Food Chemistry 382 (2022) 132271

Contents lists available at ScienceDirect

Food Chemistry

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

Handling multiblock data in wine authenticity by sequentially orthogonalized one class partial least squares

Adriano A. Gomes[a,b,*], Liudmyla Khvalbota[b], Larisa Onça[a], Andrea Machyňáková[b], Ivan Špánik[b,*]

[a] Institute of Chemistry, Federal University of Rio Grande do Sul, Bento Gonçalves Avenue, 9500, 91501-970 Porto Alegre, RS, Brazil
[b] Institute of Analytical Chemistry, Faculty of Chemical and Food Technology, Slovak University of Technology in Bratislava, Radlinského 9, 812 37 Bratislava, Slovakia

ARTICLE INFO

Keywords:
Multiblock data
Tokaj wine
One class classification

ABSTRACT

New approach to deal with food authentication by modelling methods based on data recorded from different sources is proposed and called OC-PLS, combines an orthogonalization step between the different data sets to eliminate redundant information followed by definition of an acceptance area for a target class by OC-PLS. The proposed method was evaluated in two case studies. The first study used a controlled scenario with simulated data. In the second case study, the approach was applied using UV–VIS and IR data, in order to differentiate Slovak Tokaj Selection wines of high quality from other lower market value wines from the Slovak Tokaj wine region. In both cases, better results were reached than when individual blocks of data were achieved. The proposed method proved to be effective in properly exploring common and distinct information in each data block. The best compromise between sensitivity and selectivity in the prediction step was achieved.

1. Introduction

Food integrity analysis is the most comprehensive term used to cover various aspects of food analysis, such as quality, safety, confirmation of the use of declared processing technology, fraud detection, or supplement of valuable food constituents by cheaper synthetic ones with lower dietary value. The first three aspects can be confirmed by detailed inspection to ensure that the food complies with the current regulations (Robson et al., 2021). The last two are related to mitigation and crime prevention. In other words, the idea of food integrity is to prove that a food is what it claims to be (Montgomery et al., 2020; Robson et al., 2021).

In recent years, much effort has been devoted to development of effective methods to ensure and confirm food integrity (Montgomery et al., 2020). These methods cover various analytical approaches, like evaluation of food dietary properties including macro and micronutrients (Araújo et al., 2021; Gamela et al., 2020), detection and quantification of toxic substances naturally occurring due to degradation processes (Pinto et al., 2016), detection of adulterants (Reile et al., 2020; Xie et al., 2021), authentication with respect to geographical origin (Arndt et al., 2021; Qi et al., 2021; Ríos-Reina et al, 2020), brand (de Lima et al., 2020) or the prescribed production technology employed (Sandler et al., 2021). With regard to food integrity, an important aspect is related to fraud; this group of non-conformities is deliberately produced in order to obtain undue profits (Montgomery et al., 2020). Therefore, an integral product must be completely fraud free.

Nowadays, analytical methods must fulfil common requirements such as having a fast analysis time; being cost-effective, efficient and robust; if possible being non-destructive and / or non-invasive and eco-friendly with low waste and hazardous residues generation, involving less solvent consumption and ideally combining with multivariate approaches. Despite progress achieved during past 15 years, there are still gaps that need to be filled due to the wide complexity of food matrices. Additionally, frauds are highly dynamic, sophisticated and are developing rapidly, so inspection methods need some time to uncover novel “fraud approaches” (Ulberth, 2020).

Within the context of the use of chemometrics tools for food authenticity analysis, two ways have shown a significant development in the last decade. Those are multiblock (Næs et al., 2013; Smilde et al., 2017) and one class classification (Oliveri & Downey, 2012). The first one is related to simultaneous analysis of several data blocks from different sources (Mishra et al., 2020; Campos & Reis, 2020). The combined use of multiblocks of data can provide an enhanced and comprehensive understanding of the common and the distinct

* Corresponding authors.
E-mail addresses: araujo.gomes@ufrgs.br (A.A. Gomes), ivan.spanik@stuba.sk (I. Špánik).

https://doi.org/10.1016/j.foodchem.2022.132271
Received 3 August 2021; Received in revised form 24 January 2022; Accepted 25 January 2022
Available online 30 January 2022

information coming from different sources. Furthermore, it can guarantee higher efficiency, sensitivity, and selectivity. In many cases, two or more data blocks are just placed side by side in a row-wise augmented structure and treated as if they were a single block by conventional tools such as principal component analysis (PCA), partial least-squares (PLS). (Alamar et al., 2020; Jurado-Campos et al., 2020). Unfortunately, this strategy does not consider the intrinsic characteristics of the multiblock systems resulting in worse classification efficiency. For adequate modelling of multiblocks, a series of methods has been proposed in the literature that explicitly consider the structure of the common and the distinct data information (such as Common Dimensions (ComDim) (Cariou et al., 2018), Distinct and common simultaneous component analysis (DISCO-SCA) (Schouteden et al., 2013), Multi-block principal component analysis (MB-PCA) (Trygg & Wold, 2003), Multi-block partial least-squares (MB-PLS) (Næs et al., 2013), Sequential orthogonalized partial-least squares regression (SO-PLS) (Biancolillo et al., 2015) and others as described in (Smilde et al., 2003)).

The second one is based on the use of class modelling tools to solve food integrity problems (Rodionova et al., 2016). Although the vast majority of published works use discriminating methods like linear (LDA), quadratic (QDA) and PLS (PLS-DA) discriminants, the literature over past 10 years has shown that these approaches are conceptually misleading (Brereton, 2011; Rodionova et al., 2016). Discriminative methods are based on the definition of a boundary between classes; this means that all classes contributed to the definition of the decision rule equally and therefore must be properly sampled. In authentication studies, this is not a trivial or even possible task. On the other hand, class modelling strategies focus on delimiting borders for each class individually. They are also called one class methods, because each class is modelled separately (Oliveri et al., 2021). In an authentication study, the authentic class is treated as a target class, and modelled with no non-authentic samples' contribution. In the prediction stage, if an unknown sample does not match the target class, it is considered as non-authentic. The great advantage is that this approach does not require inclusion of all possible forms of fraudulent samples.

To the best of our knowledge the contributions found in the literature when using one class in food authentication have not included data from different sources; on the other hand, when data fusion/multiblock methods are used, the modeling is always discriminate (Azcarate et al, 2021; Borrás et al, 2015). Ultimately, when modeling methods are used, they are conducted in a compatible mode as described elsewhere (Rodionova et al., 2016). Some recent multiblock/data fusion and class-modelling in food authenticity applications are listed in Table 1S to describe the problem mentioned above.

Authentication studies must be performed by modelling methods and the combination of multiple data blocks could significantly improve the results. In this work, we propose to explore the advantages of multiblock models to consistently deal with food authenticity problems via one-class approaches. The proposed method, called sequentially orthogonalized one class partial least square SO-OC-PLS, includes a sequential orthogonalization step to treat the different data blocks that are modelled via one class PLS in a rigorous way. Our approach was evaluated in two case studies. The first one was a controlled scenario with simulated data. Furthermore, it was also applied to the classification of Tokaj Selection wines, a noble sweet wine produced in Tokaj region that falls close to the border between Slovakia and Hungary. The approach combined ultra-violet and infrared spectroscopy data.

2. Theory

2.1. Notation

In following text, matrices, vectors and scalars will be denoted by bold capital letters, bold lowercase letters and italic characters, respectively. The *T* superscript indicates the transpose of a vector or matrix.

