensemble methods. The application of SPORT to different datasets shows the advantage of associating pre-treatments in an ensemble method. It also shows that our boosting approach, on the data sets used, is more efficient than the stacking approach, already published. On the other hand, the influence of the order according to which the individual blocks are processed was also studied. It appears that this order may, as expected, influence which blocks are combined in order to build the final model, but it has a rather small impact on the predictive performances. The use of SPORT in other, more complex data sets and applications, such as discrimination, will allow for a more in-depth study of its properties.

References

[1] Zeaiter, M., Rutledge, D.N., in Tauler, R., Walczak, B., & Brown, S. D. (2009). Comprehensive chemometrics: chemical and biochemical data analysis. Elsevier, 121-227

[2] Barnes, R. J.; Dhanoa, M. S.; Lister, S. J. (1993). Correction of the Description of Standard Normal Variate (SNV) and De-Trend Transformations in Practical Spectroscopy with Applications in Food and Beverage Analysis, 2nd ed. J. Near Infrared Spectrosc. 1, 185–186.

[3] Eilers, P. (2004). Parametric time warping. Analytical Chemistry, 76, 404-411.

[4] Rabatel, G., Marini, F., Walczak, B., & Roger, J. M. (2019). VSN: Variable sorting for normalization. Journal of Chemometrics, e3164.

[5] Isaksson,T.;Næs,T. (1988). The Effect of Multiplicative Scatter Correction (MSC) and Linearity Improvement in NIR Spectroscopy. Appl. Spectrosc. 42, 1273–1284.

[6] Savitzky, A.; Golay, M. J. E., (1972). Smoothing and differentiating of data by

249 simplified least squares procedure. Analytical Chemistry, 11 (44), 1906.
250 [7] Roger, J-M.; Chauchard, F.; Bellon-Maurel., V. (2003). EPO–PLS external parameter
251 orthogonalisation of PLS. Application to temperature-independent measurement of sugar
252 content of intact fruits. *Chemometrics and Intelligent Laboratory Systems*, *66*, 191– 204.
253 [8] Andrew, A.; Fearn, T. (2004). Transfer by orthogonal projection: making near-infrared
254 calibrations robust to between-instrument variation. *Chemometrics and Intelligent*
255 *Laboratory Systems*, *72*(1), 51-56.
256 [9] Engel, J., Gerretzen, J., Szymańska, E., Jansen, J. J., Downey, G., Blanchet, L., &
257 Buydens, L. M. (2013). Breaking with trends in pre-processing?. *TrAC Trends in Analytical*
258 *Chemistry*, *50*, 96-106.
259 [10] Dietterich, T. G. (2000, June). Ensemble methods in machine learning.
260 In *International workshop on multiple classifier systems* (pp. 1-15). Springer, Berlin,
261 Heidelberg.
262 [11] Wong, S. C., Gatt, A., Stamatescu, V., & McDonnell, M. D. (2016, November).
263 Understanding data augmentation for classification: when to warp?. In *2016 international*
264 *conference on digital image computing: techniques and applications (DICTA)* (pp. 1-6).
265 IEEE.
266 [12] Ni, W., Brown, S. D., & Man, R. (2009). Stacked partial least squares regression
267 analysis for spectral calibration and prediction. *Journal of Chemometrics: A Journal of the*
268 *Chemometrics Society*, *23*(10), 505-517.
269 [13] Xu, L., Zhou, Y. P., Tang, L. J., Wu, H. L., Jiang, J. H., Shen, G. L., & Yu, R. Q. (2008).
270 Ensemble preprocessing of near-infrared (NIR) spectra for multivariate
271 calibration. *Analytica chimica acta*, *616*(2), 138-143.
272 [14] Reda R., Saffaj T., Ilham B., Saidi O., Issam K., Brahim L., El Hadrami E. M. (2019). A
273 comparative study between a new method and other machine learning algorithms for soil

organic carbon and total nitrogen prediction using near infrared spectroscopy, *Chemometrics and Intelligent Laboratory Systems*, 195, 103873.

[15] Næs T., Tomic O., Mevik B.H., Martens H. (2011). Path modelling by sequential PLS regression. *Journal of Chemometrics: A Journal of the Chemometrics Society. 25,* 28–40.

[16] Biancolillo, A., & Næs, T. (2019). The Sequential and Orthogonalized PLS Regression for Multiblock Regression: Theory, Examples, and Extensions. *Data Handling in Science and Technology, 31*, 157-177.

[17] Nielsen, J. P., Pedersen, D. K., & Munck, L. (2003). Development of nondestructive screening methods for single kernel characterization of wheat. Cereal chemistry, 80(3), 274-280.

[18] Borggaard, C., & Thodberg, H. H. (1992). Optimal minimal neural interpretation of spectra. Analytical chemistry, 64(5), 545-551.

[19] Dyrby, M., Engelsen, S. B., Nørgaard, L., Bruhn, M., & Lundsberg-Nielsen, L. (2002). Chemometric quantitation of the active substance (containing C≡ N) in a pharmaceutical tablet using near-infrared (NIR) transmittance and NIR FT-Raman spectra. Applied Spectroscopy, 56(5), 579-585.

[20] R.D. Snee (1977). Validation of regression models: methods and examples. Technometrics, 19 415-428.

[21] Biancolillo A., Næs T., Bro R., Måge I. (2014). Extension of SO-PLS to multi-way arrays: SO-N-PLS, *Chemometrics and Intelligent Laboratory System*, 164, 113–126.

[22] T. Næs, O. Tomic, B.-H. Mevik, H. Martens. (2011). Path modelling by sequential PLS regression,J. Chemometr. 25, 28–40.