Analytica Chimica Acta xxx (2017) 1–10

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Selection of robust variables for transfer of classification models employing the successive projections algorithm

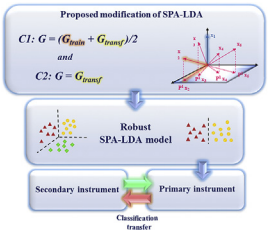
Karla Danielle Tavares Melo Milanez [a], Thiago César Araújo Nóbrega [a], Danielle Silva Nascimento [b], Roberto Kawakami Harrop Galvão [c], Márcio José Coelho Pontes [a, *]

[a] Departamento de Química, Universidade Federal da Paraíba, João Pessoa, PB, Brazil
[b] Laboratorio FIA, INQUISUR-CONICET, Departamento de Química, Universidad Nacional del Sur, Av. Alem 1253, B8000CPB Bahía Blanca, Buenos Aires, Argentina
[c] Instituto Tecnológico de Aeronáutica, Divisão de Engenharia Eletrônica, São José dos Campos, São Paulo 12228-900, Brazil

HIGHLIGHTS

- Two new criteria for selection of robust variables for classification transfer.
- Use of linear discriminant analysis coupled to successive projections algorithm.
- Identification of adulteration in extra virgin olive oil and alcohol fuel.
- Comparison of the proposed criteria with DS and PDS methods.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 17 February 2017
Received in revised form
4 July 2017
Accepted 17 July 2017
Available online xxx

Keywords:
Multivariate classification transfer
Robust modeling
Successive projections algorithm
Standardization methods
UV–Vis spectroscopy
NIR spectroscopy

ABSTRACT

Multivariate models have been widely used in analytical problems involving quantitative and qualitative analyzes. However, there are cases in which a model is not applicable to spectra of samples obtained under new experimental conditions or in an instrument not involved in the modeling step. A solution to this problem is the transfer of multivariate models, usually performed using standardization of the spectral responses or enhancement of the robustness of the model. This present paper proposes two new criteria for selection of robust variables for classification transfer employing the successive projections algorithm (SPA). These variables are then used to build models based on linear discriminant analysis (LDA) with low sensitivity with respect to the differences between the responses of the instruments involved. For this purpose, transfer samples are included in the calculation of the cost for each subset of variables under consideration. The proposed methods are evaluated for two case studies involving identification of adulteration of extra virgin olive oil (EVOO) and hydrated ethyl alcohol fuel (HEAF) using UV–Vis and NIR spectroscopy, respectively. In both cases, similar or better classification transfer results (obtained for a test set measured on the secondary instrument) employing the two criteria were obtained in comparison with direct standardization (DS) and piecewise direct standardization (PDS). For the UV–Vis data, both proposed criteria achieved the correct classification rate (CCR) of 85%, while the best CCR obtained for the standardization methods was 81% for DS. For the NIR data, 92.5% of CCR was obtained by both criteria as well as DS. The results demonstrated the possibility of using either of the

* Corresponding author. Universidade Federal da Paraíba, Departamento de Química – Laboratório de Automação e Instrumentação em Química Analítica/Quimiometria (LAQA), CEP 58051-970 João Pessoa, PB, Brazil.
E-mail address: marciocoelho@quimica.ufpb.br (M.J.C. Pontes).

http://dx.doi.org/10.1016/j.aca.2017.07.037

criteria proposed for building robust models as an alternative to the standardization of spectral responses for transfer of classification.

1. Introduction

Multivariate models have been widely used in analytical problems involving quantitative and qualitative analyzes of a variety of matrices [1–10]. Besides the well established multivariate methods in the literature, new chemometric tools are constantly being proposed in order to improve information acquisition and provide increasingly robust multivariate models.

The efficiency of the multivariate models is associated with the quantity of samples analyzed (as well as their chemical/physical composition and representativeness), instrumental used and laboratory conditions [11]. In addition, the actual construction and validation steps of the model are also fundamental for achieving a good performance [12]. However, even though these parameters are carefully controlled, there are cases in which a model is not applicable to spectra of samples obtained under new experimental conditions or in an instrument not involved in the modeling step. The differences between the spectral responses may be related to changes in the chemical and/or physical composition of the samples or changes in the instrumental response, normally caused by maintenance irregularities, repairs, changes in the environment of the instrument or even natural wear [13].

A solution to this problem is the transfer of multivariate models, usually performed using the following strategies: standardization of the spectral responses or enhancement of the robustness of the model [14].

In the standardization methods, the response of a secondary instrument is adapted to correspond to the response from a primary instrument, for which the model was developed [13]. The standardization methods include Direct Standardization (DS) and Piecewise Direct Standardization (PDS) [15]. In both cases, spectra of a representative set of samples, known as *transfer samples*, are recorded in the instruments involved and used to built the standardized models [13]. In DS, the mathematical manipulations are carried out along the entire spectral range, while in the PDS, the manipulations are performed in windows of variables within the spectrum [12].

Robustness is one of the parameters most used to evaluate the performance of multivariate models. In both quantitative and qualitative analyzes, robustness is associated with the ability of a model to provide reliable results against variations not included in the modeling [16]. The robustness of multivariate models can be achieved by incorporating all the important sources of variation still in the modeling step, which can be obtained by analyzing the samples in different instruments or under different experimental conditions [17]. Besides this, pre-processing the data (which eliminates unrelated variations of the properties of interest) and the selection of variables less sensitive to experimental conditions or instrumental variations can also be used to increase the robustness of a model, eliminating the need for standardization when the differences between the response functions are relatively small [17,18].

Several studies have been reported in the literature using transfer of calibration models [11,19–25]. However, few studies have been directed towards classification transfer problems. Myles et al. [26] investigated different strategies of transfer of classification models developed for discrimination of coffee beans analyzed by NIR spectroscopy. Di Anibal et al. [27] applied piecewise direct standardization (PDS) to UV–Vis spectra to discriminate different culinary spices adulterated. Silva et al. [28] used direct standardization (DS) method to transfer MIR spectra of gasoline samples recorded on three different instruments. Milanez et al. [29] used DS and PDS standardization methods for NIR spectra employed in the identification of adulteration of fuel ethanol samples. In that work, linear discriminant analysis (LDA) [30] models, with previous selection of variables by the SPA [31] algorithm, and partial least squares - discriminant analysis (PLS-DA) [32] were developed and the classification results were evaluated in terms of the correct classification rate before and after standardization, where a substantial increase in model performance was observed.

In this context, this paper proposes two new criteria for the selection of robust variables for transfer classification employing the successive projections algorithm (SPA). SPA have been successfully employed, coupled to linear discriminant analysis (LDA), in problems of classification involving instrumental techniques [33–37], mainly to minimize collinearity problems of the LDA model. Recently Soares et al. [38] proposed a new validation criterion for SPA based on the cost function adaptation, for situations in which the quantity of sample is limited. In cases involving classification transfer, the selection of robust variables may be of value for the construction of LDA models with low sensitivity with respect to the differences between the instruments under consideration. In this paper, the increase of the robustness of the model is achieved with the inclusion of transfer samples in the calculation of the cost for each subset of variables under consideration.

Two data sets involving simulated adulteration are employed to evaluate the proposed criteria. The first data set consists of extra virgin olive oil (EVOO) samples (unadulterated and adulterated) analyzed in Brazil (primary instrument) and Argentina (secondary instrument) using UV–Vis spectroscopy. The second data set consists of hydrated ethyl alcohol fuel samples (unadulterated and adulterated). The NIR spectra were recorded under the same experimental conditions in two different spectrometers. In both case studies, the results obtained with the two criteria were compared with the strategies of direct standardization (DS) and piecewise direct standardization (PDS) of the test set measured on the secondary instrument.