2.2. One class partial least square – OC PLS

One class partial least square (OCPLS) proposed by Xu et al. (2013, 2014), combines one-class modeling suitable for authentication problems with the versatility of PLS. The regression equation is given by:

$$\mathrm{y} = \mathbf{X}_{(\mathrm{I}\times\mathrm{j})} \times \mathbf{b}_{(\mathrm{j}\times 1)} + \mathbf{e}_{(\mathrm{I}\times 1)} \tag{1}$$

In traditional discriminant PLS, **X** is fitted to a dummy matrix (**Y**) where 1 indicates that the i*th* sample belongs to the n*th* class, and zero indicates the opposite. However, in the OC-PLS training step, just samples belonging to the target class are present. In Equation (1), **y** represents a vector of elements equal to 1 and **X** are matrices of instrumental responses recorded in *j* feature for *i* training samples, **b** is the regression vector of the PLS model and **e** are the residues of the model. It is important to note that **X** cannot be column centered; this would make all column vectors in **X** orthogonal to **y**. The standard deviation (σ_e) by leave one out cross validation (LOOCV) is used to guide the selection of the suitable number of latent variables or factors (*A*), it is computed by Eq. (2).

$$\sigma_e = \sqrt{\frac{\sum_{i=1}^{I}\left(1 - \widehat{y}_i - \widehat{\mu}_e\right)^2}{dof}} \tag{2}$$

Where $\widehat{y}_i$ is the predicted response by the model; $\widehat{\mu}_e$ is mean of training errors and *dof* are degrees of freedom (*i*-1). After selecting an appropriate number of factors based on minimizing the σ_e value, the final model is computed, while two different types of distance are used to estimate the acceptance area of the target class. The first one is the absolute centered residual (**acr**) that can be interpreted as a dispersion measure of the projection onto the regression coefficients estimated by OC PLS for *A* factors (Xu et al., 2013). The **acr** vector can be calculated by Eq. (3).

$$\mathbf{acr} = \left|1 - \widehat{y}_i - \widehat{\mu}_e\right| \tag{3}$$

The **acr** is assumed to have a normal zero-centered distribution. The upper confidence limit (*acr*$_{ul}$) can be estimated by the following Eq. (4) (Xu et al., 2011):

$$acr_{ul} = Z_{\alpha/2} \times_{\widehat{\sigma_e}} \tag{4}$$

The second type of distance is computed in the space of the scores (score distance – *SD*, see eq. (5)) that has Hotelling's T^2 distribution and measures the distance of a sample to the center of the target class in the space of the scores defined by the *A* factors.

$$\mathrm{SD} = \sum_{i=1}^{A} \frac{\left(\boldsymbol{t}_{i,a} - \bar{t}\right)_i^2}{s_i^2} \tag{5}$$

where $\mathbf{t}_{i,a}$ is the *ith* score vector included in the model, $\bar{\mathbf{t}}_i$ and s_i^2 are the mean and variation of the $\mathbf{t}_i$ vector respectively. Since *SD* follows a T^2 statistic, the upper confidence limit (*SD*$_{ul}$) can be calculated by Eq. (6) (Xu et al., 2011) considering *F*-distribution critical value (degrees of freedom: A, I – A).

$$SD_{ul} = \frac{(\mathrm{I}-1)\times \mathrm{A}}{\mathrm{I}\times(\mathrm{I}-\mathrm{A})} \times F_{\alpha(\mathrm{A},\mathrm{I}-\mathrm{A})} \tag{6}$$

Then *acr*$_{ul}$ and *SD*$_{ul}$ are combined to define a rectangular acceptance area for samples of the target class at a statistical significance (α1 and α2 *acr*$_{ul}$ and *SD*$_{ul}$ respectively). In the prediction group, unknown samples that fall within the acceptance area are considered authentic and belong to the target class; otherwise samples that do not fall within the acceptance area are rejected in the target class. Note that the area of acceptance is defined considering only the target class; this type of model is known as a rigorous approach (Brereton, 2011; Oliveri et al., 2021; Rodionova et al., 2016; Xu et al., 2013, 2014).

2.3. Proposed Method: SO-OC-PLS

The proposed Sequentially Orthogonalized One Class Partial Least Squares (SO-OC-PLS) method combines multiblock data processing by means of a sequentially orthogonalized (Brereton, 2011) approach in latent variable space with one class classifier. The following is a description of SO-OC-PLS (Ballabio, 2015; López et al., 2014) for two hypothetical blocks **X1** and **X2**; this approach, however, can be easily generalized to three or more data blocks. The contribution of two blocks to predicted response ($\widehat{\mathbf{y}}$) can be described as indicated in Eq. (7).

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}1_{(\mathrm{I}\ \times\ \mathrm{J})} \times \boldsymbol{b}1_{(\mathrm{j}\times 1)} + \boldsymbol{X}2_{(\mathrm{I}\ \times\ \mathrm{k})} \times \boldsymbol{b}2_{(\mathrm{k}\times 1)} \tag{7}$$

Where **X1** and **X2** are the blocks of data that were recorded for the same set of samples and came from different sources. They can have different column dimensions, in these cases expressed as *J* and *K* respectively. The **b1** and **b2** are the regression vectors computed sequentially by PLS. Firstly, the model is fitted over block **X1** (see Eq. (1)) and then block **X2** is orthogonalized with respect to the first one in the score space to remove redundant information in the final model as shown in Eq. (8).

$$\boldsymbol{X}2_{\mathrm{ort}} = \boldsymbol{X}2\ -\ \boldsymbol{T}_{X1}\left(\boldsymbol{T}_{X1}^{T}\boldsymbol{T}_{X1}\right)^{-1}\boldsymbol{T}_{X1}^{T}\ \times \boldsymbol{X}2 \tag{8}$$

where $\mathbf{X2}_{\mathrm{ort}}$ corresponds to the second data block after orthogonalization in the space of the **X1** scores, contained in the $\mathbf{T}_{\mathbf{X1}}$ matrix. Then $\mathbf{X2}_{\mathrm{ort}}$and **y**-deflected (computed according to Eq. (9)) are fitted by PLS as shown in Eq. (10).

$$\mathbf{y}\mathrm{defleted} = \mathbf{y} - \mathbf{X}1\mathbf{b}1 \tag{9}$$

$$\widehat{\mathbf{y}}\mathit{defleted} = \mathbf{X}2_{\mathrm{ort}} \times \mathbf{b}_2 + \mathbf{e}_{2(\mathrm{I}\times 1)} \tag{10}$$

The appropriate number of factors for block 1 (A_1) and block 2 (A_2) are estimated using LOOCV employing an inner loop from 1 to A_{1max}; and 1 from A_{2max}. A_{max} refers to the maximum number of latent variables set by the user for each pair A_1, A_2. The standard deviation σ_e as defined in Eq. (2) is calculated. The final model is estimated considering the optimal values of A_1 and A_2. The final answer considering both blocks is given by Eq. (11).

$$\widehat{\mathbf{y}} = \mathbf{X}1_{(\mathrm{I}\times\mathrm{J})} \times \mathbf{b}1_{(\mathrm{j}\times 1)} + \mathbf{X}2_{\mathrm{ort}(\mathrm{I}\times\mathrm{k})} \times \mathbf{b}2_{(\mathrm{k}\times 1)} \tag{11}$$

Note that **b1** and **b2** are the vectors of regression coefficients obtained by sequential PLS models (see Eq. (12)). In the case of **b1**, computation is based on the model fitted between **X1** and y. The b2 vector, on the other hand, is estimated to fit to the PLS model between **X2ort** (Eq. (8)) and ydefleted (Eqs. (9) and (10)).

$$\boldsymbol{b} = \boldsymbol{W}\left(\boldsymbol{P}^{\mathrm{T}}\boldsymbol{W}\right)^{-1}\mathrm{q}^{\mathrm{T}} \tag{12}$$

Where **W**, **P** and **q** are the PLS loadings weights, loadings and regression coefficients in score space respectively. A step-by-step description on how to compute so-pls can be found elsewhere (Næs, at al 2010).

