Received: 30 July 2017 | Revised: 31 January 2018 | Accepted: 7 February 2018

DOI: 10.1002/cem.3030

TUTORIAL

WILEY Journal of CHEMOMETRICS

Multiclass partial least squares discriminant analysis: Taking the right way—A critical tutorial

Alexey L. Pomerantsev | Oxana Ye. Rodionova

Semenov Institute of Chemical Physics RAS, Kosygin str. 4, 119991 Moscow, Russia

Correspondence
Alexey L. Pomerantsev, Semenov Institute of Chemical Physics RAS, Kosygin str. 4, 119991 Moscow, Russia.
Email: forecast@chph.ras.ru

Funding information
Russian State assignment, Grant/Award Number: AAAA-A18-118020690203-8; International Atomic Energy Agency, Grant/Award Number: D5240 and G42007

Abstract

Here, the theory of the multi-class partial least squares discriminant analysis (PLS-DA) is presented. A distinct feature of this theory is that it does not utilize PLS scores but is entirely based on the predicted dummy responses. It is shown that the results of the multi-class PLS-DA can be presented in a straightforward way by projecting the response matrix on the "super-score" space by means of principal component analysis. Two approaches to discrimination are considered: the hard and the soft way of allocation. Correspondingly, 2 versions of PLS-DA are presented: the conventional hard PLS-DA, and the newly introduced soft PLS-DA that seems to be a novel approach in chemometrics. The quality of classification is assessed using the figures of merit (sensitivity, specificity, and efficiency). It is shown how these characteristics are used for the selection of the model complexity. A number of practical problems are investigated, such as unbalanced sizes of classes, comparison of the discriminant and the class-modeling methods and authentication by the "one against all" strategy. The paper is illustrated by real-world and simulated examples.

KEYWORDS

authenticity, multiclass discrimination, partial least squares discriminant analysis, PLS-DA, soft and hard classification

1 | INTRODUCTION

Partial least squares discriminant analysis (PLS-DA) is an enormously popular method in various scientific areas: genomics,[1] proteomics,[2] metabolomics,[3] as well as in food[4] and pharmaceutical[5] sciences. A search in Scopus returns approximately 3500 papers for this keyword, which makes it even stranger puzzling to see very few theoretical papers devoted to this method. We can only mention a handful. The first is a pioneer research[6] by Stahle and Wold. A valuable contribution to the PLS-DA theory was made by Barker and Rayens in paper.[7] At this juncture, we can mark an excellent research piece[8] by Indahl, Martens, and Næs, which is extensively used in our study, and paper[9] by Nocairi et al. The authors of these works have showed that the dummy-regression based PLS-DA can serve as a feature extractor from high-dimensional X space into low-dimensional Y space.

Reflecting on the PLS-DA basics, we can mention 3 major issues that are worthy of discussion and improvement. The first one does not actually present a problem to solve, but a point, which should be mentioned. The application of the PLS scores for classification can lead to incorrect results and wrong interpretations. As early as in 2008, Westerhuis and co-authors[10] noted that "the PLSDA score plot therefore does not give a good representation of class difference

between the groups". Later, Kjeldahl and Bro repeated this warning in paper.[11] In spite of these clear alarms, plenty of researchers still employ the PLS-DA scores as the main source for conclusions. So, we are repeating this again—be very careful in making conclusions regarding classification on the base of the PLS scores alone, without proper validation with a relevant test set!

The second issue is much more interesting and inspiring for new investigation. It is the fact that most papers are concern with a binary PLS-DA when only 2 classes are considered. It looks like a multi-class PLS-DA never existed, or that it is too sophisticated for practical implementation. In attempts to avoid the actual multi-class discrimination, researchers invent very complex schemes that split a multi-class task into a set of binary classification problems.[12] Nevertheless, the multi-class PLS-DA exists, and we present its theory, which, in fact, is not more complex than the binary version.

The last issue that is presented in this paper is of methodological nature. In Rodionova et al,[13] we discuss that PLS-DA is an inappropriate method of authentication. In fact, PLS-DA is a good method when used according to its intended purpose, which is discrimination. At the same time, PLS-DA has a serious shortcoming being a hard classification tool. In general, 2 approaches to classification can be considered: the hard and soft way of allocation. The first method presumes that each sample is mandatorily attributed to 1 and only 1 class. The second one allows a sample to be allocated into more than 1 class, or even left unclassified. Based on this concept, we suggest 2 methods of PLS-DA: the conventional hard PLS-DA, and the newly introduced soft PLS-DA that seems to be a novel approach in chemometrics. We show that in some cases the soft PLS-DA can be an appropriate tool for authentication.

PLS-DA provides us with a wide selection of related topics that could be discussed. Many of them have been left out of scope—some deliberately, others due to the limitation of the paper size. In particular, due to the previously mentioned reasons, we do not discuss methods based on the PLS scores analysis. We also decided not to touch on the OPLS-DA technique,[14] because this interesting approach deserves intended separate consideration. One more appealing issue that is not considered here is application of PLS-DA for the analysis of the importance of the variables.

The concluding remark is just a technical note—in this paper, we use row-wise vector notation.

2 | THE DATASETS

The paper is illustrated with the following real-world datasets.

Dataset *Pills* consists of the NIR spectra (4000–12 500 cm^{-1} range with resolution of 8 cm^{-1}) of uncoated tablets of calcium channel blockers, produced by 7 different manufacturers and denoted as *A1*, *A2*, .., *A7*. All producers make the tablet with the same quantity (10 mg) of the active pharmaceutical ingredient originated from the same source. Each manufacturer is represented by a set of batches ranging from 3 to 10. The sizes of classes *A1* to *A7* are correspondingly: 30, 50, 70, 50, 30, 50, and 100. Overall, there are 380 tablets in the dataset. The detailed description and the class modeling results are presented in Rodionova et al.[15] Initially, classification models were built and validated using test samples of the same class. In this study, training sets of various classes are used for building classification and discrimination models. Data from extraneous classes are used for demonstration models' specificity in various cases.

Dataset *Olives* is composed of the NIR spectra (4000–10 000 cm^{-1} range, at 4 cm^{-1} resolution) of olives in brine. Three classes comprise cultivars: *Taggiasca*, (*T*, 111 samples), *Leccino* (*L*, 72), and *Coquillo* (*C*, 50). Data summary and classification models are presented in Oliveri et al[16] and Rodionova et al.[17]

Dataset *Juices* consists of 38 samples of juices of different botanic origin. They are divided into 3 classes: citrus (*C*, 20 samples), apple (*A*, 7), and super juices (*S*, 11). Fifteen variables represent various bio-chemical (eg, antioxidant assays) and physicochemical (eg, pH, acidity) properties of the samples. Details can be found in Fidelis et al.[18]

In addition, 2 simulated datasets are considered in the Section "9", where all details are given.

3 | BASICS OF PLS-DA

We introduce a general multi-class PLS-DA concept, in which I samples are allocated to K groups, which are called the target classes. It is known that PLS-DA is a conventional regression approach, where the $(I \times J)$ feature matrix $\mathbf{X}$ is utilized as a predictor matrix, and the $(I \times K)$ dummy matrix $\mathbf{Y}$ is used as a response matrix. The predictor matrix $\mathbf{X}$ is composed of I samples (rows) and J variables (columns). The samples are split into K groups of sizes $I_1 + I_2 + ... + I_K = I$, with the index groups $\omega(1), \omega(2),..., \omega(K)$, which indicate belonging of sample i to class k, so that $i \in \omega(k)$. Without loss of generality, we can assume that $\omega(1) = \{1, .., I_1\}$, $\omega(2) = \{I_1 + 1, .., I_1 + I_2\}$, etc.