2. Background and theory

The Successive Projections Algorithm (SPA) was originally proposed in Ref. [39] in the context of multivariate calibration employing multiple linear regression (SPA-MLR). For this purpose, the instrumental response data were arranged in a matrix **X**, with rows and columns corresponding to the objects and variables, respectively. The choice of suitable variables was then addressed as a selection procedure involving the columns of **X**. Each column was employed to initiate a sequence of projection operations that resulted in the formation of candidate subsets of variables with increasing cardinality (i.e. number of elements). At each iteration of this procedure, the subset was augmented by choosing the column displaying the least collinearity with respect to those selected in the previous iterations. For this purpose, each column was projected onto the subspace spanned by the columns already selected.

Please cite this article in press as: K.D.T.M. Milanez, et al., Selection of robust variables for transfer of classification models employing the successive projections algorithm, Analytica Chimica Acta (2017), http://dx.doi.org/10.1016/j.aca.2017.07.037

Table 1
Number of training and test samples in each class for the two data sets. *N* indicates the number of samples in each class.

Class	EVOO data set			HEAF data set		
	N	Training	Test	*N*	Training	Test
Unadulterated	49	34	15	52	32	20
Adulterated	40	28	12	50	30	20
Total	89	62	27	102	62	40

The projected column with largest norm was then selected. By restarting this procedure from each column of **X**, different subsets of variables were obtained, with cardinality varying from one up to the number of rows (if the columns were not mean-centered) or up to the number of rows minus one (if the columns were mean-centered). The best subset was selected on the basis of the resulting root-mean-square error of prediction in a separate validation data set (*RMSEV*). This cost metric was calculated as

$$RMSEV = \sqrt{\frac{1}{K_{val}} \sum_{k=1}^{K_{val}} \left(y_{val,k} - \hat{y}_{val,k} \right)^2} \quad (1)$$

where $y_{val,k}$ and $\hat{y}_{val,k}$ denote the reference and predicted values of the parameter under consideration in the kth validation object and K_{val} is the overall number of validation objects. An in-depth description of this algorithm, with a review of applications in the analytical chemistry literature, can be found in Ref. [40]. Additional details, including the mathematical expressions employed in the projection operations and the associated computational code, can be found in Ref. [41].

In a subsequent paper [31] SPA was adapted for the selection of variables in classification problems using linear discriminant analysis (SPA-LDA). In this case, the problem consists of assigning each given object to one of C possible classes. The basic difference with respect to the SPA-MLR formulation consists of the criterion employed in the selection of the best candidate subset of variables. In SPA-LDA, it is assumed that each object belongs to a known class with index Ik ($k = 1, 2, \ldots, C$). The sample mean $\mathbf{m}_{Ik}$ of each class and a pooled covariance matrix **S** are calculated, for each candidate subset of variables, by using the available training data set, as in the standard LDA procedure [42]. The risk of misclassification of the kth validation object can then measured by using the following expression:

$$g_{val,k} = \frac{r^2 \left(\mathbf{x}_{val,k}, \boldsymbol{\mu}_{Ik} \right)}{\min_{Ij \neq Ik} r^2 \left(\mathbf{x}_{val,k}, \boldsymbol{\mu}_{Ij} \right)} \quad (2)$$

where $r^2(\mathbf{x}_{val,k}, \boldsymbol{\mu}_{Ik}) = (\mathbf{x}_{val,k} - \mathbf{m}_{Ik})\mathbf{S}^{-1}(\mathbf{x}_{val,k} - \mathbf{m}_{Ik})^T$ is the squared Mahalanobis distance between the kth validation object $\mathbf{x}_{val,k}$ and the sample mean of its true class $\boldsymbol{\mu}_{Ik}$ (both row vectors). The denominator in (2) corresponds to the squared Mahalanobis distance between $\mathbf{x}_{val,k}$ and the center of the nearest wrong class. Ideally the objects should be close to the center of their corresponding classes and distant from the centers of the other classes. Therefore, the criterion employed in SPA-LDA consists of selecting the candidate subset of variables corresponding to the smallest average value of (2), which is calculated as

$$G_{val} = \frac{1}{K_{val}} \sum_{k=1}^{K_{val}} g_{val,k} \quad (3)$$

This metric can be regarded as a cost function to be minimized in the variable selection process.

A alternative that dispenses with the requirement of a separate validation set was proposed in Ref. [38]. The idea consists of evaluating the risk of misclassification (2) by using the same training objects that were employed to calculate the class means $\mathbf{m}_{Ik}$ and the covariance matrix **S**. In order to avoid possible overfitting issues resulting from this repeated use of the training set, the expression (3) is modified as

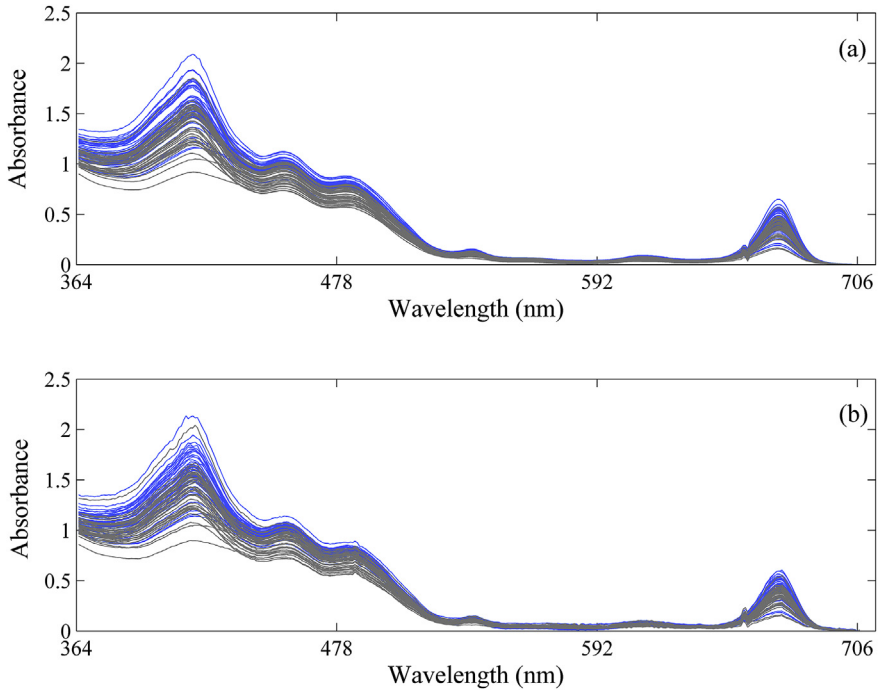


Fig. 1. UV–Vis spectra of the EVOO samples acquired in the (a) primary and (b) secondary instruments. Unadulterated (—) and adulterated (—) samples.

Please cite this article in press as: K.D.T.M. Milanez, et al., Selection of robust variables for transfer of classification models employing the successive projections algorithm, Analytica Chimica Acta (2017), http://dx.doi.org/10.1016/j.aca.2017.07.037