The acceptance area is defined according to procedure described in section 2.2 *One class partial least square – OC PLS*, but Eq. (5) is modified to consider the contribution of both blocks given Eq.13.

$$\mathrm{SD} = \sum_{b=1}^{B}\left(\sum_{\mathrm{i}=1}^{\mathrm{A}b}\frac{\left(\boldsymbol{t}_{\mathrm{i,a}} - \bar{t}\right)_{i}^{2}}{s_{i}^{2}}\right)_{b} \tag{13}$$

Where *B* is the number of blocks being considered in multiblock modeling, and A_b the number of latent variables in the B*th* block of data. It is important to note that in SO approaches, the second block is orthogonalized with respect to the first in sequential mode making the two PLS models independent of each other. As the models are orthogonal to each other, the final answer will be additive (Næs, at al 2010), making it possible to compute the distance from the sample to the center of the training in score space set as shown in (13).

Once the α_1 and α_2 values are defined, the acceptance area of the target class is built. Test set samples are processed according to Eq. (7), and the values of *acr* and *SD* are calculated. The test samples that present values of both metrics compatible with the training set will fall within the acceptance area and can be considered as a part of the target class (or regular sample). On the other hand, samples that exhibit higher values for one of the distance measures or both, will be classified as target class outliers. Samples not belonging to the target class can be categorized into three different types of outliers: type 1 with large *SD* and small *acr*; type 2 with small *SD* and large *acr*; and type 3 with both large *SD* and *acr* (Ballabio, 2015; López et al., 2014; Pomerantsev & Rodionova, 2013). The sequentially orthogonalized partial least squares approach takes information sequentially from each block of data, for that reason it is important to treat the chosen order of the blocks by SO-PLC to deliver different results. These differences will occur because the spaces spanned by Tx1/X2ort and Tx2 and X1ort are not the same. A good practice is to permute the blocks and check the quality of the results in order to choose the most convenient order (Campos et al., 2017).

3. Experimental section

3.1. Simulated data

The data set was simulated in order to demonstrate an algorithm of the proposed method and to highlight its superiority in relation to the application of typical methods used for processing as two independent data blocks placed side by side or row-wise agumnetated matrix. Firstly, six Gaussian profiles were generated, as shown in Fig. 1Sa using Eq. (14).

$$\boldsymbol{p} = \sum_{i=1}^{I} e^{-\frac{(\boldsymbol{x}-\mu)}{\sigma}} \tag{14}$$

where **X** is a 1 × 100 vector, μ and σ define center and width of Gaussian profile. If *i* is greater than 1, multimodal profiles were generated. In addition to the target class (*C1*), two non-target classes (*C*2 and *C*3) were created by combining the profiles in Fig. 1Sa according to Table 2S. Class 1 (target) samples are distinguished from Class *C*2 samples by the information contained in block1 and from Class 3 based on information from block 2. The training set contains 100 target samples while test set 250 samples of the target class and 150 samples for each non-target class. The dataset was created by the sum of the profiles specified in Table 1S with different weights (*w*) generated by the MatLab normrnd function.

The mean values and standard deviation of the weights (*w*) for each profile were selected to generate up to 10 % overlap between the target and the non-target classes. The training and test sets (see Fig. 1Sb) are composed of two matrices named $\mathbf{X}_{\mathbf{trainblock1}}/\mathbf{X}_{\mathbf{trainblock2}}$ and $\mathbf{X}_{\mathbf{testblock1}}/\mathbf{X}_{\mathbf{testblock2}}$, each sized 100 × 100 and 500 × 100 respectively. For all samples, normally distributed noise of 1% was added to the data.

3.2. Wine classification

3.2.1. Samples

Varietal Tokaj wines, Tokaj cuvée, different Tokaj “putňa” selections and Tokaj essence of vintages (1959–2017) were included in this study. All 58 samples were stored in bottles at 4 °C. After opening, each wine aliquot was transferred into a 20 mL vial and equilibrated at 20 °C before measurements. The samples were obtained directly from producers.

3.2.2. UV–VIS spectra

UV–VIS spectra were recorded using the Shimadzu UV-1800 Spectrometer with tungsten-halogen and deuterium lamps. UV PROBE 2.33 software was used for spectral acquisition and data processing. Samples were placed into the conventional 10 × 10 × 45 mm quartz cell. UV–VIS spectra of diluted wine samples (1% w/w in water) were obtained in wavelength range 190–1100 nm with 1.0 nm of sampling interval and slit width of 1.0 nm.

3.2.3. *Infrared spectra*

IR spectra were recorded with the Customs Laboratory of the Financial Directorate of the Slovak Republic which is accredited according to ISO/IEC 17025:2017. The IR spectra of the Tokaj wines samples were obtained using a Shimadzu IR Prestige-21 infrared spectrometer and the ATR method (attenuated total reflectance) in the region of 4000–650 cm^{-1} on average of 16 scans, with a resolution 4 cm^{-1}. Sample was measured directly without pretreatment. The measured amount of the sample was 1 mL. The wine samples were dried for 4 h in an oven at 105 °C until constant mass.

3.3. *Software*

All calculation was carried out in a MatLab 2010a environment. The training set contained 30 wine samples form target group of 40 wines selected by by the Kernnard-Stone algorithm (Xu et al., 2011). Additionally, a test set contained 10 target and 18 non-target wine samples. For the UV–VIS spectra the following region of interest (ROI) were defined as 199 to 490 nm; base line IR spectra were corrected by first derivative Savitzky-Golay filter with 15 points and second order polynomial (derivative spectra can be visualized on Fig. 3S). PCA32 and OC-PLS30 calculations were carried out using toolboxes available at https://michem.unimib.it/ and http://www.tinyupload.org/q3fvf84wj6g respectively. All other calculations were performed using homemade

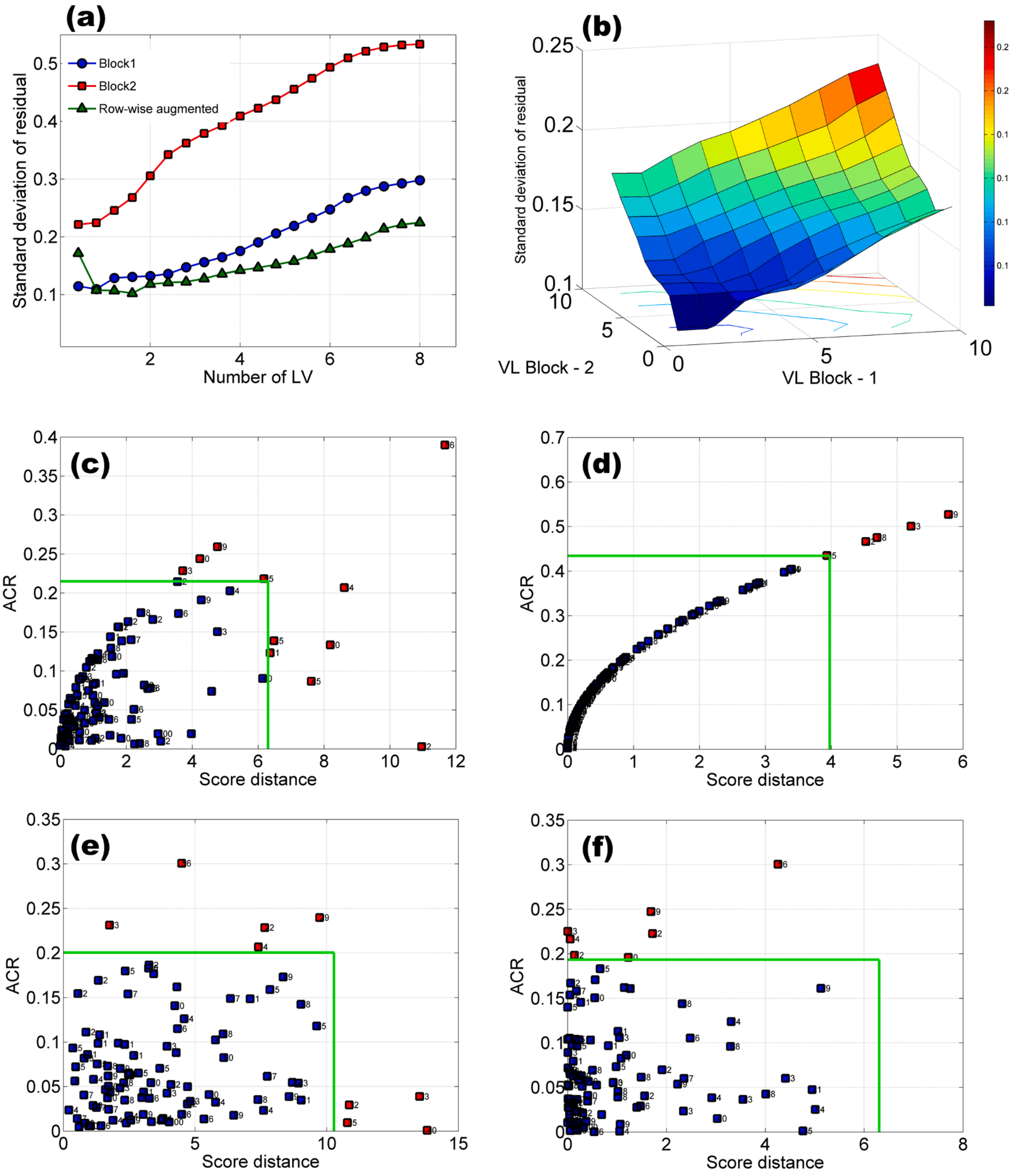

Fig. 1. Standard deviation of residuals in cross validation (a) for individual and row-wise augmented data, (b) the proposed method and acceptance area considering α1 (*SD*) = 0.05 and α2 (*acr*) = 0.05 for (c) block 1, (d) block 2, (e) row-wise augmented and (f) proposed method. Blue squares are the true positives and red squares are the false negatives. The green lines represent the limits of the acceptance area. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

routines, available by request from the corresponding authors.

The proposed method was evaluated in relation of individual data blocks and row-wise augmented data considering the sensitivity (López et al., 2014) (*SEN*), selectivity (*SPC*) and efficiency (*Eff*) as shown in Eqs. (15), 16 and 17, where TP, TN, FP and FN mean true positive and negative, false positive and negative.

$$\mathbf{SEN} = \frac{\mathbf{TP}}{(\mathbf{TP}+\mathbf{FN})} \tag{15}$$

$$\mathbf{SPC} = \frac{\mathbf{TN}}{(\mathbf{TN}+\mathbf{FP})} \tag{16}$$

$$\mathbf{Eff} = \frac{(\mathbf{TP}+\mathbf{TN})}{(\mathbf{TP}+\mathbf{FP}+\mathbf{TN}+\mathbf{FN})} \tag{17}$$

4. Results and discussion

4.1. Simulated system

Simulated data set was designed to show the benefits of the proposed method. In the first step, both blocks 1 and 2 were modelled by OC-PLS followed by reorganization to side by side row-wise augmented structure. The prepared dataset underwent OC-PLS treatment and finally was processed by SO-OC-PLS. For all cases, the same criteria were adopted to select an appropriate number of latent variables as shown in Fig. 1.

The minimization of standard deviation of residuals calculated by LOOCV (σ_{cv}), depicted for both blocks and row-wise augmented data (see Fig. 1a), was used to guide the choice of the number of latent variables in each data set. The best number of factors were 2 for block 1, 1 for block 2 and 4 for the row-wise augmented matrix. In our proposed methods, the LOOCV process consists of a nested loop, thus σ_{cv} is a function of the latent variables of both blocks, as can be seen in Fig. 1b. The lowest σ_{cv} was found for 2 and 1 LVs in blocks 1 and 2 respectively. Considering the number of factors mentioned above for each case, the distances *acr* and *SD* were calculated and the acceptance areas (Fig. 1) were built for both α equal to 0.05.

Since all samples belonged to the target class, only two status are possible, those that fall within the acceptance area are true positives, otherwise they are false negatives. The sensitivities reached were of 0.89, 0.95, 0.91 and 0.93. In general, the results are satisfactory considering both α level 0.05. The proximity of the sensitivity values close to 1-α indicates the appropriate choice of the number of latent variables, indicating that the sensitivity in training set is in agreement with the *a priori* α value.

When the boundaries of the acceptance area are set to an α value in a large data set, as in this simulated case, some samples will appear naturally outside of the acceptance area. This means that samples with 3σ (in terms of *acr* or *SD*) value belong to the target class but are located outside region of interest, thus they are considered as extreme samples of the selected population (Pomerantsev & Rodionova, 2013). Note that only outlier type 1 was identified in this simulated data set using the calculated area of acceptance (Fig. 1f). In addition, when both data blocks were combined, a more comprehensive description of the samples was achieved. None of the type 3 outliners was observed, suggesting that multiblocks can improve quality of results especially when they are treated properly.

Afterwards, all modes of predictive capacity were evaluated to classify target and non-target samples in the independent data test set. The location of the test set samples in relation to the acceptance area, defined as a 95 % confidence level, are shown in Fig. 2S. The sensitivity, specificity and efficiency for different α values is summarized in Table 1. As can be seen, individual blocks did not achieve a high efficiency classification, even if the data were combined in a row-wise augmented matrix. This finding reinforces the idea that multiblocks should be treated together and not as a single set of data. In general, the proposed method achieved results in compliance with the level of overlap defined in data setup for all α values. Efficiency was practically similar when the levels of 0.01 and 0.05 were compared, but showed a decrease when α was equal to 0.1. It was expected, that the orthogonalization process would reduce the number of latent variables when comparing the model that crossed both blocks of data.

Table 1
Statistical summary of fit and test data sets.

Method (LV)*	Data	α1/α2	SEN_{Train}	SEN_{Test}	SPC_{Test}	Eff_{Test}
OC-PLS (1)	*Block 1*		0.98	0.97	0.52	0.79
OC-PLS (2)	*Block 2*		1.00	0.86	0.42	0.68
OC-PLS (4)	*Row-wise*	0.01	0.99	0.49	0.93	0.66
SO-OC-PLS (2/1)	***Multiblock***		**0.99**	**0.88**	**0.96**	**0.90**
OC-PLS (1)	*Block 1*		0.89	0.94	0.57	0.79
OC-PLS (2)	*Block 2*		0.95	0.75	0.52	0.66
OC-PLS (4)	*Row-wise*	0.05	0.91	0.30	0.98	0.57
SO-OC-PLS (2/1)	***Multiblock***		**0.93**	**0.88**	**0.96**	**0.90**
OC-PLS (1)	*Block 1*		0.85	0.86	0.60	0.75
OC-PLS (2)	*Block 2*		0.89	0.66	0.56	0.62
OC-PLS (4)	*Row-wise*	0.10	0.79	0.20	0.98	0.51
SO-OC-PLS (2/1)	***Multiblock***		**0.87**	**0.77**	**0.98**	**0.85**

*Latent variables.