The dummy matrix **Y** composes categorical (dummy) variables {0,1} that describe class memberships. It is constructed as follows. Consider a unit matrix of size K that can be presented as a column of row-vectors **e**

$$\mathbf{E} = \begin{vmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \cdots \\ \mathbf{e}_K \end{vmatrix} = \begin{vmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & 1 \end{vmatrix} \tag{1}$$

Each vector $\mathbf{e}_k$, $k = 1,..,K$, is a pattern response for class k. Matrix **Y** can also be represented as a column that comprises the row vectors $\mathbf{y}_i$, $i = 1,.., I$

$$\mathbf{Y} = \begin{vmatrix} \mathbf{y}_1 \\ \cdots \\ \mathbf{y}_i \\ \cdots \\ \mathbf{y}_I \end{vmatrix}; \quad \text{where } \mathbf{y}_i = \mathbf{e}_k, \text{when } i \in \omega(k) \tag{2}$$

The predictor **X** and the response **Y** matrices are related by the PLS2 regression, which is employed to calculate the predicted responses $\hat{\mathbf{Y}}$. We are not discussing the PLS method, as it is explained in numerous textbooks, eg, in Martens and Naes.[19] The number of PLS2 components, which are referred to as the latent variables (LV), is selected using a validation approach; it is explained below.

The data pre-treatments for PLS2 regression are rather common: **X** matrix is always centered (column-wise) and may be scaled in dependence on the nature of the variables; **Y** matrix is only centered.

A discrimination rule is based on the comparison of each row $\hat{\mathbf{Y}}_i$ of matrix $\hat{\mathbf{Y}}$ with each pattern response vector $\mathbf{e}_k$. Sample i is attributed to that class k, which pattern is closer. To evaluate the distance between a sample and a class pattern, it is natural[7,8] to treat matrix $\hat{\mathbf{Y}}$ as the input data set for classification. However, this cannot be done directly, because this matrix has a rank of $K - 1$, and the corresponding covariance matrix is singular.

4 | PLS-DA GEOMETRY

The space spanned by $\hat{\mathbf{Y}}$ matrix is evidently unique, but it can be parameterized in different ways. In Indahl et al,[8] it was suggested to "obtain the reduced dummy matrix by elimination of one column from **Y**". In our turn, we suggest employing the principal components analysis (PCA) to reduce matrix $\hat{\mathbf{Y}}$.[20] This approach provides us with several obvious benefits, such as orthogonality, independence, etc., but the main advantage is a simple geometrical structure of the PCA space.

The geometry of PCA applied to the PLS responses is illustrated in Figure 1 for the case of 3 classes. The elements of $\hat{\mathbf{Y}}$ matrix are shown by the colored marks (dots, squares, and triangles), which belong to the 3 classes. All these points are located on a plane (the gray shaded triangle), which passes through the class pattern points—$\mathbf{e}_1 = (1,0,0)$, $\mathbf{e}_2 = (0,1,0)$, and $\mathbf{e}_3 = (0,0,1)$—shown by the crosses. This plane is actually the PCA score space; its PC axes are presented by the magenta colored arrows. These properties are proven in Theorem 1 in Appendix.

Before application of PCA, matrix $\hat{\mathbf{Y}}$ is centered**.** PCA gives us the following decomposition:

$$\hat{\mathbf{Y}} = \mathbf{u}^{\mathrm{t}}\mathbf{m} + \mathbf{T}\mathbf{P}^{\mathrm{t}} \tag{3}$$

where $\mathbf{m} = (m_1,...,m_K)$, is the $(1\times K)$ vector of column-wise mean values of **Y**

$$m_k = \frac{1}{I}\sum_{i=1}^{I} y_{ik} = \frac{1}{I}\sum_{i=1}^{I} \hat{y}_{ik} = \frac{I_k}{I} \tag{4}$$

and **u** is a vector of units of the appropriate dimensionality. Here, **u** is the $(1\times K)$ vector.

The number of the PCA components is $K - 1$; that is why Equation 3 is an exact relationship. The $(I \times (K - 1))$ scores matrix **T** represents a new data set for which a classification method can be employed.

The PLS-DA geometry can be explained in another way. Matrix $\hat{\mathbf{Y}}$ has an inherited property– the sum of all elements in a row is 1. This means that all points $\hat{\mathbf{Y}}_i$ $(i = 1,..., I)$ belong to a simplex, which rest upon the pattern vectors $\mathbf{e}_k$. The

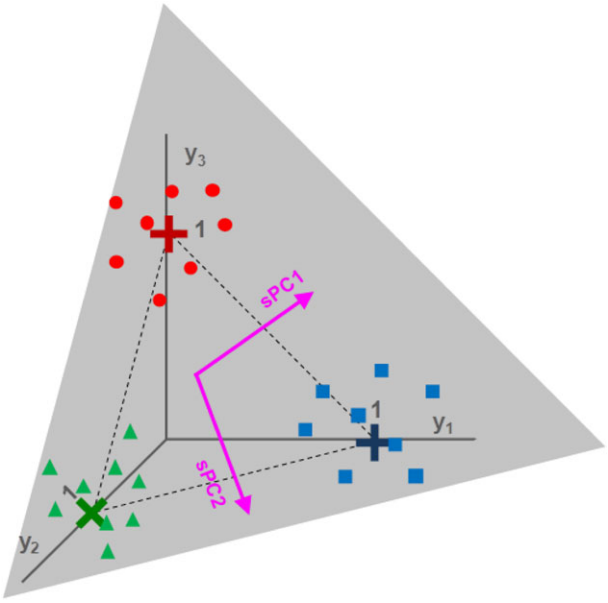

FIGURE 1 Geometry of PLS-DA

PCA projection does not change this $(K-1)$ dimensional simplex space, but only introduces new coordinates there. Different sizes of the training classes influence on the PCA loadings, and, therefore, on the sample positions, but that variability has no effect on the simplex space itself.

5 | HARD PLS-DA

In general, there are many ways to classify **T** using different concepts: deterministic,[7] or probabilistic[8]; various methods: LDA, quadratic discriminant analysis (QDA), k-nearest neighbors, tree methods, etc.[8] For the didactic reasons, we use the most straightforward LDA approach based on the total covariance matrix.

In this case, the application of LDA has specific features. Firstly, the pulled covariance matrix is known from PCA; it is a diagonal matrix.

$$\mathrm{cov}(\mathbf{T}) = \mathbf{T}^{\mathrm{t}}\mathbf{T} = \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_{K-1}) \tag{5}$$

Secondly, the class centers (centroids), $\mathbf{c}_k$, $k = 1,..,K$, are not the means of classes as they are in the conventional LDA. In our approach, the class centers are known in advance, and they constitute the $(K \times K-1)$ matrix **C**, which is the projection of the class pattern matrix **E** onto the PCA score space

$$\mathbf{C} = \left(\mathbf{E} - \mathbf{u}^{\mathrm{t}}\mathbf{m}\right)\mathbf{P} = \begin{vmatrix} \mathbf{c}_1 \\ \cdots \\ \mathbf{c}_K \end{vmatrix} \tag{6}$$

Here, **u** is the $(1 \times K)$ vector of units.

Therefore, the distance from sample i to target class k is given by a formula

$$d_{ik} = (\mathbf{t}_i - \mathbf{c}_k)\,\Lambda^{-1}(\mathbf{t}_i - \mathbf{c}_k)^{\mathrm{t}}. \tag{7}$$

The discrimination rule is very simple and says that sample i belongs to the class which is closer by metric given in Equation 7.

Setting $d_k = d_l$ in formula (7), we obtain an equation of a hyperplane that separates classes k and l. The quadratic terms are reduced, and the equation becomes linear

$$(\mathbf{w}_k - \mathbf{w}_l)\mathbf{t}^{\mathrm{t}} = v_k - v_l \tag{8}$$

where

$$\mathbf{w}_k = \mathbf{c}_k\mathbf{\Lambda}^{-1}, \qquad v_k = 0.5\,\mathbf{c}_k\mathbf{\Lambda}^{-1}\mathbf{c}_k^{\mathrm{t}}. \tag{9}$$

An interesting property of this version of PLS-DA is that all hyperplanes cross at a common point (see Theorem 2 in Appendix).

$$\mathbf{t}_0 = \mathbf{vP\Lambda}. \tag{10}$$

In Figure 2, we present the results of PLS-DA (12 LVs, $\alpha = 0.05$) applied to *Juices* data set, which has 3 classes: *A*, *C*, and *S*. The colored marks show the projections of the corresponding $\hat{\mathbf{Y}}$ values onto the PCA score space, and the solid black lines demonstrate the separation of this space into 3 acceptance areas as given in Equation 8. The large crosses mark the class centers $\mathbf{c}_k$. In this plot, several wrong attributions can be seen: samples A_1 and S_1 to class *C*, and sample C_1 to class *A*.