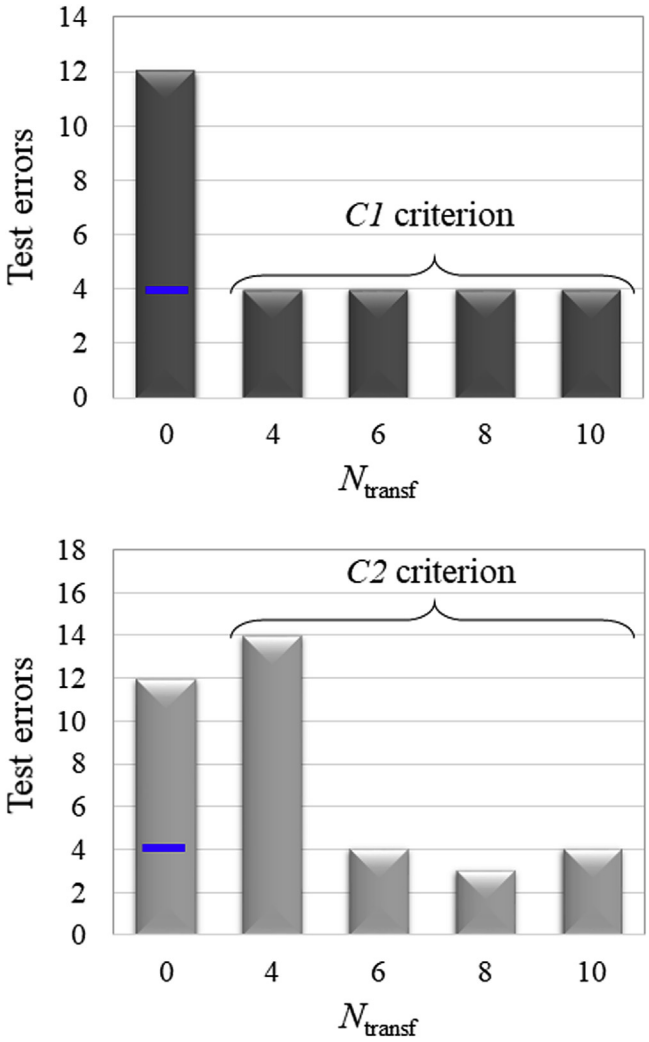


Fig. 2. Results in terms of number of errors obtained by the SPA-LDA models applied to the test samples measured on both instruments. The SPA-LDA model in usual form corresponds to $N_{transf} = 0$. The blue line on the bar corresponding to $N_{transf} = 0$ represents the test (P) errors while the bars represent the test (S) errors. The bars under the braces correspond to the results obtained by using SPA-LDA with the proposed criteria. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

$$G_{train} = \frac{1}{K_{train} - L - C} \sum_{k=1}^{K_{train}} g_{train,k} \tag{4}$$

where L denotes the number of variables in the candidate subset under evaluation, while C and K_{train} correspond to the number of classes in the data set and the total number of training objects, respectively. By doing so, a subset with a large number of variables L will only be selected if the corresponding values of $g_{train,k}$ ($k = 1, 2, \ldots, K_{train}$) are sufficiently small to compensate for the increase in the complexity of the LDA model.

Remark: If the training objects are centered in the mean of the respective classes prior to the projection operations, the candidate subsets of variables generated by SPA-LDA will have one up to ($K_{train} - C$) elements [38]. By using the G_{train} index in (4), candidate subsets with a number L of variables close to this limit will not be favoured, as the denominator in (4) will be close to zero. The underlying rationale consists of avoiding the use of too many variables, which may lead to poor conditioning in the construction of the linear discriminant model. In particular the use of the maximum number of variables that can be selected by SPA-LDA ($L = K_{train} - C$) will be avoided altogether, as the value of G_{train} will be infinitely large.

It is worth noting that both SPA-MLR and SPA-LDA were initially developed without explicit mechanisms for transfer of the resulting models. In Ref. [11], a modified version of SPA-MLR was proposed to enable the selection of variables that were robust with respect to the differences between two instruments. The modification consisted of including a set of K_{transf} transfer samples in the selection of the best candidate subset of variables. The proposed cost metric was defined as

$$E = \frac{1}{2}(RMSEV + RMSET) \tag{5}$$

with *RMSEV* calculated by using validation objects measured in the primary instrument, as in (1) and *RMSET* (root-mean-square error of prediction in the transfer set) defined as

$$RMSET = \sqrt{\frac{1}{K_{transf}} \sum_{k=1}^{K_{transf}} \left(y_{transf,k} - \hat{y}_{transf,k}\right)^2} \tag{6}$$

where $y_{transf,k}$ is the reference value of the parameter under consideration in the kth transfer sample and $\hat{y}_{transf,k}$ denotes the corresponding predicted value obtained by applying the MLR model to the measurements recorded in the secondary instrument. By choosing the candidate subset of variables that leads to the smallest value of (5), both the predictive ability (evaluated by *RMSEV*) and the robustness (evaluated by *RMSET*) of the model are taken into account.

Remark: As pointed out in Ref. [11], the transfer samples employed in the evaluation of the *RMSET* metric (6) only need to be measured at the secondary instrument. This is an important advantage over standardization methods such as DS and PDS, especially if the primary and secondary instruments are not located in the same laboratory, or if the primary instrument is no longer available.

2.1. Proposed modification of SPA-LDA

The modification of SPA-LDA proposed in the present work follows the lines of the SPA-MLR formulation for calibration transfer described above. More specifically, given a set of K_{transf} transfer objects with known class indexes, an average risk of misclassification is calculated as

$$G_{transf} = \frac{1}{K_{transf}} \sum_{k=1}^{K_{transf}} g_{transf,k} \tag{7}$$

where $g_{transf,k}$ is defined as in (2) by using the kth transfer object $\mathbf{x}_{transf,k}$ recorded in the secondary instrument, instead of the kth validation object $\mathbf{x}_{val,k}$. After calculating the values of G_{train} and G_{transf} as in (4) and (7), respectively, for each candidate subset of variables, two possible criteria will be considered here in:

- Criterion *C1*: Select the candidate subset of variables corresponding to the smallest value of the cost $G = (G_{train} + G_{transf})/2$.
- Criterion *C2*: Select the candidate subset of variables corresponding to the smallest value of the cost G_{transf}.

Criterion *C1* is a direct extension of (5) in a classification framework. By using this criterion, the goal consists of obtaining a model with good classification performance on both the primary and secondary instruments. The use of criterion *C2* will be

Please cite this article in press as: K.D.T.M. Milanez, et al., Selection of robust variables for transfer of classification models employing the successive projections algorithm, Analytica Chimica Acta (2017), http://dx.doi.org/10.1016/j.aca.2017.07.037