The proposed method was also performed with respect to simulated data with different noise levels (Table 2S) and different overlapping degrees between target and non-target classes (Table 3S). In the first scenario, three noise levels were used, 1%, 5% and 15%, keeping the level of overlap between target samples constant and not exceeding 10% and with the acceptance area delimited to 95% of statistical confidence. Between 1% and 5%, the method proved to be stable, exhibiting an efficiency of classification close to the level of overlap between classes. When the data showed high residuals, however, a drop in the effectiveness of the proposed method was observed. On the other hand, when the noise level was kept at 1% and different degrees of overlap (10%, 20% and 30%) were adopted, it was observed that the method was able to classify the samples with the expected effectiveness.

4.2. Wine classification

Tokaj wines are a very special type of sweet botrytized wine produced in the Tokaj region of Hungary and Slovakia. In addition, it is one of the most famous Slovak commodities with protected designation of origin. Thus, in this case study, the proposed method was applied to differentiate Slovak Tokaj wines. The target class included the most expensive and rare types of Tokaj wine called Tokaj Selection 2- to 6-putňa index and Tokaj Essence. On the other hand, the non-target samples included in the test set were Slovak Tokaj wines of lower commercial value with the intent to imitate possible fraudulent scenarios (Furdíková et al., 2021).

In order to achieve high classification efficiency, UV–VIS and IR spectra were combined (Fig. 2a). The UV–VIS absorption spectra convey information related to organic compounds such as polyphenols originated from the grapes and subsequent fermentation or aging processes. On the other hand, IR spectra reflect important information about compounds like sugars, glycerol and many different organic compounds such as organic acids, various alcohols, esters, carbonyl compounds, and terpenes (Gomes et al., 2021; Machyňáková et al., 2021). In order to assess synergy between the two data sets, the UV–VIS and IR blocks were treated by PCA decomposition separately. The results are shown in Fig. 2b and 2c. The most important information is that both data sets contain chemical information able to distinguish the target class from other types of Tokaj wine. In both cases, however, some degree of

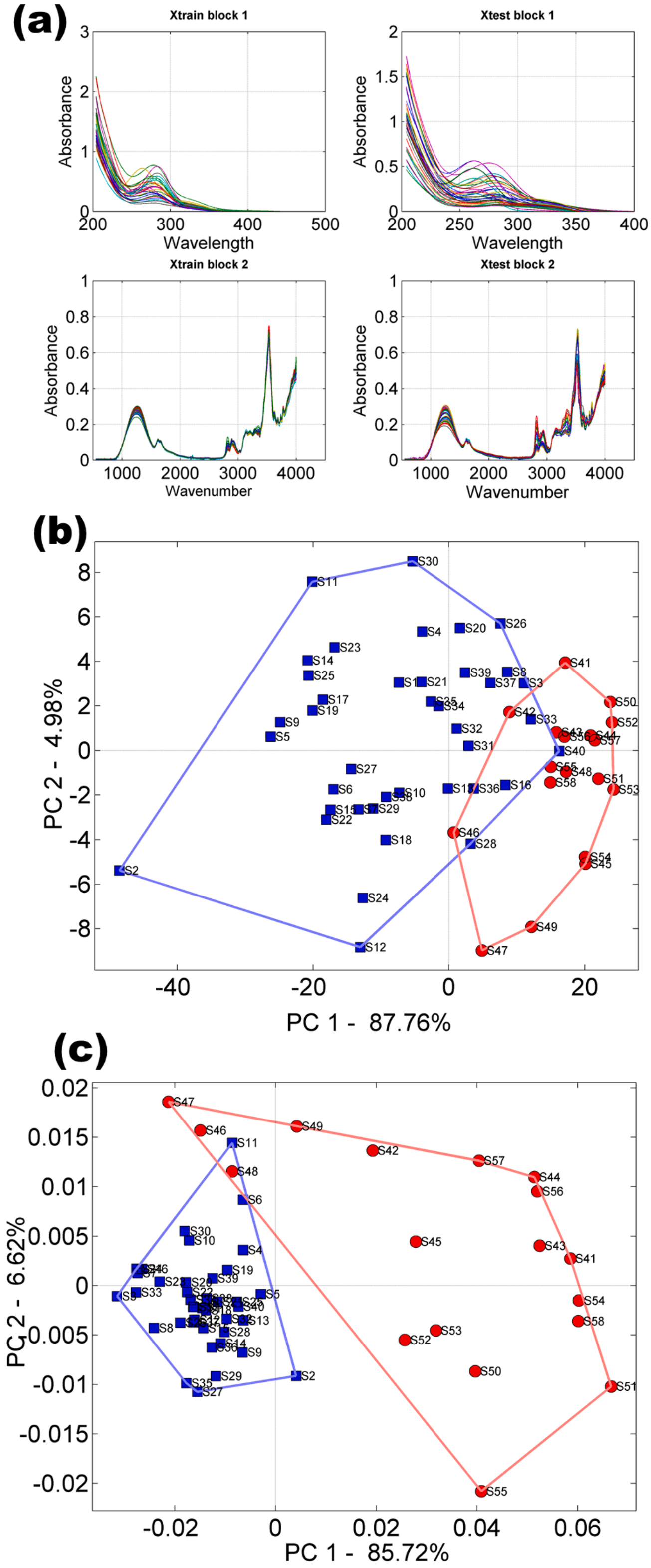

Fig. 2. Wine classification (a) data sets: at the top, upper side, the UV–Vis spectra are displayed, on the bottom, lower side the infrared spectra are displayed. The training and test sets are on the right and left sides, respectively. PCA score plots for (b) UV–Vis and (c) IR raw spectra, derivative spectra used to build the models can be viewed in Supplementary Material Fig. 3S. The numbering of samples is the same in all figures.

overlap between the classes is visible. Each type of data shows overlap between different samples indicating that their combination in one multiblock could improve results.

In previous our work, UV spectra (Gomes et al., 2021) were used to distinguish Slovak Tokaj wine samples from those of Hungarian and Ukrainian origin, while IR spectra (Machyňáková et al., 2021) made it possible to classify Slovak samples into ordinary wines and special selections. However, in both cases, samples of the type Tokaj essence were misclassified.

Afterward individual blocks, row-wise augmented matrix and multiblock structure required by proposed method underwent LOOCV treatment, in order to access the appropriate number of latent variables, as can be seen in Fig. 3. Ten and eight latent variables were selected for individual blocks respectively. When both blocks were processed together, fewer factors were required to minimize the standard deviation of residuals in cross validation. When the data were processed in the row-wise agumnetated structure, only 4 latent variables were required; and finally only 3 (1 for UV data and 2 IR data) were used in our proposed method. As mentioned earlier, SO-OC-PLS is based on a sequentially orthogonalization process and therefore the order of the blocks influences the results. Both possibilities were evaluated, but UV–VIS was chosen as block 1 and IR as block 2, considering the increased complexity of the information in each data set. Opposite block order showed only 67% of classification efficiency for predicting the test set (both α_1 and $\alpha_2 = 0.05$).

Fig. 3a and 3b confirm that a large number of latent variables are needed to minimize standard deviation of residuals in cross validation if individual blocks are used. However, when row-wise augmented (Fig. 3c) and multiblock orthogonally sequenced data (Fig. 3d) were processed by OC-PLS, the number of latent variables was significantly reduced. This observation supports our idea that both data sets were able to improve the classification results.