It is worth to remind that we do not use the PLS scores and loadings. Therefore, to emphasize that PCA scores $\mathbf{t}$ are not the PLS2 scores, we use special notation for the plot axes: sPC1 and sPC2 meaning that those are the "superscores" obtained according to the following scheme

$$\mathbf{X}, \mathbf{Y} \xrightarrow{\text{PLS2}} \hat{\mathbf{Y}} \xrightarrow{\text{PCA}} \mathbf{T} \tag{11}$$

6 | SOFT PLS-DA

In our paper,[13] we criticized the PLS-DA method for wrong interpretation of new objects. The main drawback of this approach is its inability of proper classification of the samples, which do not belong to any of the predefined classes. The reason is the absence of soft decisions such as "the sample does not belong to any class at all".

In an attempt to improve on this weakness, we suggest a soft version of PLS-DA, which is based on the QDA[20] applied to $\mathbf{T}$ data set defined in Equation 11. We consider the PCA scores for each class separately and assume that they form the normally distributed subsets with the known means $\mathbf{c}_k$. Using the within-class covariance matrices

$$\mathbf{S}_k = \frac{1}{I_k}\sum_{i\in\omega(k)}(\mathbf{t}_i - \mathbf{c}_k)^{\mathrm{t}}(\mathbf{t}_i - \mathbf{c}_k) \tag{12}$$

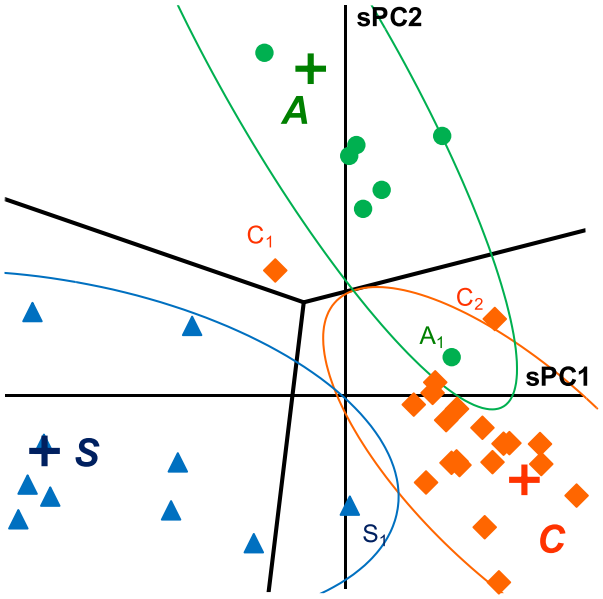

FIGURE 2 PLS-DA applied to the *Juices* data set. Hard and soft versions of PLS-DA

we can calculate new (Mahalanobis) distances between sample i and class k

$$d_{ik} = (\mathbf{t}_i - \mathbf{c}_k)\mathbf{S}_k^{-1}(\mathbf{t}_i - \mathbf{c}_k)^{\mathrm{t}}. \tag{13}$$

Assuming that these distances follow the chi-squared distribution, a new soft discrimination rule can be established. It says that sample i belongs to class k if the distance given in Equation 13 is less than a threshold

$$d_{\mathrm{crit}} = \chi^{-2}(1-\alpha, K-1). \tag{14}$$

where χ^{-2} is the quantile of the chi-squared distribution with $K-1$ degrees of freedom. Value α stands for a given type I error. According to this rule, a sample can simultaneously be attributed to several classes. Moreover, it may be not allocated at all, in case all the distances are greater than the threshold.

The acceptance area for each class can be represented by ellipsoids depicted around the corresponding class centers $\mathbf{c}_k$ as it is shown in Figure 2. In this case, we observe the following allocation of the selected samples. Sample A1 is correctly attributed to class *A*, but it is simultaneously marked as a member of *C*. Sample C1 is not classified at all, and C2 is wrongly attributed to *A*. Sample S1 is correctly allocated in *S*.

Discussing the advantage of the hard and soft PLS-DA, we suggest an example with the *Pills* dataset (3 LVs, $\alpha = 0.01$) shown in Figure 3. In this example, classes *A2*, *A5*, and *A7* are used for the PLS-DA modeling (hard and soft), and classes *A4* and *A6* are then utilized as new objects. In the frame of the hard PLS-DA, which is presented by the solid black lines, the entire class *A6* belongs to *A5*, and class *A4* is shared between *A7* and *A2*. Using the soft PLS-DA, which is shown by the ellipses outlined by the solid lines, we happily conclude that class *A6* consists of alien objects, which do not belong to any class. Class *A4* is partially located (34%) inside class *A7*, and this is a misclassification.

An additional benefit of the soft approach can be obtained in case we utilize the outlier thresholds that are shown by the dashed ellipses in Figure 3. These thresholds are established introducing the outlier significance level, γ, by a formula

$$d_{\mathrm{out}} = \chi^{-2}\left((1-\gamma)^{1/I_k}, K-1\right). \tag{15}$$

We can see that, for $\gamma = 0.05$, sample A5_8 is an outlier that should be removed.

To assess the performance of the hard and soft methods of PLS-DA, we should compare their figures of merit, which are explained in the next section. The actual values for this example are given in Table 1. Now we can only mention that, at the training stage, the total efficiencies of the soft and hard approaches are almost equal to 99% and 100%, respectively. However, at the prediction stage, the total efficiency of the soft version outperforms the hard one as 83% to 0%.

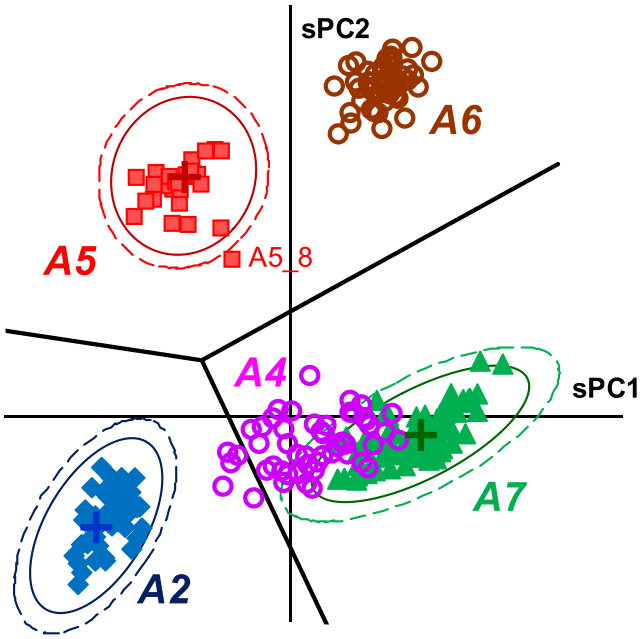

FIGURE 3 PLS-DA applied to the *Pills* data set

TABLE 1 *Pills* data set. Figures of merit for the hard and soft PLS-DA

Hard PLS-DA				**Soft PLS-DA**			
Training							
	A2	*A5*	*A7*		*A2*	*A5*	*A7*
A2	50	0	0	*A2*	50	0	0
A5	0	30	0	*A5*	0	29	0
A7	0	0	100	*A7*	0	0	96
CSNS	100%	100%	100%	CSNS	100%	97%	96%
CSPS	100%	100%	100%	CSPS	100%	100%	100%
CEFF	100%	100%	100%	CEFF	100%	98%	98%
TSNS	100%			TSNS	97%		
TSPS	100%			TSPS	100%		
TEFF	100%			TEFF	98%		
Prediction of extraneous objects							
	A2	*A5*	*A7*		*A2*	*A5*	*A7*
A4	5	0	45	*A4*	0	0	17
A6	0	50	0	*A6*	0	0	0
CSPS	95%	50%	55%	CSPS	100%	100%	83%
TSPS	0%			TSPS	83%		

7 | FIGURES OF MERIT

Many figures of merit (FoM) exist to assess the quality of classification. These characteristics are calculated for samples, which class membership is known a priori. In PLS-DA, we prefer using 3 FoMs: sensitivity, specificity, and efficiency.[17] These values are introduced using the confusion matrix **N**, which elements, n_{kl} represent the number of samples from class k, predicted as members of the target class l. Speaking of "the number of samples from class k" we mean the samples in a specific (training, or test, or validation) set for which FoMs are calculated. Introducing the FoMs values, we always indicate what data set is used for their calculation. The examples of the confusion matrix are given in Table 1 in the previous section.

In case of the hard PLS-DA, the following relation is hold

$$\sum_{l=1}^{K} n_{kl} = I_k \tag{16}$$

where I_k is the number of objects in class k. In case of the soft PLS-DA, Equation 16 is not satisfied because in this method an object could be simultaneously attributed to several classes (+1 to each of those classes), or not classified at all (+0 to all classes).

The FoM values can be considered in relation to a particular target class k, or they can be calculated for the entire model. The following definitions are given for the training set, while the features of the FoM calculation for the test and validation (new) sets are presented below.

Class sensitivity, CSNS(k), is defined for each target class k as the percentage of samples of this class, which are correctly recognized as the members of this class. It can also be defined as the rate of true positives, and, therefore, in soft PLS-DA it is complementary to the type I error α.

$$\mathrm{CSNS}(k) = n_{kk}/I_k. \tag{17}$$

Class specificity, CSPS(k), is defined for each target class k as the percentage of samples from other classes (not k), which are correctly attributed as inconsistent with the target class. This value is complementary to the rate of false positives.

$$\mathrm{CSPS}(k) = 1 - \sum_{l \neq k} n_{kl} / \sum_{l \neq k} I_l \tag{18}$$

To characterize an overall quality of classification with respect to class k, the *class efficiency*, CEFF(k), is usually introduced[21] as the geometric mean of sensitivity and specificity,

$$\mathrm{CEFF}(k) = \sqrt{\mathrm{CSNS}(k) \cdot \mathrm{CSPS}(k)} \tag{19}$$

The FoMs that characterize the performance of the entire PLS-DA model are defined as follows. The *total sensitivity*, TSNS, is given by a formula

$$\mathrm{TSNS} = \frac{1}{I} \sum_{k=1}^{K} n_{kk} \tag{20}$$

The *total specificity*, TSPS, is defined by a formula

$$\mathrm{TSPS} = 1 - \frac{1}{I} \sum_{k \neq l}^{K} n_{kl}. \tag{21}$$

The *total efficiency*, TEFF, is defined similar to Equation 19

$$\mathrm{TEFF} = \sqrt{\mathrm{TSNS} \cdot \mathrm{TSPS}}. \tag{22}$$

In case FoMs are calculated for the test set, or for the validation set, the formulae given in Equations 17 to 22 should be modified. Firstly, the class sizes, $I_1 + I_2 + \ldots + I_K = I$, should be replaced with the numbers, which represent the actual number of samples in the corresponding sets. Secondly, if a new set consists of extraneous (non-target objects), only specificity can be calculated; therefore, CSPS(k) = CEFF(k).

Table 1 contains the FoM values obtained in the example with *Pills* data set presented in the previous section. At the training stage, the first 3 rows show the confusion matrices, and the subsequent rows demonstrate the obtained FoM values. The hard PLS-DA has ideal results. The samples are properly attributed to their own classes, and all FoMs are equal to 100%. In the soft case, classification for *A2* is perfect, and CSNS(*A2*) = 50/50 = 100%. One sample of *A5* is not accepted (outlier), so, in line with Equation 18, CSNS(*A5*) = 29/30 ≈ 97%. In class *A7*, 4 samples out of 100 are located outside the ellipse, so CSNS(*A7*) = 96/100 = 96%. We have no wrongly accepted samples (false positives) in training; thus, all specificity values (CSPS) are equal to 100%. The class efficiencies are calculated in line with Equation 19 as follows: CEFF(*A2*) = $(100 \times 100)^{1/2}$ = 100%, CEFF(*A5*) = $(97 \times 100)^{1/2}$ ≈ 98%, CEFF(*A7*) = $(96 \times 100)^{1/2}$ ≈ 98%. In the soft version, total sensitivity TSNS = (50 + 29 + 95)/(50 + 30 + 100) ≈ 97%. Total efficiency TEFF = $(97 \times 100)^{1/2}$ ≈ 98%.

The second part of Table 1 contains the results obtained in prediction. Again, 2 first rows represent the confusion matrix, and the following rows show the FoM values. Because *A4* and *A5* (50 and 50 samples) are extraneous classes, the sensitivity values cannot be calculated. For the hard PLS-DA, we have the following results. Five samples from *A4*, and no samples from *A6* are wrongly attributed to class *A2*; therefore, the class specificity is CSPS(*A2*) = 1–(5 + 0)/(50 + 50) = 95%. No samples from *A4* and 50 samples from *A6* are false positives with respect to class *A5*, CSPS(*A5*) = (0 + 50)/100 = 50%. Similarly, CSPS(*A7*) = 1–(45 + 0)/100 = 55%. In the soft case, classes *A2* and *A5* are free from aliens, so CSPS(*A2*) = CSPS(*A5*) = 100%, while 17 samples from *A4* are wrongly accepted as the members of class *A7*, therefore CSPS(*A7*) = 1–(0 + 17)/100 = 83%. Total specificities are obtained in the following way. In the hard case, TSPS = 1–(5 + 50 + 45)/100 = 0%. In the soft version, TSPS = 1–17/100 = 83%.

8 | PLS-DA COMPLEXITY

The FoMs are of great importance for validation of the PLS-DA model. Typically, they are considered with respect to 2 data sets: the training and the test sets. The latter one could be a real, external set, or it can be simulated in the course of cross-validation procedure.[10] In any case, to assess the model quality, the FoM values should be calculated both for the training and test sets, and then considered in parallel.

In particular, FoMs can be used for the selection of the model complexity, which, in case of PLS-DA, is the number of the latent variables (PLS components). To illustrate this procedure, we employ the *Olives* data with 3 classes, which are arbitrarily split into the training and test sets: (*T*) 83 + 28 = 111, (*L*) 59 + 13 = 72, and (*C*) 45 + 5 = 50. Both the hard and

soft ($\alpha = 0.01$) PLS-DA models are developed using a different number of LVs ranging from 3 to 20. In each case, the total efficiency values given in Equation 22 are calculated for the training and test sets. The results are shown in Figure 4.

In case of hard PLS-DA, the appropriate number of LVs equals 9, because at this point, we observe the convergence of 2 curves. In case of soft PLS-DA, the optimal number of LVs is 13, as at this point both curves are close to the projected efficiency that is $100(1 - \alpha) = 99\%$.

The selected number of LVs seems to be too high in comparison with the "common" number in a regular PLS calibration, which takes value between 3 and 5, and very seldom 7. In our opinion, the high values of LVs in PLS-DA cannot be considered as overfitting, as this problem is more complex than an ordinary calibration. In PLS-DA, each class takes approximately 2 to 3 LVs for the internal modeling, plus 1 to 2 LVs are necessary to describe the external links between classes. In the 3-class discrimination, this means, at least, $3 \times 2 + 3 = 9$ LVs, which is the number we obtained in the hard version. Analyzing the rate of the Y-variance explained by PLS in this example, we can notice that it grows slowly with LV: 5 LVs—0.81, 10 LVs—0.92, 15 LVs—0.96, and 20 LVs—0.99. Thus, we can conclude that a high number of LVs in PLS-DA is reasonable.

9 | TWO CLASS PLS-DA

The binary classification is the most popular approach within PLS-DA. As we mentioned earlier, the majority of publications regard this approach as a sole version of PLS-DA. Certainly, the case of $K = 2$ can be easily implemented in the frame of both soft and hard methods presented previously. However, there is another interesting option that relates to a different way of coding of the dummy matrix. Instead of the 2-column **Y** matrix, given in Equation 2, a 1 column matrix (vector **y**) is conventionally employed in the 2-class PLS-DA. It is defined as follows

$$y_i = \begin{cases} +1, & i \in \omega(1) \\ -1, & i \in \omega(2) \end{cases}, \quad i = 1, ..., I. \tag{23}$$

Obviously, with the 1-column coding, PLS regression, not PLS2, should be applied. The predicted response vector, $\hat{\mathbf{Y}}$, can be directly used for classification with a hard threshold equal to 0. In case of the soft PLS-DA, the acceptance areas are built as the confidence intervals around the class centers $\bar{y}_k$ with standard deviations s_k, which are calculated by formulae

$$\bar{y}_k = \frac{1}{I_k} \sum_{i \in \omega(k)} \hat{y}_i, \; s_k^2 = \frac{1}{I_k} \sum_{i \in \omega(k)} (\hat{y}_i - \bar{y}_k)^2, \; k = 1, 2 \tag{24}$$

Barker and Rayens[7] claimed, "in fact, our experiences have been that the classification results for different coding are almost identical". In our turn, we can only agree with this conclusion—the coding methods given in Equation 2 and Equation 23 lead to very similar results.

To illustrate the explained binary PLS-DA method, we present an interesting case study, which additionally aims at the clarification of the assertion[22] that unequal sizes of classes always lead to an inappropriate discrimination. In Krakowska et al,[23] the authors artificially reduce one of the classes in order to use classes with an equal number of samples for the PLS-DA modeling. We consider 2 classes from *Olives* dataset, *C* and *T*, which include 50 and 100 samples,

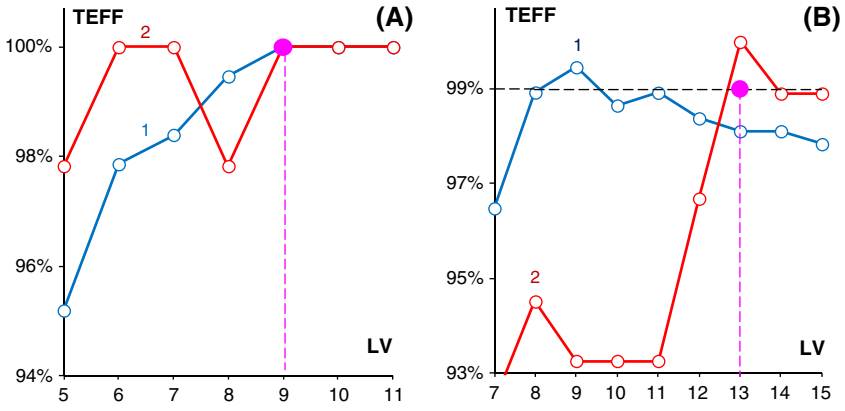

FIGURE 4 Selection of the number of LVs in *Olives* dataset. Curves show the total efficiencies: (1) training set, (2) test set. Plot (A): Hard PLS-DA, plot (B): Soft PLS-DA

correspondingly. Our plan is to reduce the size of *T* class systematically observing the effect of this diminution on the separation of classes.

The results (4 LVs), are shown in Figure 5, where the size of class *T* is varied as 100, 50, and 25 samples. The hard discrimination is marked by the solid line drawn at $\hat{Y} = 0$ (threshold), and the soft method ($\alpha = 0.01$) is presented by the rectangles outlined with the thin lines. This plot demonstrates that, in fact, the size of class *T* has a minor influence on the discrimination outcomes. The FoM values given in Table 2 confirm this conclusion.

From this table, we see that the class and total efficiencies vary insignificantly with the size of class *T*. In fact, this is an obvious conclusion because PLS-DA utilizes a regression approach, the efficiency of which mostly depends on the design of experiment, rather than on the size of the data. Reducing the *T* size in this example, we always selected a representative subset, and this approach gave us the presented result.

10 | PLS-DA AND SIMCA

This is a very popular opinion that PLS-DA better separates classes than SIMCA does. Various explanations of this consideration can be found in numerous publications. Contributing to this discussion, we have to make an important remark that, in general, this is not a fair comparison, because SIMCA and PLS-DA have absolutely different goals and ways of modeling.[13] SIMCA is a one-class classifier[21,24-26] that produces a description of a target class of objects and then detects whether a new object resembles this class or not. The rigorous version of SIMCA[17] does not utilize any information about the non-target (extraneous) classes even when the data regarding such extraneous classes are available. On the contrary, PLS-DA makes a description of several sets of objects that represent the predefined classes and then determines the membership of an object in one of these classes. Therefore, in our opinion, it is not consistent to compare the methods that have different objectives and employ various amounts of the modeling information. Nevertheless, we will

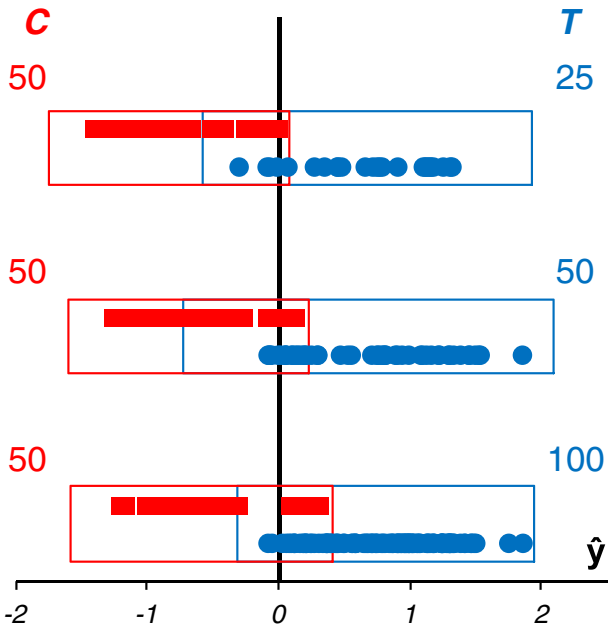

FIGURE 5 Two classes from olives dataset. The influence of the unbalanced class sizes on the PLS-DA

TABLE 2 The class and total efficiencies of PLS-DA for the different sizes of class *T*

Class sizes (*T*/*C*)	Hard PLS-DA			Soft PLS-DA		
	T	*C*	Total	*T*	*C*	Total
100/50	92%	92%	94%	92%	86%	88%
50/50	91%	91%	91%	75%	81%	78%
25/50	91%	91%	93%	91%	89%	90%

make this comparison using the PLS-DA rules of play: an exhaustive list of classes, and a compliant usage of the data sets. For this purpose, we present 2 didactic examples, in which data are organized in the simplest way: 2 variables, 2 classes, and normally distributed samples.

In the first example, we independently simulate 2 normal variables, x_1 and x_2, with zero mean and unit variance. Class 1 comprises 100 samples for which ${x_1}^2 + {x_2}^2 < 1$, and class 2 holds another 100 samples for which ${x_1}^2 + {x_2}^2 \geq 1$. In the result, we obtain a Russian pastry "vatrushka" (or "danish") shaped data, where class 1 is located in the middle, while class 2 occupies the periphery area. The results of the PLS-DA and SIMCA modeling (class 1 is used as the target class) at $\alpha = 0.05$ are shown in Figure 6.

Numerical outcomes are presented in Table 3. The SIMCA method is presented in 2 instances regarding the choice of the target class: Class 1 and Class 2.

We can conclude that, in this example, SIMCA is more efficient than PLS-DA. This is not a surprise because the "vartushka" data represent a nonlinear case, so the linear PLS-DA method fails to discriminate them. On the contrary, SIMCA is a quadratic approach, which perfectly models this specific data layout. In a real-world case, such type of data may occur in the following situation. One class is rather tight, eg, it comprises healthy biological tissues, and the other class is very broad, and, in fact, does not form a specific class at all, eg, it contains tissues damaged in various ways.

In our second example, the independent variables, x_1 and x_2, are also distributed normally. In both classes, we have 100 samples with the following parameters of $N(m, \sigma)$ distributions. In class 1: $m_1 = 0$, $\sigma_1 = 1$, $m_2 = +0.1$, $\sigma_2 = 0.01$. In class 2: $m_1 = 0$, $\sigma_1 = 1$, $m_2 = -0.1$, $\sigma_2 = 0.01$. In the result, we have data shaped as a Dutch "stroopwafel" (or French "macaron"), where 2 similar classes (wide but flat) are slightly shifted one against the other. The results of the PLS-DA and SIMCA modeling are shown in Figure 7.

Numerical outcomes are presented in Table 4, where 2 SIMCA models, regarding each target class, are built.

The results of the second example show that in this case SIMCA is worse than the hard PLS-DA. The former demonstrates a high sensitivity (96%) but low specificity (44%), while the latter has them balanced at 84% and 84% in the hard version. This can be explained by the different approaches to the data interpretation in the 2 methods. SIMCA looks for

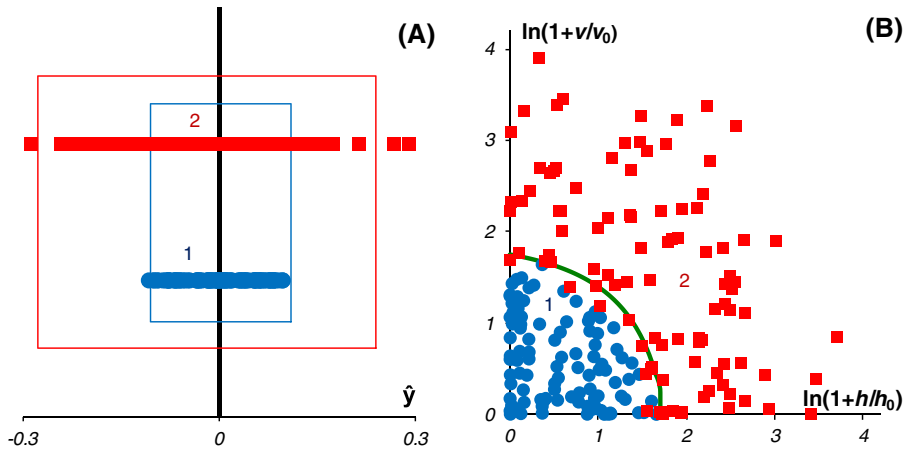

FIGURE 6 The "vatrushka" data classification. Plot (A): PLS-DA, plot (B): SIMCA, class 1 is used as the target class

TABLE 3 Figures of merit for the first didactic example (vatrushka)

Figures of merit	SIMCA		Hard PLS-DA		Soft PLS-DA	
	Class 1	Class 2	Class 1	Class 2	Class 1	Class 2
Class sensitivity	100%	93%	51%	54%	99%	96%
Class specificity	86%	52%	54%	51%	55%	0%
Class efficiency	93%	70%	52%	52%	74%	0%
Total sensitivity	97%		53%		98%	
Total specificity	73%		53%		28%	
Total efficiency	84%		53%		52%	

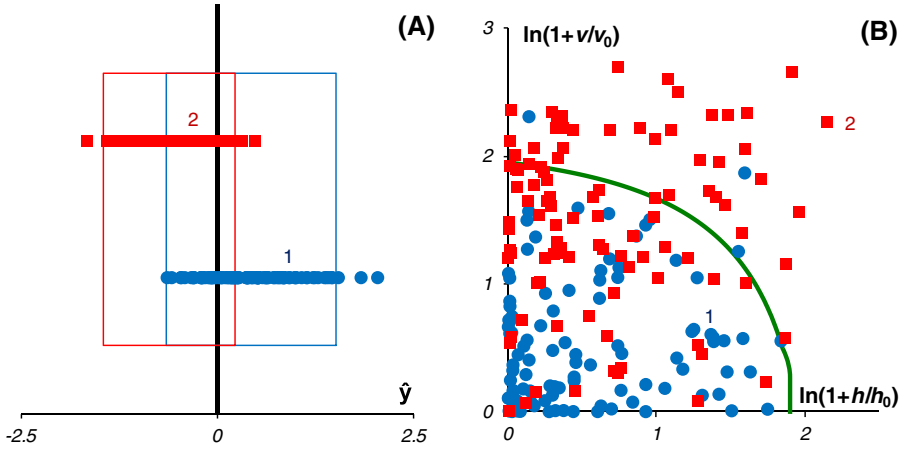

FIGURE 7 The "stroopwafel" data classification. Plot (A): PLS-DA, plot (B): SIMCA, class 1 is used as the target class

TABLE 4 Figures of merit for the second didactic example, "stroopwafel"

Figures of merit	SIMCA		Hard PLS-DA		Soft PLS-DA	
	Class 1	Class 2	Class 1	Class 2	Class 1	Class 2
Class sensitivity	96%	96%	76%	91%	97%	96%
Class specificity	39%	49%	91%	76%	48%	60%
Class efficiency	61%	69%	83%	83%	68%	76%
Total sensitivity	96%		84%		97%	
Total specificity	44%		84%		54%	
Total efficiency	65%		84%		72%	

features that better explain class unity, so it is mostly focused on the wide spread of variable x_1. On the contrary, PLS-DA concerns with the variables that better explain the diversity of the class, so it mostly cares about variable x_2. In a real-world case, such data may be obtained, eg, when we try to separate 2 mixtures, which have the same major components but different impurities.

A great advantage of PLS-DA is that it provides the loadings and scores, which give insight into the variables and samples, and this is really what makes the method special. A similar approach in SIMCA is possible, but it has not been developed so far.

It is interesting that the soft PLS-DA stands in between being slightly better than SIMCA, but worse than the hard PLS-DA. This effect will be explained in the following section, in which we show that the soft PLS-DA has features that make it similar to SIMCA.

11 | ONE AGAINST "ALL"

This is a very popular PLS-DA strategy, when 1 target class (genuine) is discriminated against a collection of all available alternative classes (aliens). It is believed that such an approach could solve the authenticity problem better than a one-class classifier. We will investigate this method using *Pills* dataset, in which class *A4* is considered as the target, whereas other classes, *A1 to A3* plus *A5 to A6*, are used together as the second "all" class. Class *A7* is kept separately and is employed as a new class that should be predicted using the established "one vs. all" model. Plot (A) in Figure 8 shows the graphical results of PLS-DA modeling with 5 LVs. In the soft version $\alpha = 0.05$.

Numerical values are presented in Table 5.

It can be seen that the hard method perfectly discriminates the "one" class *A4* from the "all" class, but it wrongly attributes a new class *A7* as a member of class *A4*. The soft PLS-DA looks much better. At the training stage, it develops an appropriate model with an efficiency of 97% that fits into the given α value, because $0.97 \approx 1 - 0.05$. At the prediction

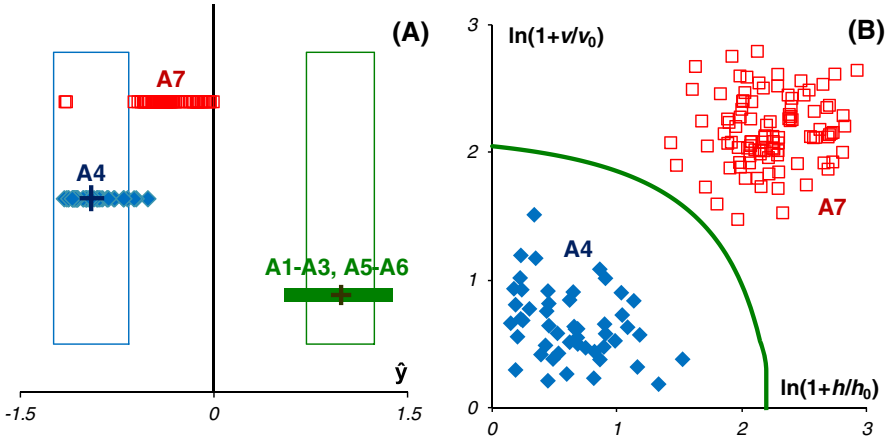

FIGURE 8 The "one against all" classification. Plot (A): PLS-DA, plot (B): SIMCA, 3 PCs

TABLE 5 Figures of merit for the "one against all" example

Figures of merit	Hard PLS-DA A4	Hard PLS-DA All	Soft PLS-DA A4	Soft PLS-DA All	SIMCA A4
	Training of A4 vs. all				Training of A4
Class sensitivity	100%	100%	93%	94%	100%
Class specificity	100%	100%	100%	100%	100%
Class efficiency	100%	100%	97%	97%	100%
Total sensitivity	100%		94%		100%
Total specificity	100%		100%		100%
Total efficiency	100%		97%		100%
	Prediction of A7- an extraneous class				
Class specificity	0%	100%	100%	100%	100%
Total specificity	0%		98%		100%

stage, the soft PLS-DA correctly classifies 98% of objects in class *A7* as aliens, rejecting them both from the "one" class, and from the "all" class. Therefore, we can conclude that, in this case, the soft PLS-DA is able to solve the authenticity problem, but the hard version cannot do this.

The last issue we have to explore is the superiority of (soft) PLS-DA over SIMCA in authentication. Plot (B) in Figure 8 presents the graphical results of the SIMCA-based authentication. In this model, only class *A4* is used at the training stage; the class *A7* data are utilized at the prediction stage and the "all" class data are not involved at all. The SIMCA model is developed for the following parameters: the number of PCA PCs = 3, type I error $\alpha = 0.0001$. Note the extremely low value of α, which ensures the correspondingly high level of sensitivity, as large as 99.99%. At the same time, the theoretically calculated[27] value of the type II error, β, for set *A7* equals 0.004, which corresponds to specificity of 99.6%.

Summarizing, we conclude that PLS-DA strategy of "one against all" may be a reasonable method for authentication when the soft version is applied, but not in the case of the hard one. Even though, SIMCA remains a better approach for solving the authenticity problems.

12 | SOFTWARE

So far, the proposed PLS-DA methods are not implemented as the Matlab programs—this is work in progress.[28] Meanwhile, those readers who practice Chemometric Add-In for Excel[20] are kindly directed to web page[29] where all relevant information (templates, supplementary Excel add-in, instructions) is presented.

13 | CONCLUSIONS

1. We propose the multi-class version of PLS-DA, which, in fact, is not more complex than the conventional binary (2-class) PLS-DA. The method does not utilize the PLS scores but is entirely based on the predicted dummy responses. To get around the degeneracy of this matrix, we suggest using PCA that converts the response matrix $\hat{\mathbf{Y}}$ into the score matrix **T.** These scores can be employed for classification by any appropriated method. Therefore, PLS-DA should be considered as a method of feature extraction from high-dimensional X space into low-dimensional T space than a method of discrimination.
2. As examples, we introduce 2 discrimination method based on the **T** data. The first is a conventional hard PLS-DA approach based on LDA. We also propose the novel soft version of the PLS-DA method, which is based on QDA applied to the **T** data. In this version, discrimination rule employs the Mahalanobis distances and a threshold, which is calculated for a given type I error. According to this rule, a sample can be simultaneously attributed to several classes, or it may be not allocated at all. It was demonstrated that the soft PLS-DA is able to avoid misclassification, in case a new object is not a member of any target class.
3. The principal measures of classification quality (sensitivity, specificity, and efficiency) are defined for the multi-class PLS-DA. It is also shown how these characteristics are used for the selection of the complexity of the model, which, in case of PLS-DA, is the number of the PLS latent variables.
4. A popular opinion that an equal number of objects in the training classes is preferred for a good PLS-DA model is analyzed and found to be wrong. In fact, PLS-DA utilizes a regression approach, the efficiency of which depends primarily on the design of the experiment, rather than on the size of data.
5. The comparison of the discriminant (PLS-DA) and the class-modeling (SIMCA) methods is conducted using the simulated and real-world examples. In particular, it is shown that SIMCA is better when 1 class is tight, eg, it comprises healthy biological tissues, and the other class is broad, and, in fact, does not form a class at all, eg, it contains tissues damaged in various ways. On the contrary, PLS-DA is preferable in cases when we separate 2 classes with the same major components but different impurities.
6. We considered a very popular PLS-DA strategy, when 1 target class is discriminated against a collection of all available alternative classes. It is demonstrated that this approach may be a reasonable method for authentication when the soft PLS-DA is applied, but not in the case of the hard one. Nevertheless, SIMCA remains a better approach for solving the authentication problems.

Finally, we can repeat our notion presented in Rodionova et al.[13] "The "best" classification method does not exist. Every task at hand requires an application of a pertinent chemometric method best suited to answer the posed question."

ACKNOWLEDGEMENTS

We acknowledge partial funding from the International Atomic Energy Agency in the frame of projects D5240 and G42007, and a partial support within the Russian state assignment AAAA-A18-118020690203-8.

ORCID

Alexey L. Pomerantsev http://orcid.org/0000-0001-7402-4011
Oxana Ye. Rodionova http://orcid.org/0000-0002-0146-8284

REFERENCES

1. Boulesteix A-L, Strimmer K. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Brief Bioinform*. 2007;8:32-44.
2. Verhoeckx KC, Gaspari M, Bijlsma S, et al. In search of secreted protein biomarkers for the anti-inflammatory effect of beta2-adrenergic receptor agonists: application of DIGE technology in combination with multivariate and univariate data analysis tools. *J Proteome Res*. 2005;4(6):2015-2023.
3. Bijlsma S, Bobeldijk I, Verheij ER, et al. Large-scale human metabolomics studies: a strategy for data (pre-) processing and validation. *Anal Chem*. 2006;78(2):567-574.
4. Granato D, Margraf T, Brotzakis I, Capuano E, van Ruth SM. Characterization of conventional, biodynamic, and organic purple grape juices by chemical markers, antioxidant capacity and instrumental taste profile. *J Food Sci*. 2015;80:55-65.

5. Sacré P-Y, Deconinck E, De Beer T, et al. Comparison and combination of spectroscopic techniques for the detection of counterfeits. *J Pharm Biomed Anal*. 2010;53(3):445-453.
6. Stahle L, Wold S. Partial least squares analysis with cross-validation for the two-class problem: a Monte Carlo study. *J Chemometr*. 1987;1(3):185-196.
7. Barker M, Rayens WS. Partial least squares for discrimination. *J Chemometr*. 2003;17(3):166-173.
8. Indahl UG, Martens H, Næs T. From dummy regression to prior probabilities in PLS-DA. *J Chemometr*. 2007;21(12):529-536.
9. Nocairi H, Qannari EM, Vigneau E, Bertrand D. Discrimination on latent components with respect to patterns. Application to multicollinear data. *Comput Stat Data Anal*. 2005;48(1):139-147.
10. Westerhuis JA, Hoefsloot HCJ, Smit S, et al. Assessment of PLSDA cross validation. *Metabolomics*. 2008;4(1):81-89. https://doi.org/10.1007/s11306-007-0099-6
11. Kjeldahl K, Bro R. Some common misunderstandings in chemometrics. *J Chemometr*. 2010;24(7-8):558-564.
12. Perez NF, Ferre J, Boque R. Multi-class classification with probabilistic discriminant partial least squares (p-DPLS). *Anal Chim Acta*. 2010;664(1):27-33.
13. Rodionova OY, Titova AV, Pomerantsev AL. Discriminant analysis is an inappropriate method of authentication. *Trends Anal Chem*. 2016;78(4):17-22.
14. Bylesjo M, Rantalainen V, Cloarec O, Nicholson JK, Holmes E, Trygg J. OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. *J Chemometr*. 2006;20(8-10):341-351.
15. Rodionova OY, Balyklova KS, Titova AV, Pomerantsev AL. Quantitative risk assessment in classification of drugs with identical API content. *J Pharm Biomed Anal*. 2014;98:186-192.
16. Oliveri P, López MI, Casolino MC, et al. Partial least squares density modeling (PLS-DM)—a new class-modeling strategy applied to the authentication of olives in brine by near-infrared spectroscopy. *Anal Chim Acta*. 2014;851:30-36. https://doi.org/10.1016/j.aca.2014.09.013
17. Rodionova OY, Oliveri P, Pomerantsev AL. Rigorous and compliant approaches to one-class classification. *Chemom Intel Lab Syst*. 2016;159:89-96.
18. Fidelis M, Santos JS, Kincheski Coelho AL, Rodionova OY, Pomerantsev A, Granato D. Authentication of juices from antioxidant and chemical perspectives: a feasibility quality control study using chemometrics. *Food Control*. 2017;73:796-805.
19. Martens H, Naes T. *Multivariate Calibration*. New York: Wiley; 1989.
20. Pomerantsev AL. *Chemometrics in Excel*. Hoboken NJ: John Wiley & Sons; 2014.
21. Oliveri P, Downey G. Multivariate class modeling for the verification of food-authenticity claims. *TrAC—Trends Anal Chem*. 2012;35:74-86.
22. Brereton RG, Lloyd GR. Partial least squares discriminant analysis: taking the magic away. *J Chemometr*. 2014;28(4):213-225.
23. Krakowska B, Custers D, Deconinck E, Daszykowski M. The Monte Carlo validation framework for the discriminant partial least squares model extended with variable selection methods applied to authenticity studies of Viagra ® based on chromatographic impurity profiles. *Analyst*. 2016;141(3):1060-1070.
24. Forina P, Oliveri S. Class-modeling techniques, classic and new, for old and new problems. *Chemom Intel Lab Syst*. 2008;93(2):132-148.
25. Pomerantsev A. Acceptance areas for multivariate classification derived by projection methods. *J Chemometr*. 2008;22(11-12):601-609.
26. Pomerantsev AL, Rodionova OY. Concept and role of extreme objects in PCA/SIMCA. *J Chemometr*. 2014;28(5):429-438.
27. Pomerantsev AL, Rodionova OY. On the type II error in SIMCA method. *J Chemometr*. 2014;28(6):518-522.
28. MatLab for PLS-DA. https://github.com/yzontov/pls-da
29. PLS-DA templates. http://chemometrics.chph.ras.ru/PLSDA/

How to cite this article: Pomerantsev AL, Rodionova OY. Multiclass partial least squares discriminant analysis: Taking the right way—A critical tutorial. *Journal of Chemometrics*. 2018;32:e3030. https://doi.org/10.1002/cem.3030

APPENDIX

PROVING OF STATEMENTS

Notations

Row-wise vector notation is used in the appendix in the same manner as in the main text. Matrix $\mathbf{U}$ and vector $\mathbf{u}$ consist of units, matrix $\mathbf{E}$ is the identity matrix. Matrix $\mathbf{M} = \mathbf{u}^t\mathbf{m}$, where $\mathbf{m}$ is the vector of mean values of $\mathbf{Y}$. Dimensionality of

these objects depends on the context. The $(I \times K)$ matrix $\hat{\mathbf{Y}}$ contains the PLS2 regression dummy responses predicted at the training stage. Matrix $\mathbf{Z} = \hat{\mathbf{Y}} - \mathbf{M}$. The $(I \times K-1)$ matrix $\mathbf{T}$ and the $(K \times K-1)$ matrix $\mathbf{P}$ represent the PCA decomposition $\mathbf{Z} = \mathbf{TP}^t$. $\mathbf{\Lambda} = \mathbf{T}^t\mathbf{T}$ is the $(K-1 \times K-1)$ diagonal matrix. The $(I \times K-1)$ matrix $\mathbf{C} = (\mathbf{E} - \mathbf{M})\mathbf{P}$ is the projection of matrix $\mathbf{E}$ onto the PCA space. The $(K \times K-1)$ matrix

$$\mathbf{W} = \begin{vmatrix} \mathbf{w}_1 \\ \cdots \\ \mathbf{w}_K \end{vmatrix}$$

consists of vectors $\mathbf{w}_k$, the $(1 \times K)$ vector $\mathbf{v} = (v_1,..,v_K)^t$ consists of values v_k. They are defined in Equation 9. The $(1 \times K)$ vector $\mathbf{t}_0$ is an intersection of all hyper planes defined in Equation 8.

Properties of Z

$\mathbf{ZU} = \mathbf{0}$ and rank$(\mathbf{Z}) = K - 1$. Dimension of the null space of $\mathbf{Z}$, null$(\mathbf{Z})$, is 1, and vector $\mathbf{u}$ forms its basis, $\mathbf{Zu} = \mathbf{0}$.

Properties of P

$$\mathbf{UP} = \mathbf{0} \text{ and } \mathbf{P}^t\mathbf{P} = \mathbf{E}$$

Lemma $\mathbf{PP}^t = \mathbf{E} - q\mathbf{U}$, where $q = 1/K$.

Proof $\mathbf{\Pi} = \mathbf{PP}^t$ is a projection matrix because $\mathbf{\Pi}^2 = \mathbf{\Pi}$ and $\mathbf{\Pi}^t = \mathbf{\Pi}$. Because $\mathbf{Z} - \mathbf{TP}^t = \mathbf{0}$, then $\mathbf{Z}(\mathbf{E} - \mathbf{\Pi}) = \mathbf{0}$, and $\mathbf{E} - \mathbf{\Pi} = \mathbf{u}^t\mathbf{q}$, where $\mathbf{q} = (q_1, q_2,...,q_K)$. Because $\mathbf{\Pi}$ is symmetric, then $q_1 = q_2 = ... = q_K = q$. Therefore, $\mathbf{\Pi} = \mathbf{E} - q\mathbf{U}$. Because $\mathrm{tr}(\mathbf{\Pi}) = \mathrm{rank}(\mathbf{\Pi}) = K - 1$, then $q = 1/K$.

Theorem 1. Matrix $\mathbf{C}$ belongs to the PCA space.

Proof We should prove that $\mathbf{M} + \mathbf{CP}^t = \mathbf{E}$. Using Lemma we obtain
$\mathbf{M} + \mathbf{CP}^t = \mathbf{M} + (\mathbf{E} - \mathbf{M})\mathbf{PP}^t = \mathbf{M} + (\mathbf{E} - \mathbf{M}) \times (\mathbf{E} - q\mathbf{U}) = \mathbf{M} + (\mathbf{E} - \mathbf{M}) - q\,(\mathbf{E} - \mathbf{M})\mathbf{U} = \mathbf{E} - q\,(\mathbf{U} - \mathbf{U}) = \mathbf{E}$.

Theorem 2. $\mathbf{t}_0 = \mathbf{vP\Lambda}$.

Proof The system of Equation 8

$$(\mathbf{w}_k - \mathbf{w}_l)\mathbf{t}^t = v_k - v_l, \quad k = 1, ..., K-1; l>k$$

can be represented using a matrix notation.

$$\mathbf{DWt}^t = \mathbf{Dv}^t \tag{A1}$$

The $(L \times K)$ matrix $\mathbf{D}$ $(L = K(K-1)/2)$ has a special structure. Each row of matrix $\mathbf{D}$ contains zeros except 2 columns: k that contains +1, and l that contains −1; $k = 1,..., K-1$; $l = k+1,.., K$. It can be shown that

$$\mathbf{D}(\mathbf{E} - \mathbf{M}) = \mathbf{D}; \quad \mathbf{D}^t\mathbf{D} = K\mathbf{E} - \mathbf{U} \tag{A2}$$

Considering that $\mathbf{W} = (\mathbf{E} - \mathbf{M})\mathbf{P\Lambda}^{-1}$ we get

$$\mathbf{DW} = \mathbf{D}(\mathbf{E} - \mathbf{M})\mathbf{P\Lambda}^{-1} = \mathbf{DP\Lambda}^{-1}$$

Multiplying Equation A1 by $\mathbf{P}^t\mathbf{D}^t$ and accounting for Equation A2, we obtain

$$\mathbf{P}^t(K\mathbf{E} - \mathbf{U})\mathbf{P\Lambda}^{-1}\mathbf{t}^t = \mathbf{P}^t(K\mathbf{E} - \mathbf{U})\mathbf{v}^t$$

Because $\mathbf{UP} = \mathbf{0}$ and $\mathbf{P}^t\mathbf{P} = \mathbf{E}$,

$$\mathbf{t} = \mathbf{vP\Lambda}.$$