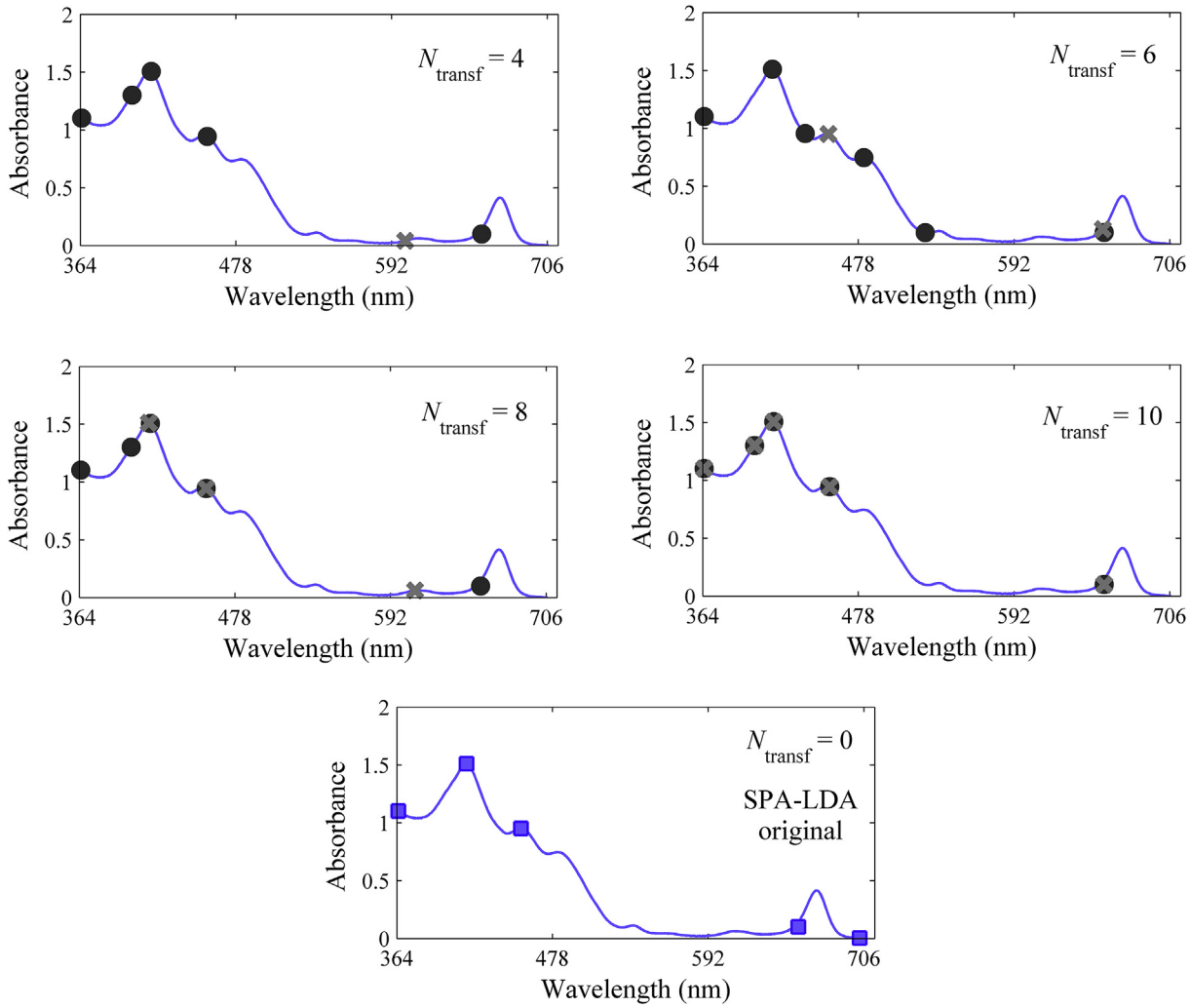


Fig. 3. Average spectrum of the data set with indication of the variables selected by the SPA-LDA algorithm in classification models involving different number of transfer samples. Variables selected by *C1:* ● and *C2:* ✖ criteria.

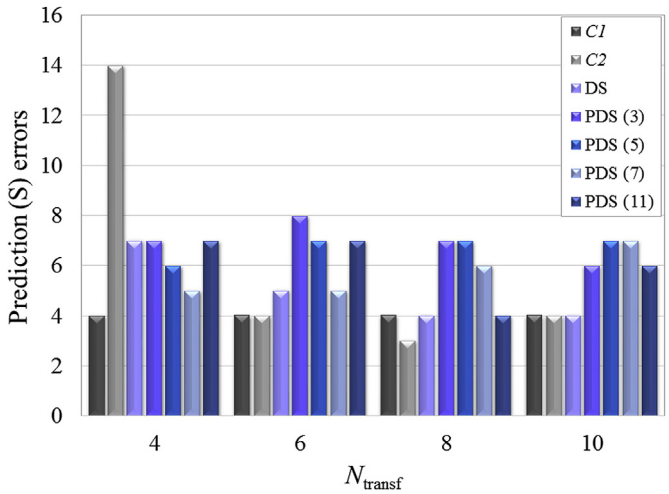


Fig. 4. Results in terms of number of errors obtained by SPA-LDA with the *C1* and *C2* criteria and SPA-LDA in usual form after performing the standardization procedures, applied to the test samples measured by the secondary instrument. The width of the window employed in the PDS standardization method is indicated in parenthesis.

investigated as an alternative that places larger emphasis in the classification accuracy of the transfer samples, in the understanding that the resulting model is to be used only in the secondary instrument.

As in the case of SPA-MLR for calibration transfer [11], the transfer samples employed in the evaluation of the G_{transf} metric (7) only need to be measured at the secondary instrument, which is an advantage over standardization methods such as DS and PDS.

3. Experimental

3.1. Extra virgin olive oil data set

The first data set consists of 89 extra virgin olive oil (EVOO) samples (49 unadulterated samples and 40 adulterated samples). The unadulterated samples were acquired in local commerce with different lots. The manufacturer was chosen based on an investigation performed by a Brazilian Association of Consumer Protection [43], which evaluates the quality of commercially available products (more details regarding this investigation can be found in http://www.proteste.org.br/azeite).

In order to evaluate the authenticity of the unadulterated EVOO samples used in this study, the determination of specific extinction

Please cite this article in press as: K.D.T.M. Milanez, et al., Selection of robust variables for transfer of classification models employing the successive projections algorithm, Analytica Chimica Acta (2017), http://dx.doi.org/10.1016/j.aca.2017.07.037

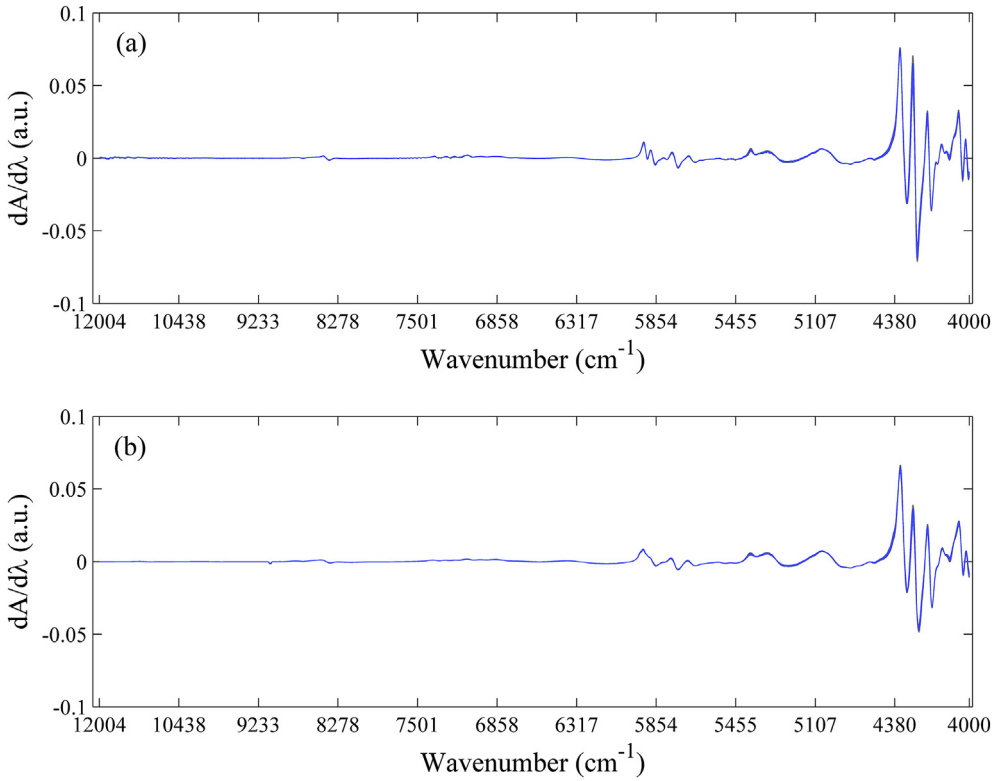


Fig. 5. NIR spectra of the EVOO samples acquired in the (a) primary and (b) secondary instruments. Unadulterated (▬) and adulterated (▬) samples.

(K) by absorption in the ultraviolet region (AOCS Official Method Ch 5–91) was performed using cyclohexane of spectrophotometric grade as solvent, as described in Ref. [44].

The results demonstrated that all unadulterated EVOO samples presented values of specific extinction at the wavelengths of 232 and 270 nm (K_{232} and K_{270}) and specific extinction variation (ΔK) equal or lower than the limits established by International Olive Council [45], reinforcing the results obtained by Ref. [43].

The adulterations were prepared by addition of soybean oil at different levels: 1.0%, 5.0%, 10.0%, 15.0%, 20.0%, 25.0% and 30.0% (w/w). All samples were stored in amber glass bottles, protected from light and kept at a temperature of approximately 23 ± 2 °C until time of analysis. No sample pretreatment was performed.

The EVOO samples were analyzed in Brazil (primary instrument) and Argentina (secondary instrument) using a Hewlett-Packard model HP 8453 UV–VIS spectrophotometer. In both cases, the spectrometer was equipped with a quartz cell (10 mm optical path) and the spectra were recorded in the range 190–1100 nm, with 1 nm resolution. The adjustment of the transmittance signal was performed using isooctane as blank. The spectra measured in both instruments presented a systematic variation of baseline that was corrected with application of baseline offset.

3.2. Ethanol fuel data set

The second data set consists of 102 hydrated ethyl alcohol fuel (HEAF) samples (52 unadulterated samples and 50 adulterated samples). The adulterations were prepared with methanol in the range 1.86–13% (w/w) as described in Ref. [30].

The NIR spectra were recorded in the range 12004–4000 cm^{-1} under the same experimental conditions by using a Spectrum GX FTIR spectrometer (Perkin Elmer), which was considered as the primary instrument, and a dispersive NIR spectrometer (Foss AnalyticalXDS), used as secondary instrument. In order to circumvent the problem of systematic variations in the baseline, first-derivative spectra were calculated using a Savitzky–Golay filter [46] with second-order polynomial and 21-point window.

3.3. Chemometric procedures

The samples analyzed in the primary instrument were divided into training (EVOO data: 70%; HEAF data: 60%) and test (EVOO data: 30%; HEAF data: 40%) sets by using the classic Kennard – Stone (KS) algorithm [47]. Table 1 presents the number of training and test samples in each class for both data sets.

The training sets were employed in the modeling procedures, including SPA variable selection for LDA, whereas the test sets (measured on both instruments) were only used in the final evaluation of the classification models. The LDA models were developed based on the variables selected by the successive projections algorithm (SPA) adapted for internal validation.

The KS algorithm was also used to select subsets of transfer samples from the training set measured on both instruments. Different numbers of transfer samples (4, 6, 8 and 10) were investigated.

The classification models were developed from the SPA-LDA described by Ref. [38] with the introduction of the two criteria proposed herein. As mentioned in section 2.1, these criteria involve the inclusion of the transfer samples in the calculation of the cost for each subset of variables under consideration. The results obtained for the three models (without the use of transfer samples and with each of the two criteria for using the transfer samples) were compared in terms of the number of errors for the test sets

Please cite this article in press as: K.D.T.M. Milanez, et al., Selection of robust variables for transfer of classification models employing the successive projections algorithm, Analytica Chimica Acta (2017), http://dx.doi.org/10.1016/j.aca.2017.07.037

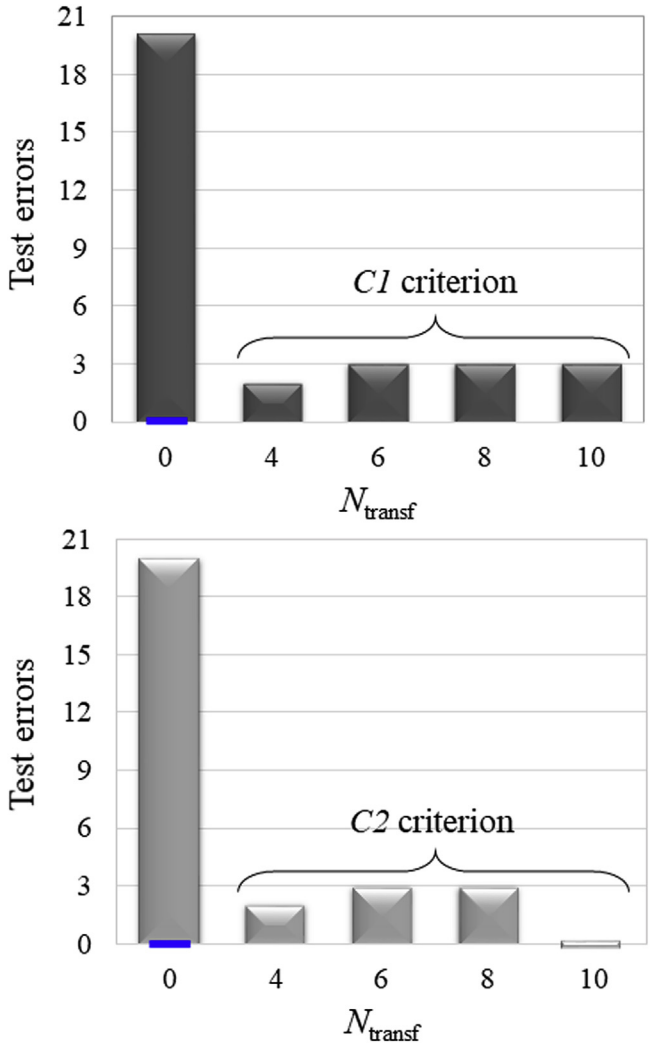


Fig. 6. Results in terms of number of errors obtained by the SPA-LDA models applied to the test samples measured on both instruments. The SPA-LDA model in usual form corresponds to $N_{transf} = 0$. The blue line on the bar corresponding to $N_{transf} = 0$ represents the test (P) errors while the bars represent the test (S) errors. The bars under the braces correspond to the results obtained by using SPA-LDA with the proposed criteria. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and correct classification rate (CCR) for test and training sets.

For the purpose of comparison, two strategies of standardization were employed: direct standardization (DS) and piecewise direct standardization (PDS) of the test set measured on the secondary instrument. The PDS algorithm was run with different window sizes (3, 5, 7 and 11). DS and PDS algorithms were performed using the PLS Toolbox (version 3.5) for Matlab.

In the discussions of the results, the notation "test (P) errors" will be used to represent the test errors obtained by applying the classification model to the primary instrument data, while "test (S) errors" will refer to the test errors obtained by applying the classification model to the secondary instrument data. The proposed criteria will be represented by symbol *C1* (when referring to the cost function involving training and transfer samples) and *C2* (when the cost function considers only the transfer samples).

All calculations were carried out by using the MATLAB® 2010a software.

4. Results and discussion

4.1. Extra virgin olive oil data set

Fig. 1 presents the corrected UV–Vis spectra of EVOO samples acquired in each of the two instruments. As can be seen, the spectra (Fig. 1a and b) present similar profiles. In both cases, it is possible to observe three clearly defined peaks (at approximately 410, 450 and 480 nm) corresponding to carotenoid absorption in the blue range [48]. Those absorptions are associated to electronic transitions from the ground state S_0 to the S_2 state, with energy of the S_2 state depending on the extent of π-electron conjugation of the carotenoids [49]. The peak at approximately 670 nm, near the red range, corresponds to chlorophyll's absorption and involves the electronic transitions between the states S_0 and S_1 [48].

Fig. 2 presents the classification results in terms of the number of test errors obtained by the three classification models (SPA-LDA in usual form, *C1* and *C2*), for different values of N_{transf} (number of transfer samples). The SPA-LDA model in usual form corresponds to $N_{transf} = 0$. The bars indicate the test (S) errors, whereas the blue line on the first bar indicates the test (P) errors. By using the SPA-LDA model in the usual form ($N_{transf} = 0$), the number of test (S) errors is much larger than the number of test (P) errors, which indicates the need for a classification transfer procedure.

As can be seen in Fig. 2, when the models obtained with the *C1* and *C2* criteria were applied to the secondary instrument data, there was a substantial decrease of the test (S) errors. Indeed, if six or more transfer samples are employed, both criteria lead to test (S) errors compatible with the test (P) errors. These findings are in agreement with the expected outcome, in that the use of transfer samples enables the selection of variables that are more robust against differences between the primary and secondary instruments. As a result, the resulting classification performance in the secondary instrument is improved.