The distances *acr* and *SD* were calculated, and the limits of the acceptance area were estimated for each scenario considering both α 0.05. If the UV–VIS and IR blocks were treated separately, all samples of the training set would fall within the acceptance area, reaching maximum sensitivity (Fig. 4a and 4b). The models involving both blocks (Fig. 4c and 4d) were more parsimonious in a number of latent variables, which resulted in a decrease of sensitivity in the training set. However, in the proposed method, it is important to note that the reduced number of latent variables provide consistent results, in terms of *a prior* (1-α) and estimated sensitivity are close each other, consequently indicating absence of both over and under fittings.

In order to investigate stability of the proposed algorithm, various α values were tested and results are shown at Table 2. The models remained stable with high sensitivity to values of α equal to 0.01 and 0.05 in the training stage; however, as expected, a decrease in sensitivity was observed for α equal to 0.1. All models were considered well fitted; and they were used to predict an independent test set composed by both target and non-target samples in order to validate them. Table 2 shows high sensitivity for all cases regardless of confidence levels used. The position of the test samples in relation to the acceptance area is shown in Fig. 4S. Completely different results were observed for specificity. The data for UV–VIS (Fig. 4Sa) achieved a fairly reasonable specificity for α of 0.1, while maintaining high sensitivity.

IR data (Fig. 4Sb) and row-wise augmented data structure (Fig. 4Sc) show similar results, however the UV–VIS and IR data treated together showed similar result using only 2 variables. Sole IR data achieved similar efficiency with 8 latent variables. If only 2 latent variables were used for IR data, a dramatic drop in efficiency would be observed. This conclusion reinforces the idea that both blocks carry information capable of distinguishing the target class from the other samples; treating them as a single set of data, however, did not enable exploration of overall synergy between them.

On the other hand, when the data blocks were treated adequately by discarding redundant information in the orthogonalization stage and

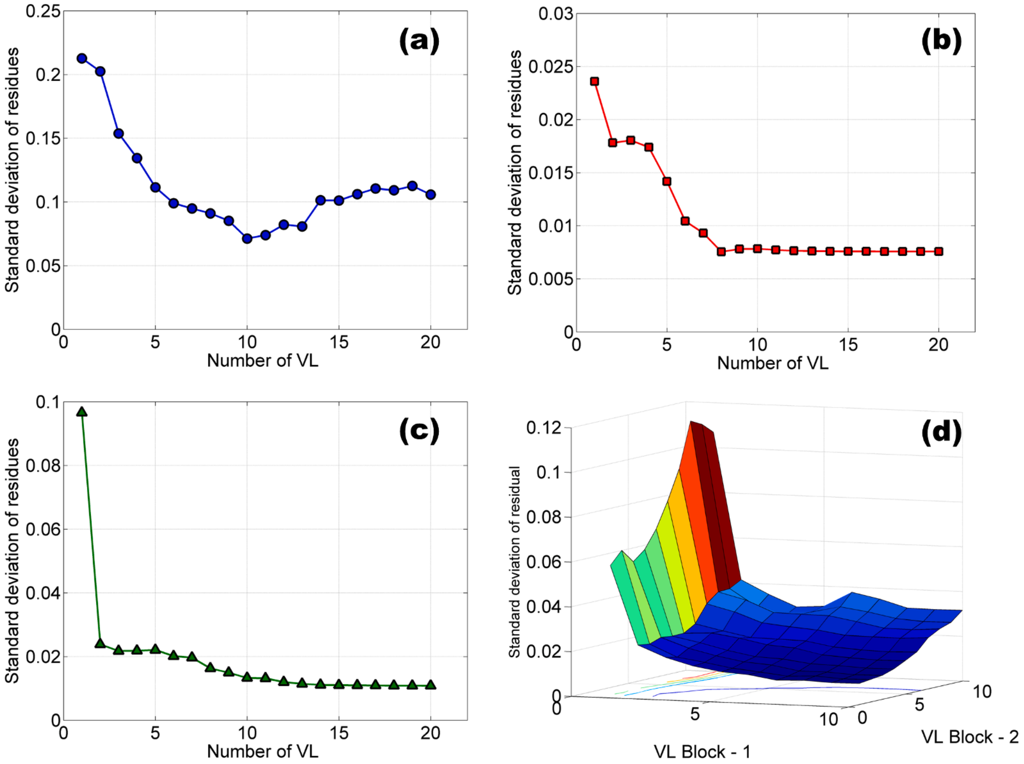

Fig. 3. Standard deviation of residuals in cross validation (a) for UV–VIS; (b) IR; (c) row-wise augmented data; (d) the proposed method.

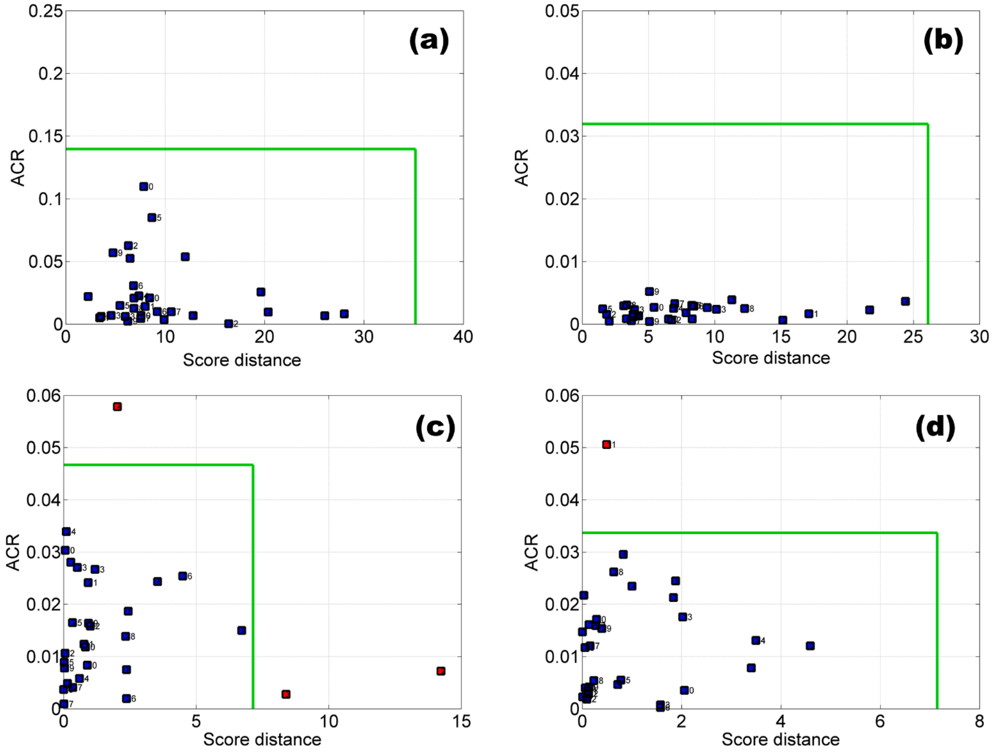

Fig. 4. Acceptance area considering α1 (*SD*) = 0.05 and α2 (*acr*) = 0.05 for (a) UV–Vis, (b) IR, (c) row-wise augmented and (d) proposed method. Blue squares are the true positives and red squares are the false negatives. The green lines represent the limits of the acceptance area. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

only the distinct information included in the final model, a good compromise between sensitivity and selectivity was achieved in the test set (Fig. 4Sd). When α equal to 0.05 is selected, the sensitivity reaches highest values, and the specificity reaches 0.94. Going even further, for α equal to 0.1, the proposed method is able to adequately exploit the useful information for target and non-target class reaching 100 %

Table 2
Statistical summary of fit and test for Slovak Tokaj selection classification.

Method (VL)*	Data	α1/ α2	$\text{SEN}_{\text{Train}}$	SEN_{Test}	SPC_{Test}	Eff_{Test}
OC-PLS (10)	*UV*		1.00	1.00	0.06	0.39
OC-PLS (8)	*IR*		1.00	1.00	0.77	0.85
OC-PLS (2)	*Row-wise*	0.01	0.96	1.00	0.77	0.85
SO-OC-PLS (1/ 2)	***Multiblock***		**0.97**	**1.00**	**0.89**	**0.93**
OC-PLS (10)	*UV*		1.00	1.00	0.06	0.39
OC-PLS (8)	*IR*		1.00	1.00	0.88	0.92
OC-PLS (2)	*Row-wise*	0.05	0.90	0.90	0.88	0.89
SO-OC-PLS (1/ 2)	***Multiblock***		**0.97**	**1.00**	**0.94**	**0.96**
OC-PLS (10)	*UV*		1.00	1.00	0.40	0.61
OC-PLS (8)	*IR*		0.93	1.00	0.88	0.92
OC-PLS (2)	*Row-wise*	0.10	0.86	0.90	0.88	0.89
SO-OC-PLS (1/ 2)	***Multiblock***		**0.93**	**1.00**	**1.00**	**1.00**

*Latent variables.

specificity and sensitivity in the test set (see Table 2). Thus, even Tokaj essence type samples could be properly classified.

5. Conclusions

In this work, one class model based on PLS regression coupled to sequential orthogonalization was presented as a new method to deal with multiblock data processing for food integrity issues. The proposed method was evaluated in a simulated data case study and also the use of UV–VIS and IR data fusion was explored for the classification of Slovak Tokaj Selection wines. In both cases, the proposed method improved the results in relation to a general disregard for the structure of multiblocks. Properly treated multiblocks have been reported to significantly improve food classification. In the proposed method, the orthogonalization of a block in relation to the others is used to discard the common and retain only distinct information. This permitted estimation of an acceptance area with high efficiency in the prediction step.

CRediT authorship contribution statement

Adriano A. Gomes: Data curation, Methodology, Writing – original draft, Writing – review & editing. **Liudmyla Khvalbota:** Investigation, Methodology, Writing – original draft. **Larisa Onça:** Data curation, Writing – original draft, Writing – review & editing. **Andrea Machyňáková:** Investigation, Methodology, Writing – original draft. **Ivan Špánik:** Conceptualization, Writing – original draft, Writing – review & editing, Funding acquisition, Supervision, Project administration.

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

We would like to thank to Grant Agency of the Ministry of Education of Slovak Republic for contract VEGA 1/0521/19. Great thanks belong to Tokaj wineries Ostrožovič spol. s.r.o., Anna Nagyová – Zlatý Strapec and Tokaj & Co., s.r.o. for the cooperation and granting the samples. Brazilian authors are grateful to Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES (Finance code 001), Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq and National Scholarship Programme of the Slovak Republic-NPS (Adriano A. Gomes's CNPq grants 438734/2018-1 and NPS scholarship).

The English text of this paper has been revised by Sidney Pratt, Canadian, MAT (The Johns Hopkins University), RSAdip – TESL (Cambridge University).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.foodchem.2022.132271.

References

Alamar, P. D., Caramês, E. T. S., Poppi, R. J., & Pallone, J. A. L. (2020). Detection of fruit pulp adulteration using multivariate analysis: Comparison of NIR, MIR and data fusion performance. *Food Analytical Methods, 13*(6), 1357–1365. https://doi.org/10.1007/s12161-020-01755-x

Araújo, A. S., Castro, J. P., Sperança, M. A., Andrade, D. F., de Mello, M. L., & Pereira-Filho, E. R. (2021). Multiway Calibration Strategies in Laser-Induced Breakdown Spectroscopy: A Proposal. *Analytical Chemistry, 93*(16), 6291–6300. https://doi.org/10.1021/acs.analchem.0c04722

Arndt, M., Rurik, M., Drees, A., Ahlers, C., Feldmann, S., Kohlbacher, O., & Fischer, M. (2021). Food authentication: Determination of the geographical origin of almonds (*Prunus dulcis* Mill.) via near-infrared spectroscopy. *Microchemical Journal, 160*, 105702. https://doi.org/10.1016/j.microc.2020.105702

Azcarate, S. M., Ríos-Reina, R., Amigo, J. M., & Goicoechea, H. C. (2021). Data handling in data fusion: Methodologies and applications. *Trends in Analytical Chemistry, 143*, 116355. https://doi.org/10.1016/j.trac.2021.116355

Ballabio, D. (2015). A MATLAB toolbox for Principal Component Analysis and unsupervised exploration of data structure. *Chemometrics and Intelligent Laboratory Systems, 149*, 1–9. https://doi.org/10.1016/j.chemolab.2015.10.003

Biancolillo, A., Måge, I., & Næs, T. (2015). Combining SO-PLS and linear discriminant analysis for multi-block classification. *Chemometrics and Intelligent Laboratory Systems, 141*, 58–67. https://doi.org/10.1016/j.chemolab.2014.12.001

Borrás, E., Ferré, J., Boqué, R., Mestres, M., Aceña, L., Busto, Olga. Data fusion methodologies for food and beverage authentication and quality assessmente – A review. (2015). Analytica Chimica Acta 891, 1-14. https://doi.org/10.1016/j.aca.2015.04.042.

Brereton, R. G. (2011). One-class classifiers. *Journal of Chemometrics, 25*, 225–246. https://doi.org/10.1002/cem.1397

Campos, M. P., Sousa, R., & Reis, M. S. (2018). Establishing the optimal blocks' order in SO-PLS: StepwiseSO-PLS and alternative formulations. *Journal of Chemometrics., 32* (8), e3032. https://doi.org/10.1002/cem.v32.810.1002/cem.3032

Campos, M. P., Reis, M. S. (2020). Data preprocessing for multiblock modelling – A systematization with new methods. Chemometrics and Intelligent Laboratory Systems 199, 103959. https:/ /doi.org/10.1016/j.chemolab.2020.1 03959.

Cariou, V., Qannari, E. M., Rutledge, D. N., & Vigneau, E. (2018). ComDim: From multiblock data analysis to path modeling. *Food Quality and Preference, 67*, 27–34. https://doi.org/10.1016/j.foodqual.2017.02.012

Lima, C. M.d., Fernandes, D. D. S., Pereira, G. E., Gomes, A.d. A., Araújo, M. C. U.d., & Diniz, P. H. G. D. (2020). Digital image-based tracing of geographic origin, winemaker, and grape type for red wine authentication. *Food Chemistry, 312*, 126060. https://doi.org/10.1016/j.foodchem.2019.126060

Furdíková, K., Khvalbota, L., Machyňáková, A., & Špánik, I. (2021). Volatile composition and enantioselective analysis of chiral terpenoids in Tokaj varietal wines. *Journal of Chromatography B, 1167*, 122565. https://doi.org/10.1016/j.jchromb.2021.122565

Gamela, R. R., Costa, V. C., Sperança, M. A., & Pereira-Filho, E. R. (2020). Laser-induced breakdown spectroscopy (LIBS) and wavelength dispersive X-ray fluorescence (WDXRF) data fusion to predict the concentration of K, Mg and P in bean seed samples. *Food Research International, 132*, 109037. https://doi.org/10.1016/j.foodres.2020.109037

Gomes, A. A., Khvalbota, L., Machyňáková, A., Furdíková, K., Zini, C. A., & Špánik, I. (2021). Slovak Tokaj wines classification with respect to geographical origin by means of one class approaches. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 257*, 119770. https://doi.org/10.1016/j.saa.2021.119770

Jurado-Campos, N., Arroyo-Manzanares, N., Viñas, P., & Arce, L. (2020). Quality authentication of virgin olive oils using orthogonal techniques and chemometrics based on individual and high-level data fusion information. *Talanta, 219*, 121260. https://doi.org/10.1016/j.talanta.2020.121260

López, M. I., Colomer, N., Ruisánchez, I., & Callao, M. P. (2014). Validation of multivariate screening methodology. Case study: Detection of food fraud. *Analytica Chimica Acta, 827*, 28–33. https://doi.org/10.1016/j.aca.2014.04.019

Machyňáková, A., Schneider, M. P., Khvalbota, L., Vyviurska, O., Špánik, I., & Gomes, A. A. (2021). A fast and inexpensive approach to characterize Slovak Tokaj selection wines using infrared spectroscopy and chemometrics. *Food Chemistry, 357*, 129715. https://doi.org/10.1016/j.foodchem.2021.129715

Mishra, P., Roger, J. M., Rutledge, D. N., Biancolillo, A., Marini, F., Nordon, A., & Jouan-Rimbaud-Bouveresse, D. (2020). MBA-GUI: A chemometric graphical user interface for multi-block data visualisation, regression, classification, variable selection and automated pre-processing. *Chemometrics and Intelligent Laboratory Systems, 205*, 104139. https://doi.org/10.1016/j.chemolab.2020.104139

Montgomery, H., Haughey, S. A., & Elliott, C. T. (2020). Recent food safety and fraud issues within the dairy supply chain (2015–2019). *Global Food Security, 26*, 100447. https://doi.org/10.1016/j.gfs.2020.100447

Næs, T., Tomic, O., Afseth, N. K., Segtnan, V., & Måge, I. (2013). Multi-block regression based on combinations of orthogonalisation, PLS-regression and canonical correlation analysis. *Chemometrics and Intelligent Laboratory Systems, 124*, 32–42. https://doi.org/10.1016/j.chemolab.2013.03.006

Oliveri, P., & Downey, G. (2012). Multivariate class modeling for the verification of food-authenticity claims. *TrAC Trends in Analytical Chemistry, 35*, 74–86. https://doi.org/10.1016/j.trac.2012.02.005

Oliveri, P., Malegori, C., Mustorgi, E., & Casale, M. (2021). Qualitative pattern recognition in chemistry: Theoretical background and practical guidelines. *Microchemical Journal, 162*, 105725. https://doi.org/10.1016/j.microc.2020.105725

Rodionova, O. Y., Oliveri, P., & Pomerantsev, A. L. (2016). Rigorous and compliant approaches to one-class classification. *Chemometrics and Intelligent Laboratory Systems, 159*, 89–96. https://doi.org/10.1016/j.chemolab.2016.10.002

Pinto, L., Nieto, C. H. D., Zón, M. A., Fernández, H., & de Araújo, M. C. U. (2016). Handling time misalignment and rank deficiency in liquid chromatography by multivariate curve resolution: Quantitation of five biogenic amines in fish. *Analytica Chimica Acta, 902*, 59–69. https://doi.org/10.1016/j.aca.2015.10.043

Pomerantsev, A. L., & Rodionova, O. Y. (2013). Concept and role of extreme objects in PCA/SIMCA. *Journal of Chemometrics, 28*, 429–438. https://doi.org/10.1002/cem.2506

Qi, J., Li, Y., Zhang, C., Wang, C., Wang, J., Guo, W., & Wang, S. (2021). Geographic origin discrimination of pork from different Chinese regions using mineral elements analysis assisted by machine learning techniques. *Food Chemistry, 337*, 127779. https://doi.org/10.1016/j.foodchem.2020.127779

Reile, C. G., Rodríguez, M. S., Fernandes, D. D.d. S., Gomes, A.d. A., Diniz, P. H. G. D., & Di Anibal, C. V. (2020). Qualitative and quantitative analysis based on digital images to determine the adulteration of ketchup samples with Sudan I dye. *Food Chemistry, 328*, 127101. https://doi.org/10.1016/j.foodchem.2020.127101

Ríos-Reina, R., Azzcarate, S. N., Camiña, J. M., & Goicoechea, H. C. (2020). Multi-level data fusion strategies for modeling three-way electrophoresis capillary and fluorescence arrays enhancing geographical and grape variety classification of wines. *Analytica Chimica Acta, 1126*, 52–62. https://doi.org/10.1016/j.aca.2020.06.014

Robson, K., Dean, M., Haughey, S., & Elliott, C. (2021). A comprehensive review of food fraud terminologies and food fraud mitigation guides. *Food Control, 120*, 107516. https://doi.org/10.1016/j.foodcont.2020.107516

Rodionova, O. Y., Titova, A. V., & Pomerantsev, A. L. (2016). Discriminant analysis is an inappropriate method of authentication. *TrAC Trends in Analytical Chemistry, 78*, 17–22. https://doi.org/10.1016/j.trac.2016.01.010

Sandler, C. R., Grassby, T., Hart, K., Raats, M., Sokolović, M., & Timotijevic, L. (2021). Processed food classification: Conceptualisation and challenges. *Trends in Food Science & Technology, 112*, 149–162. https://doi.org/10.1016/j.tifs.2021.02.059

Schouteden, M., Van Deun, K., Pattyn, S., & Van Mechelen, I. (2013). SCA with rotation to distinguish common and distinctive information in linked data. *Behavior Research Methods, 45*(3), 822–833. https://doi.org/10.3758/s13428-012-0295-9

Smilde, A. K., Måge, I., Naes, T., Hankemeier, T., Lips, M. A., Kiers, H. A. L., … Bro, R. (2017). Common and distinct components in data fusion. *Journal of Chemometrics, 31*(7), e2900. https://doi.org/10.1002/cem.v31.710.1002/cem.2900

Smilde, A. K., Westerhuis, J. A., & de Jong, S. (2003). A framework for sequential multiblock component methods. *Journal of Chemometrics, 17*(6), 323–337. https://doi.org/10.1002/(ISSN)1099-128X10.1002/cem.v17:610.1002/cem.811

Trygg, J., & Wold, S. (2003). O2-PLS, a two-block (X-Y) latent variable regression (LVR) method with an integral OSC filter. *Journal of Chemometrics, 17*(1), 53–64. https://doi.org/10.1002/(ISSN)1099-128X10.1002/cem.v17:110.1002/cem.775

Ulberth, F. (2020). Tools to combat food fraud – A gap analysis. *Food Chemistry, 330*, 127044. https://doi.org/10.1016/j.foodchem.2020.127044

Xie, J., Pan, Q., Li, F., Tang, Y., Hou, S., & Xu, C. (2021). Simultaneous detection of trace adulterants in food based on multi-molecular infrared (MM-IR) spectroscopy. *Talanta, 222*, 121325. https://doi.org/10.1016/j.talanta.2020.121325

Xu, L.u., Cai, C.-B., & Deng, D.-H. (2011). Multivariate quality control solved by one-class partial least squares regression: Identification of adulterated peanut oils by mid-infrared spectroscopy. *Journal of Chemometrics, 25*(10), 568–574. https://doi.org/10.1002/cem.v25.1010.1002/cem.1402

Xu, L., Goodarzi, M., Shi, W., Cai, C.-B., & Jiang, J.-H. (2014). A MATLAB toolbox for class modeling using one-class partial least squares (OCPLS) classifiers. *Chemometrics and Intelligent Laboratory Systems, 139*, 58–63. https://doi.org/10.1016/j.chemolab.2014.09.005

Xu, L., Yan, S.-M., Cai, C. B., & Yu, X. P. (2013). One-class partial least squares (OCPLS) classifier. *Chemometrics and Intelligent Laboratory Systems, 126*, 1–5. https://doi.org/10.1016/j.chemolab.2013.04.008