Analytica Chimica Acta 851 (2014) 30–36

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Partial least squares density modeling (PLS-DM) – A new class-modeling strategy applied to the authentication of olives in brine by near-infrared spectroscopy

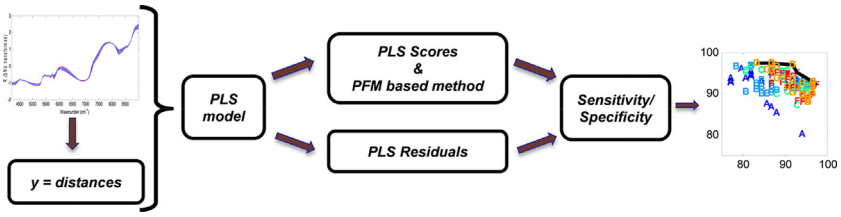
Paolo Oliveri [a,1,*], M. Isabel López [b,1], M. Chiara Casolino [a], Itziar Ruisánchez [b], M. Pilar Callao [b], Luca Medini [c], Silvia Lanteri [a]

[a] Department of Pharmacy, University of Genoa, Via Brigata Salerno, 13, I-16147 Genoa, Italy
[b] Chemometrics, Qualimetric and Nanosensors Group, Department of Analytical and Organic Chemistry, Rovira i Virgili University, Marcellí Domingo s/n, 43007 Tarragona, Spain
[c] Special Company for Professional Training and Technological and Commercial Promotion of the Chamber of Commerce of Savona, Regione Rollo, 98, I-17031 Albenga, SV, Italy

HIGHLIGHTS

- PLS-DM is presented as a new class modeling technique.
- It combines partial least squares, potential function probability and Q statistics.
- Model parameters were optimized by applying the Pareto optimality criterion.
- PLS-DM was applied to authentication of olives in brine.
- It provided more efficient and balanced results than classical modeling methods.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 30 July 2014
Received in revised form 4 September 2014
Accepted 9 September 2014
Available online 16 September 2014

Keywords:
Class-modeling
One-class classifier
Density estimation
Partial least squares (PLS)
Potential functions

ABSTRACT

A new class-modeling method, referred to as partial least squares density modeling (PLS-DM), is presented. The method is based on partial least squares (PLS), using a distance-based sample density measurement as the response variable. Potential function probability density is subsequently calculated on PLS scores and used, jointly with residual Q statistics, to develop efficient class models. The influence of adjustable model parameters on the resulting performances has been critically studied by means of cross-validation and application of the Pareto optimality criterion. The method has been applied to verify the authenticity of olives in brine from cultivar *Taggiasca*, based on near-infrared (NIR) spectra recorded on homogenized solid samples. Two independent test sets were used for model validation. The final optimal model was characterized by high efficiency and equilibrate balance between sensitivity and specificity values, if compared with those obtained by application of well-established class-modeling methods, such as soft independent modeling of class analogy (SIMCA) and unequal dispersed classes (UNEQ).

* Corresponding author. Tel.: +39 10 3532374; fax: +39 10 3532684.
E-mail addresses: oliveri@dictfa.unige.it, oliveri@difar.unige.it (P. Oliveri).
[1] These two authors contributed equally.

http://dx.doi.org/10.1016/j.aca.2014.09.013
0003-2670/

1. Introduction

Class-modeling and discriminant classification methods are widely employed to build mathematical models aimed at characterizing samples with respect to qualitative properties. Discriminant classification techniques are used to determine to which class, among a number of pre-defined classes, a sample most probably belongs, by setting a delimiter between the pairs of classes. Each new sample is then always assigned to one of the categories, even in the case of samples that do not belong to any class studied. Also, class-modeling techniques can be used for multiclass classification but, in this case, each new sample can be either assigned to one, more than one or none of the predefined classes. The suitable selection of the classification strategy depends on the problem to be solved and it may represent an important issue to be considered when working with these techniques [1]. As a matter of fact, the modeling of a single class of interest, to verify whether a sample is compatible or not with the characteristics of that class [2,3], is only allowed by the class-modeling techniques, which – for this reason – are also referred to as one-class classifiers [4] or untargeted modeling methods [5].

In multi-class classification, the discriminant approach is followed more frequently than the class-modeling one. A discriminant classification method which has gained increasing attention in the last years is based on partial least squares (PLS) regression, and it is usually referred to as discriminant PLS (D-PLS) or PLS discriminant analysis (PLS-DA) [6–8]. In the recent years, a number of attempts have been addressed to develop class-modeling techniques exploiting the advantages offered by the PLS method [9–12].