Fig. 3 shows the subsets of selected variables by the SPA algorithm in the classification models. As can be seen, variables were selected near the absorption bands of carotenoids and chlorophyll in all models. The exception was obtained for the *C2* criterion with 4 transfer samples, where only one variable was selected in a region where the spectra of the two class are overlapping, which can be seen in Fig. 1. As a consequence, the resulting number of test (S) errors was even worse compared with the SPA-LDA model in usual form, as can be seen in Fig. 2. In this particular case, the poor outcome may be ascribed to the use of an insufficient number of samples to guide the variable selection process. Indeed, the *C2* criterion uses only the transfer samples in the evaluation of the candidate subsets of variables, whereas the *C1* criterion uses both the transfer and the training samples.

Fig. 4 compares the results, in terms of the number of errors, obtained by using SPA-LDA with the *C1* and *C2* criteria and SPA-LDA in usual form after performing the standardization procedures (DS or PDS). In all cases, the models were applied to the test set measured by the secondary instrument. As can be seen, except for criterion *C2* with $N_{transf} = 4$, the results yielded by SPA-LDA with the proposed criteria were similar to or better than those obtained by using the standardization methods.

4.2. Ethanol fuel data set

Fig. 5 presents the derivative NIR spectra of HEAF samples recorded in the range of 12.004–4.000 cm^{-1} for each instrument. As can be seen, the difference between the spectra of the two instruments is small, being more easily observed in the region around 5.854 cm^{-1}. NIR spectra show bands assigned to first overtones (5.000–6.000 cm^{-1}) and combination regions (4.600–4.000 cm^{-1})

Please cite this article in press as: K.D.T.M. Milanez, et al., Selection of robust variables for transfer of classification models employing the successive projections algorithm, Analytica Chimica Acta (2017), http://dx.doi.org/10.1016/j.aca.2017.07.037

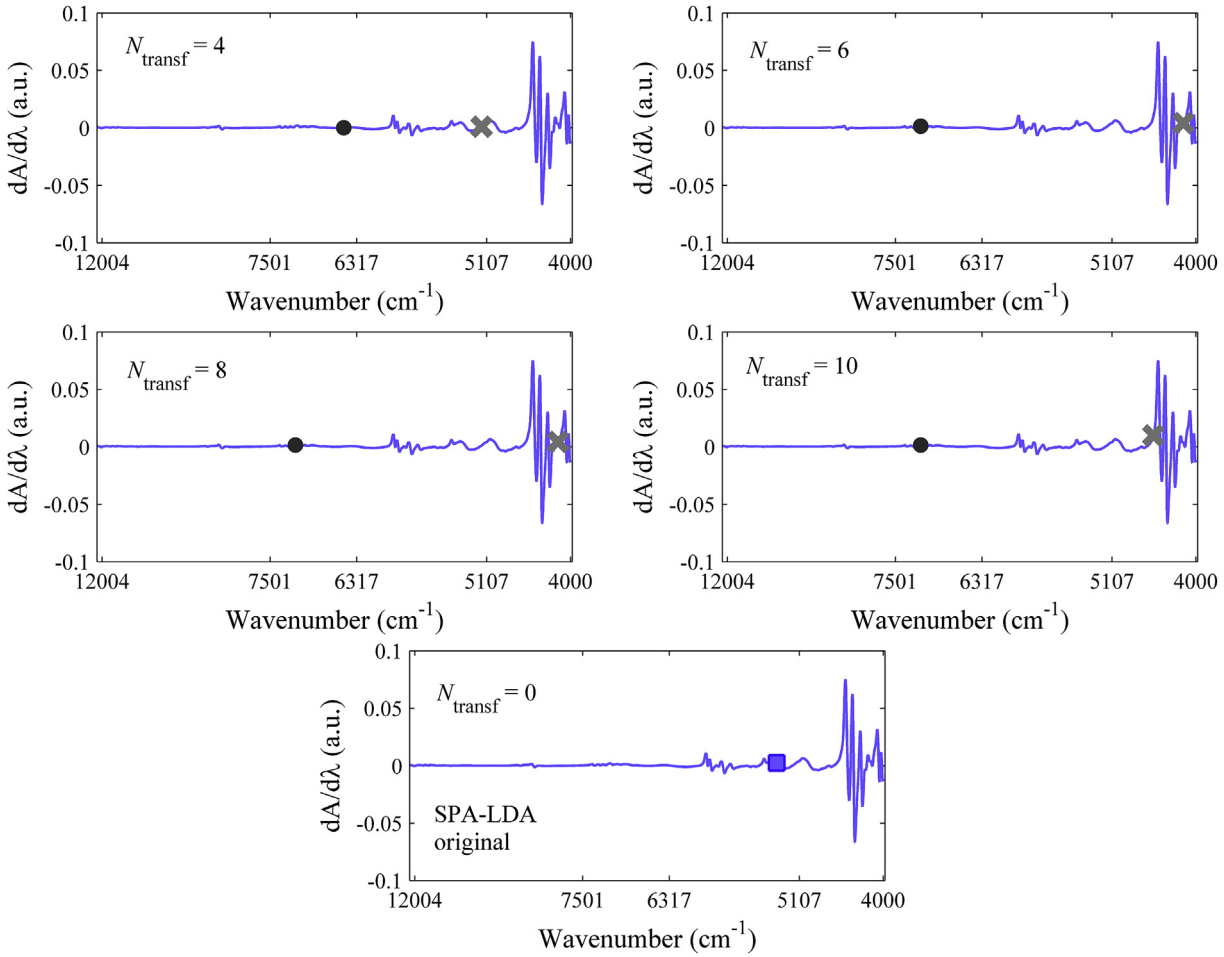


Fig. 7. Average spectrum of the data set with indication of the variables selected by the SPA-LDA algorithm in classification models involving different number of transfer samples. Variables selected by *C1:* ● and *C2:* ✖ criteria.

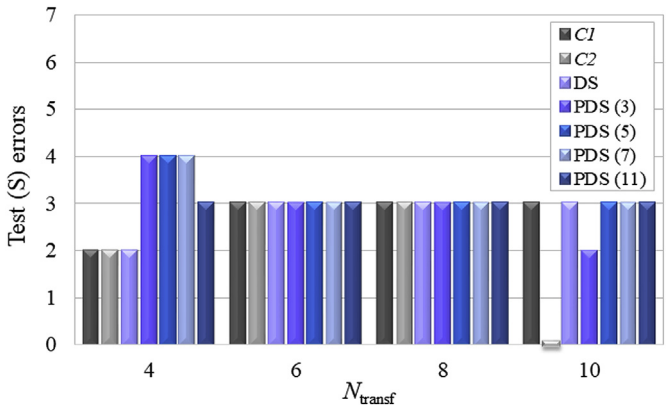


Fig. 8. Results in terms of number of errors obtained by SPA-LDA with the *C1* and *C2* criteria and SPA-LDA in usual form after performing the standardization procedures, applied to the test samples measured by the secondary instrument. The width of the window employed in the PDS standardization method is indicated in parenthesis.

of C-H stretching [50]. Additionally, at 5.128 cm^{-1} occurs combination bands of stretching and angular bending of the O-H in the alcohol samples [51].

Fig. 6 presents the classification results in terms of the number of test errors obtained by the three classification models (SPA-LDA in usual form, *C1* and *C2*), for different values of N_{transf} (number of transfer samples). As can be seen, no test (P) errors were obtained by SPA-LDA model in usual form. However, once again the number of test (S) errors is much larger than the number of test (P) errors.

The classification transfer carried out in the HEAF data set showed similar results to those found for the EVOO data set. By using the proposed criteria, a substantial decrease in the test (S) errors was observed. In particular, when the *C2* criterion was applied with $N_{transf} = 10$, the number of test (P) errors was reduced to zero, i.e. all the test samples were correctly classified.

As can be seen in Fig. 7, the variable selection process always resulted in a single variable. However, the use of transfer samples, either with criterion *C1* or criterion *C2*, affected the position of the selected variable.

Fig. 8 compares the results, in terms of the number of test (S) errors, obtained by using SPA-LDA with the *C1* and *C2* criteria and SPA-LDA in usual form after performing the standardization procedures (DS or PDS). In this case, the SPA-LDA model adapted with the *C1* and *C2* criteria provided results that were generally similar to those obtained by standardization methods. The highlight, however, is the correct classification of all test samples when the *C2* criterion was employed with $N_{transf} = 10$.

Table 2 summarizes the results obtained for the two data sets in terms of correct classification rate (CCR) obtained for test (S) set. Comparing the CCR values obtained for the test (S) errors set before and after the classification transfer procedures (with $N_{transf} = 6$) it is possible to observe a substantial increase in these values.

Please cite this article in press as: K.D.T.M. Milanez, et al., Selection of robust variables for transfer of classification models employing the successive projections algorithm, Analytica Chimica Acta (2017), http://dx.doi.org/10.1016/j.aca.2017.07.037

Table 2
Results obtained by the models in terms of the CCR (%) for the test (S) set before and after classification transfer procedures ($N_{transf} = 6$).

Set	Correct classification rate (%)							
	EVOO data set				HEAF data set			
	Original	*C1*	*C2*	DS	Original	*C1*	*C2*	DS
Test (S)	56	85	85	81	50	92.5	92.5	92.5

5. Conclusion

This paper proposed two new criteria for selection of robust variables using the successive projections algorithm coupled to linear discriminant analysis for transfer of classification models. The proposed criteria are based on the inclusion of transfer samples in the cost calculation for each subset of variables under consideration, in order to minimize the number of errors obtained in the classification of samples measured on the secondary instrument.

Two case studies involving data sets obtained from different analytical techniques were presented. In both cases, the proposed criteria substantially reduced the number of errors obtained in the classification of the samples measured on the secondary instrument. It was also observed that, the variables selected and the number of errors varied according to number of transfer samples. When compared to direct standardization (DS) and piecewise direct standardization (PDS) methods, the criteria showed equal or better results.

These results suggest that the proposed approach, using either of the developed criteria, is a promising alternative to full model recalibration or standardization of spectral responses. An additional advantage is the need for a low number of transfer samples to reduce the rate of misclassification.

Acknowledgments

The authors acknowledge the support of CNPq (303649/2015-1, 162930/2013-5 and 303714/2014-0), UNS – Argentina (Universidad Nacional del Sur, PGI - UNS 24Q054), CONICET – Argentina (Consejo Nacional de Investigaciones Científicas y Técnicas, PIP code 11220120100625). The authors are also grateful to Anna Luiza B. Brito for double-checking the chemometrics calculations, as well as to EMBRAPA (Campina Grande/PB) for providing the NIR spectra of HEAF samples (secondary instrument) employed in this study.

References

[1] L. Valderrama, P. Valderrama, Nondestructive identification of blue pen inks for documentoscopy purpose using iPhone and digital image analysis including an approach for interval confidence estimation in PLS-DA models validation, Chem. Intell. Lab. Syst. 156 (2016) 188–195.
[2] I. Nejadgholi, M. Bolic, A comparative study of PCA, SIMCA and Cole model for classification of bioimpedance spectroscopy measurements, Comput. Biol. Med. 63 (2015) 42–51.
[3] Ł. Górski, W. Sordoń, F. Ciepiela, W.W. Kubiak, M. Jakubowska, Voltammetric classification of ciders with PLS-DA, Talanta 146 (2016) 231–236.
[4] E.I. Geana, R. Popescu, D. Costinel, O.R. Dinca, R.E. Ionete, I. Stefanescu, V. Artem, C. Bala, Classification of red wines using suitable markers coupled with multivariate statistic analysis, Food Chem. 192 (2016) 1015–1024.
[5] H.V. Pereira, V.S. Amador, M.M. Sena, R. Augusti, E. Piccin, Paper spray mass spectrometry and PLS-DA improved by variable selection for the forensic discrimination of beers, Anal. Chim. Acta 940 (2016) 104–112.
[6] L. Pinto, C.H. Díaz Nieto, M.A. Zón, H. Fernández, M.C.U. Araújo, Handling time misalignment and rank deficiency in liquid chromatography by multivariate curve resolution: quantitation of five biogenic amines in fish, Anal. Chim. Acta 902 (2016) 59–69.
[7] I. Ramírez-Morales, D. Rivero, E. Fernández-Blanco, A. Pazos, Optimization of NIR calibration models for multiple processes in the sugar industry, Chem. Intell. Lab. Syst. 159 (2016) 45–57.
[8] B. Debus, D.O. Kirsanov, V.V. Panchuk, V.G. Semenov, A. Legin, Three-point multivariate calibration models by correlation constrained MCR-ALS: a feasibility study for quantitative analysis of complex mixtures, Talanta 163 (2017) 39–47.
[9] F. Zhang, T. Liu, X.Z. Wang, J. Liu, X. Jiang, Comparative study on ATR-FTIR calibration models for monitoring solution concentration in cooling crystallization, J. Cryst. Growth 459 (2017) 50–55.
[10] A.J. Tencate, J.H. Kalivas, A.J. White, Fusion strategies for selecting multiple tuning parameters for multivariate calibration and other penalty based processes: a model updating application for pharmaceutical analysis, Anal. Chim. Acta 921 (2016) 28–37.
[11] F.A. Honorato, R.K.H. Galvão, M.F. Pimentel, B. Barros Neto, M.C.U. Araújo, F.R. de Carvalho, Robust modeling for multivariate calibration transfer by the successive projections algorithm, Chem. Intell. Lab. Syst. 76 (2005) 65–72.
[12] R.K.H. Galvão, S.F.C. Soares, M.N. Martins, M.F. Pimentel, M.C.U. Araújo, Calibration transfer employing univariate correction and robust regression, Anal. Chim. Acta 864 (2015) 1–8.
[13] R.N. Feudale, N.A. Woody, H. Tan, A.J. Myles, S.D. Brown, J. Ferré, Transfer of multivariate calibration models: a review, Chem. Intell. Lab. Syst. 64 (2002) 181–192.
[14] F.A. Honorato, B. Barros Neto, M.N. Martins, R.K.H. Galvão, M.F. Pimentel, Transferência de calibração em métodos multivariados, Quim. Nova 30 (2007) 1301–1312.
[15] Y.D. Wang, D.J. Veltkamp, B.R. Kowalski, Multivariate instrument standardization, Anal. Chem. 63 (1991) 2750–2756.
[16] C.S. Gondim, R.G. Junqueira, S.V.C. de SOUZA, Trends in implementing the validation of qualitative methods of analysis, Rev. Inst. Adolfo Lutz 70 (2011) 433–447.
[17] O.E. Noord, Multivariate calibration standardization, Chem. Intell. Lab. Syst. 25 (1994) 85–97.
[18] E. Bouveresse, D.L. Massart, Standardisation of near-infrared spectrometric instruments: a review, Vib. Spectrosc. 11 (1996) 3–15.
[19] L. Salguero-Chaparro, B. Palagos, F. Peña-Rodríguez, J.M. Roger, Calibration transfer of intact olive NIR spectra between a pre-dispersive instrument and a portable spectrometer, Comput. Electron. Agric. 96 (2013) 202–208.
[20] N.C. da Silva, C.J. Cavalcanti, F.A. Honorato, J.M. Amigo, M.F. Pimentel, Standardization from a benchtop to a handheld NIR spectrometer using mathematically mixed NIR spectra to determine fuel quality parameters, Anal. Chim. Acta 954 (2017) 32–42.
[21] F. Seichter, J. Vogt, P. Radermacher, B. Mizaikoff, Nonlinear calibration transfer based on hierarchical Bayesian models and Lagrange Multipliers: error bounds of estimates via Monte Carlo e Markov Chain sampling, Anal. Chim. Acta 951 (2017) 32–45.
[22] B. Malli, A. Birlutiu, T. Natschläger, Standard-free calibration transfer - an evaluation of different techniques, Chem. Intell. Lab. Syst. 161 (2017) 49–60.
[23] V.H. da Silva, J.J. da Silva, C.F. Pereira, Portable near-infrared instruments: application for quality control of polymorphs in pharmaceutical raw materials and calibration transfer, J. Pharm. Biomed. 134 (2017) 287–294.
[24] J. Fonollosa, L. Fernández, A. Gutiérrez-Gálvez, R. Huerta, S. Marco, Calibration transfer and drift counteraction in chemical sensor arrays using direct standardization, Sens. Actuator B 236 (2016) 1044–1053.
[25] C. Liang, H. Yuan, Z. Zhao, C. Song, J. Wang, A new multivariate calibration model transfer method of near-infrared spectral analysis, Chem. Intell. Lab. Syst. 153 (2016) 51–57.
[26] A.J. Myles, T.A. Zimmerman, S.D. Brown, Transfer of multivariate classification models between laboratory and process near-infrared spectrometers for the discrimination of green Arabica and Robusta coffee beans, Appl. Spectrosc. 60 (2006) 1198–1203.
[27] C.V. Di Anibal, I. Ruisánchez, M. Fernández, R. Forteza, V. Cerdà, M.P. Callao, Standardization of UV–visible data in a food adulteration classification problem, Food Chem. 134 (2012) 2326–2331.
[28] N.C. Silva, M.F. Pimentel, R.S. Honorato, M. Talhavini, A.O. Maldaner, F.A. Honorato, Classification of Brazilian and foreign gasolines adulterated with alcohol using infrared spectroscopy, Forensic Sci. Int. 253 (2015) 33–42.
[29] K.D.T.M. Milanez, A.C. Silva, J.E.M. Paz, E.P. Medeiros, M.J.C. Pontes, Standardization of NIR data to identify adulteration in ethanol fuel, Microchem. J. 124 (2016) 121–126.
[30] R.A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugen. 7 (1936) 179–188.
[31] M.J.C. Pontes, R.K.H. Galvão, M.C.U. Araújo, P.N.T. Moreira, O.D. Pessoa Neto, G.E. José, T.C.B. Saldanha, The successive projections algorithm for spectral variable selection in classification problems, Chem. Intell. Lab. Syst. 78 (2005) 11–18.
[32] M.R. Almeida, D.N. Correa, W.F.C. Rocha, F.J.O. Scafi, R.J. Poppi, Discrimination between authentic and counterfeit banknotes using Raman spectroscopy and PLS-DA with uncertainty estimation, Microchem. J. 109 (2013) 170–177.
[33] M. Insausti, A.A. Gomes, F.V. Cruz, M.F. Pistonesi, M.C.U. Araújo, R.K.H. Galvão, C.F. Pereira, B.S.F. Band, Screening analysis of biodiesel feedstock using UV–vis, NIR and synchronous fluorescence spectrometries and the successive projections algorithm, Talanta 97 (2012) 579–583.
[34] S.K.B. Freitas, E.C.L. do Nascimento, A.G.G. Dionízio, A.A. Gomes, M.C.U. Araújo, R.K.H. Galvão, A flow-batch analyzer using a low cost aquarium pump for classification of citrus juice with respect to brand, Talanta 107 (2013) 45–48.
[35] A.S. Marques, E.P. Moraes, M.A.A. Júnior, A.D. Moura, V.F.A. Neto, R.M. Neto, K.M.G. Lima, Rapid discrimination of *klebsiella pneumoniae* carbapenemase 2-producing and non-producing *klebsiella pneumoniae* strains using near-

infrared spectroscopy (NIRS) and multivariate analysis, Talanta 134 (2015) 126–131.
[36] K.D.T.M. Milanez, M.J.C. Pontes, Classification of edible vegetable oil using digital image and pattern recognition techniques, Microchem. J. 113 (2014) 10–16.
[37] P.H.G.D. Diniz, M.F. Barbosa, K.D.T.M. Milanez, M.F. Pistonesi, M.C.U. Araújo, Using UV–Vis spectroscopy for simultaneous geographical and varietal classification of tea infusions simulating a home-made tea cup, Food Chem. 192 (2016) 374–379.
[38] S.F.C. Soares, R.K.H. Galvão, M.J.C. Pontes, M.C.U. Araújo, A new validation criterion for guiding the selection of variables by the successive projections algorithm in classification problems, J. Braz. Chem. Soc. 25 (2014) 176–181.
[39] M.C.U. Araújo, T.C.B. Saldanha, R.K.H. Galvão, T. Yoneyama, H.C. Chame, V. Visani, The successive projections algorithm for variable selection in spectroscopic multicomponent analysis, Chem. Intell. Lab. Syst. 57 (2001) 65–73.
[40] S.F.C. Soares, A.A. Gomes, A.R.G. Filho, M.C.U. Araújo, R.K.H. Galvão, The successive projections algorithm, Trends Anal. Chem. 42 (2013) 84–98.
[41] R.K.H. Galvão, M.C.U. Araújo, Variable selection, in: B. Walczak, R. Tauler, S. Brown (Eds.), Comprehensive Chemometrics, Elsevier Inc., Oxford, 2009, pp. 233–283.
[42] W. Wu, Y. Mallet, B. Walczak, W. Penninckx, D.L. Massart, S. Heuerding, F. Erni, Comparison of regularized discriminant analysis, linear discriminant analysis and quadratic discriminant analysis applied to NIR data, Chemom. Intell. Lab. Syst. 329 (1996) 257–265.
[43] PROTESTE - Associação Brasileira de Defesa do Consumidor, 2016. http://www.proteste.org.br/azeite. Accessed September 2016.
[44] O. Zenebon, N.S. Pascuet, P. Tiglea, Métodos físico-químicos para análise de alimentos, fourth ed., Instituto Adolfo Lutz, São Paulo, 2008 (first digital ed.).
[45] IOC - International Olive Council, Trade Standard Applying to Olive Oils and Olive-pomace Oils, 2016. COI/T.15/NC No 3/Rev. 11 July 2016, http://www.internationaloliveoil.org/estaticos/view/222-standards. Accessed 14 February 2017.
[46] A. Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem. 36 (1964) 1627–1639.
[47] R.W. Kennard, L.A. Stone, Computer aided design of experiments, Technometrics 11 (1969) 137–148.
[48] R. Tarakowski, A. Malanowski, R. Kościesza, R.M. Siegoczyński, VIS spectroscopy and pressure induced phase transitions – chasing the olive oils quality, J. Food Eng. 122 (2014) 28–32.
[49] A.J. Young, H.A. Frank, Energy transfer reactions involving carotenoids: quenching of chlorophyll fluorescence, J. Photoch. Photobio. B 36 (1996) 3–15.
[50] L.F.B. de Lira, M.S. de Albuquerque, J.G.A. Pacheco, T.M. Fonseca, E.H. de Siqueira Cavalcanti, L. Stragevitch, M.F. Pimentel, Infrared spectroscopy and multivariate calibration to monitor stability quality parameters of biodiesel, Microchem. J. 96 (2010) 126–131.
[51] A.C. Silva, L.F.B.L. Pontes, M.F. Pimentel, M.J.C. Pontes, Detection of adulteration in hydrated ethyl alcohol fuel using infrared spectroscopy and supervised pattern recognition methods, Talanta 93 (2012) 129–134.

Please cite this article in press as: K.D.T.M. Milanez, et al., Selection of robust variables for transfer of classification models employing the successive projections algorithm, Analytica Chimica Acta (2017), http://dx.doi.org/10.1016/j.aca.2017.07.037