In particular, a method called one-class PLS (OC-PLS) has been recently presented, in which a PLS model is built using a constant response ($y = 1$), i.e., identical values for all of the training samples belonging to the class of interest [13]. Hotelling's T^2 and Q statistics are used to verify compliance of test samples with the class model. Such a method proved to be efficient on different data sets [14–16]. Nonetheless, some issues have to be considered. First of all, PLS regression on column-centered x-block data leads to degenerate solutions when the response variable is constant [13]. Secondly, this strategy gives equal importance to all samples in the class model definition, without taking into account class heterogeneity. Finally, the use of T^2 statistics on the PLS scores implies the underlying hypothesis of a normal distribution.

In order to manage data with non-normal and non-uniform distributions, some class-modeling methods have been proposed. Among the most efficient, potential function methods (PFM) [17], a family of probabilistic non-parametric techniques, define the class model by empirically estimating a probability density distribution for a class of interest [18,19]. An important limitation of such techniques is related to the impossibility of direct application to data sets with high variable dimensionality. In fact, the higher the number of variables, the lower the reliability in the estimation of the probability density, as well as the establishment of the critical value for the decision rule. In order to overcome this hurdle, PFM are commonly applied after unsupervised variable reduction by means of principal component analysis (PCA) [19].

In the present study, a new PLS-based class-modeling strategy is presented, called partial least squares density modeling (PLS-DM), which combines the features of PLS and PFM, together with Q statistics, to obtain highly efficient class models. The method was applied on a set of near-infrared (NIR) spectra recorded on samples of olives in brine, with the purpose of verifying the authenticity of olives from cultivar *Taggiasca*. In this application – like in most of the cases involving verification of food authenticity claims – the focus was on a single class (cultivar *Taggiasca*). In such a case, the discriminant approach would require the collection of two sets of training samples: one representative of the *Taggiasca* olives and a second representative of the entire production of all of the olives potentially usable to make frauds. Such a condition is rarely realizable in practice, and collected sets of non-compliant samples would be under-representative of the whole non-compliance possibilities. This inadequacy would inevitably lead to biased decision rules, the outcomes of which being heavily dependent on those samples included in the non-compliant set. For this reason, in such a case, decisions regarding sample conformity based on class-modeling strategies are more robust and suitable than those based on discriminant approaches [20].

Olives from cultivars *Leccino* and *Coquillo*, being morphologically very similar to *Taggiasca*, are suspected to be used in fraudulent manufactures and, therefore, they were considered in this study as potential adulterants.

Model performances are evaluated in terms of sensitivity and specificity and by application of the Pareto optimality criterion. Results are compared with those achieved by unequal dispersed classes (UNEQ) [21], soft independent modeling of class analogy (SIMCA) [22] and OC-PLS [13]. UNEQ and SIMCA are the class-modeling techniques most commonly applied in chemometrics, while OC-PLS represents the most recent modeling method based on PLS.

2. Theory

2.1. Partial least squares

Partial least squares (PLS) is a multivariate regression technique which computes directions in the space of the predictors ($\mathbf{X}$) characterized by the maximum covariance with the response variable ($\mathbf{y}$). Such directions, called latent variables (LVs), are employed to define the regression model:

$$\mathbf{X}_{I,V} = \mathbf{T}_{I,L}[\mathbf{P}_{L,V}]^T + \mathbf{E}_{I,V} \tag{1}$$

$$\mathbf{y}_{I,1} = \mathbf{U}_{I,L}[\mathbf{q}_{L,1}]^T + \mathbf{f}_{I,1} \tag{2}$$

where I is the number of samples, V is the number of original predictor variables, $\mathbf{T}$ and $\mathbf{U}$ are the matrices of scores (i.e., projections) of $\mathbf{X}$ and $\mathbf{y}$, respectively, $\mathbf{P}$ and $\mathbf{q}$ contain the loading terms, $\mathbf{E}$ and $\mathbf{f}$ contain the error terms. The most appropriate number of LVs (L) is usually determined in a cross-validation (CV) cycle, by studying the evolution of the quality parameter – such as the prediction error – as a function of the increasing number of LVs [23].

2.2. Residuals and confidence limit

Q statistics is related to the residuals, i.e., the fraction of the information about samples not explained by the L latent variables retained in the final PLS model:

$$Q_i = \mathbf{e}_i\mathbf{e}_i^T \tag{3}$$

where $\mathbf{e}_i$ is the vector of residuals of sample i after applying the PLS model. Its confidence limit, Q_α, is computed according to Jackson [24]:

$$Q_\alpha = \theta_1\left[\frac{z_\alpha\sqrt{2\theta_2 h_0^2}}{\theta_1} + 1 + \frac{\theta_2 h_0(h_0-1)}{\theta_1^2}\right]^{1/h_0} \tag{4}$$

where z_α is the value of the standard normal deviate corresponding to the upper $(1-\alpha)$ percentile, and θ_j terms and h_0 are defined as: