PARTIAL LEAST SQUARES DISCRIMINANT ANALYSIS: A DIMENSIONALITY REDUCTION METHOD TO CLASSIFY HYPERSPECTRAL DATA

Mario Fordellone[1]
Department of Statistical Science, Sapienza University of Rome, Rome, Italty

Andrea Bellincontro, Fabio Mencarelli
Department for Innovation in Biological, Agro-food and Forest Systems, University of Tuscia, Viterbo, Italy

Abstract *The recent development of more sophisticated spectroscopic methods allows acquisition of high dimensional datasets from which valuable information may be extracted using multivariate statistical analyses, such as dimensionality reduction and automatic classification (supervised and unsupervised). In this work, a supervised classification through a partial least squares discriminant analysis (PLS-DA) is performed on the hyperspectral data. The obtained results are compared with those obtained by the most commonly used classification approaches.*

Keywords: *PLS-DA, hyperspectral data, high dimensional data, NIR, PLSR.*

1. INTRODUCTION

The recent development of more sophisticated spectroscopic approaches allows the acquisition of high dimensional datasets from which valuable information may be extracted via different multivariate statistical techniques. The high data dimensionality greatly enhances the informational content of the dataset and provides an additional opportunity for the current techniques for analyzing such data (Jimenez and Landgrebe, 1998). For example, automatic classification (clustering and/or classification) of data with similar features is an important problem in a variety of research areas such as biology, chemistry, and medicine (Galvan et al., 2006; Hardy et al., 2006). When the labels of the clusters are available, a supervised classification method is applied. Several classification techniques are available and described in the literature. However, data derived by spectroscopic detection represent a hard challenge for the researcher, who faces two crucial problems: data dimensionality larger than the observations, and high correlation levels among the variables (multicollinearity).

[1]Corresponding author: Mario Fordellone mario.fordellone@uniroma1.it

Usually, in order to solve these problems *(i)* a first data compression or reduction method, such as principal component analysis (PCA) is applied to shrink the number of variables; then, a range of discriminant analysis techniques is used to solve the classification problem, while *(ii)* in other cases, non-parametric classification approaches are used (Agrawal et al., 1998; Bühlmann and Van De Geer, 2011; Ding and Gentleman, 2005; Jimenez and Landgrebe, 1998; Kriegel et al., 2009).

In this work, the dataset consists of three different varieties of olives (*Moraiolo*, *Dolce di Andria*, and *Nocellara Etnea*) monitored during ripening up to harvest (Bellincontro et al., 2012). Samples contained olives from 162 trees (54 for each variety), and 601 spectral detections (i.e., dimensions/variables) were performed using a portable near infrared acousto-optically tunable filter (NIR-AOTF) device in diffuse reflectance mode from 1100 nm to 2300 nm with an interval of 2. The use of NIRS on olive fruits and related products is already known; applications for the determination of oil and moisture content are now considered routine analyses in comparison with relatively new methodologies, such as nuclear magnetic resonance (NMR), or more traditional analytical determinations (Cayuela and Camino, 2010; Gallardo et al., 2005; Garcia et al., 1996; León et al., 2004).

This paper is based on the use of partial least squares discriminant Analysis (PLS-DA). However, for comparison purposes, we also analyze the results obtained by other commonly used non-parametric classification models such as K-nearest neighbor (KNN), support vector machine (SVM) (Balabin et al., 2010; Joachims, 2005; Misaki et al., 2010; Tran et al., 2006), and some variants of discriminant functions for sparse data as such as diagonal linear discriminant analysis (DLDA), maximum uncertainty linear discriminant analysis (MLDA), and shrunken linear discriminant analysis (SLDA). All the three regularization techniques compute linear discriminant functions (Clemmensen et al., 2011; Dudoit et al., 2002; Fisher and Sun, 2011; Guo et al., 2006; Hastie et al., 1995; Thomaz et al., 2006).

PLS-DA is a dimensionality reduction technique, a variant of partial least squares regression (PLS-R) that is used when the response variable is categorical. It is a compromise between the usual discriminant analysis and a discriminant analysis on the principal components of the predictor variables. In particular, PLS-DA instead of finding hyperplanes of maximum variance between the response and independent variables finds a linear regression model by projecting the predicted variables and the observed variables into a new space. PLS-DA can provide good insight into the causes of discrimination via weights and

loadings, which gives it a unique role in exploratory data analysis, for example in metabolomics via visualization of significant variables such as metabolites or spectroscopic peaks (Brereton and Lloyd, 2014; Kemsley, 1996; Wehrens and Mevik, 2007).

The paper is structured as follows: in section 2 we provide a background on the most commonly used non-parametric statistical methodologies to solve the classification problem of sparse data (i.e., KNN and SVM) and an overview of different classifiers derived from linear discriminant analysis (LDA), in section 3 we focus on the PLS-DA model with a deeper examination of the PLS algorithm, in section 4 we show a comparison of the results obtained by the application of PLS-DA and those obtained by the other common classification methods, and finally in section 5 we provide some suggestions and ideas for future research.

2. BACKGROUND

In this section, we present a brief overview of different classifiers that have been highly successful in handling high dimensional data classification problems, starting with popular methods such as K-nearest neighbor (KNN) and support vector machines (SVM) (Dudoit et al., 2002; Zhang et al., 2006) and variants of discriminant functions for sparse data (Clemmensen et al., 2011). We also examine dimensionality reduction techniques and their integration with some existing algorithms (i.e., partial least squares discriminant analysis (PLS-DA)) (Brereton and Lloyd, 2014; Kemsley, 1996).

2.1. K-NEAREST NEIGHBOR (KNN)

The KNN method was first introduced by Fix and Hodges (Fix and Hodges, 1989) based on the need to perform discriminant analysis when reliable parametric estimates of probability densities are unknown or difficult to determine. In this method, a distance measure (e.g., Euclidean) is assigned between all points in the data. The data points, K-closest neighbors (where K is the number of neighbors), are then found by analyzing a distance matrix. The K-closest data points are then found and analyzed in order to determine which class label is the most common among the set. Finally, the most common class label is then assigned to the data point being analyzed (Balabin et al., 2010).

The KNN classifier is commonly based on the Euclidean distance between a test sample and the specified training samples. Formally, let x_i be an input sample with J features $(x_{i,1},\ldots,x_{i,J})$, and n be the total number of input samples $(i = 1,\ldots,n)$. The Euclidean distance between sample x_i and x_l $(l = 1,\ldots,n)$ is

defined as

$$d(x_i, x_l) = \sqrt{(x_{i,1} - x_{l,1})^2 + \cdots + (x_{i,J} - x_{l,J})^2}. \quad (1)$$

Using the latter characteristic, the KNN classification rule is to assign to a test sample the majority category label of its K nearest training samples. In other words, K is usually chosen to be odd, so as to avoid ties. The $K = 1$ rule is generally called the 1-nearest-neighbor classification rule.

Then, let x_i be a training sample and x_i^* be a test sample, and let ω be the true class of a training sample and $\hat{\omega}$ be the predicted class for a test sample ($\omega, \hat{\omega} = \ldots, \Omega$), where Ω is the total number of classes. During the training process, only the true class ω of each training sample to train the classifier is used, while during testing the class $\hat{\omega}$ of each test sample is predicted. With 1-nearest neighbor rule, the predicted class of test sample x_i^* is set equal to the true class ω of its nearest neighbor, where z_i is a nearest neighbor to x_i^* if the distance

$$d(z_i, x_i^*) = \min_j \{d(z_j, x_i^*)\}. \quad (2)$$

For the K-nearest neighbors rule, the predicted class of test sample x_i^* is set equal to the most frequent true class among the K nearest training samples.

2.2. SUPPORT VECTOR MACHINE (SVM)

The SVM approach was developed by Vapnik (Cortes and Vapnik, 1995; Suykens and Vandewalle, 1999). Synthetically, SVM is a linear method in a very high dimensional feature space that is nonlinearly related to the input space. The method maps input vectors to a higher dimensional space where a maximal separating hyperplane is constructed (Joachims, 2005). Two parallel hyperplanes are constructed on each side of the hyperplane that separates the data and maximizes the distance between the two parallel hyperplanes. An assumption is made that the larger the margin or distance between these parallel hyperplanes, the better the generalization error of the classifier will be.

SVM was initially designed for binary classification. To extend SVM to the multi-class scenario, a number of classification models were proposed (Wang and Xue, 2014). Formally, given training vectors $x_i \in \Re^J$, $i = 1, \ldots, n^*$, in two classes, and the label vector $Y \in \{-1, 1\}^{n^*}$ (where n^* in the size of the training samples), the support vector technique requires the solution of the following optimization

problem:

$$\begin{aligned} &\min_{w\in H, b\in\Re, \xi_i\in\Re} \frac{1}{2} w^T w + C\sum_{i=1}^{n^*} \xi_i, \\ &subject\ to \quad y_i(w^T\varphi(x_i)+b) \geq 1-\xi_i \\ &\qquad\qquad \xi_i \geq 0, \quad i=1,\ldots,n^*, \end{aligned} \tag{3}$$

where $w \in \Re^J$ is the weights vector, $C \in \Re_+$ is the regularization constant, and the mapping function φ projects the training data into a suitable feature space H.

For a K-class problem, many methods use a single objective function for training all K-binary SVMs simultaneously and maximize the margins from each class to the remaining ones (Wang and Xue, 2014; Weston and Watkins, 1998). An example is the formulation proposed by Weston and Watkins (Weston and Watkins, 1998). Given a labeled training set represented by $\{(x_1,y_1),\ldots,(x_{n^*},y_{n^*})\}$, where $x_i \in \Re^J$ and $y_i \in \{1,\ldots,K\}$, this formulation is given as follows:

$$\begin{aligned} &\min_{w_k\in H, b\in\Re^K, \xi\in\Re^{n^*\times K}} \frac{1}{2}\sum_{k=1}^{K} w_k^T w_k + C\sum_{i=1}^{n^*}\sum_{t\neq y_i} \xi_{i,t}, \\ &subject\ to \quad w_{y_i}^T\varphi(x_i)+b_{y_i} \geq w_t^T\varphi(x_i)+b_t+2-\xi_{i,t}, \\ &\qquad\qquad \xi_{i,t} \geq 0, \quad i=1,\ldots,n^*, \quad t\in\{1,\ldots,K\}. \end{aligned} \tag{4}$$

The resulting decision function is given in Equation 5 (Wang and Xue, 2014).

$$\underset{k}{\operatorname{argmax}}\, f_m(x) = \underset{k}{\operatorname{argmax}}(w_k^T\varphi(x_i)+b_k). \tag{5}$$

2.3. DISCRIMINANT ANALYSIS FUNCTIONS

In this section we present a comprehensive overview of different classifiers derived by Linear Discriminant Analysis (LDA), and that have been highly successful in handling high dimensional data classification problems: Diagonal Linear Discriminant Analysis (DLDA), Maximum uncertainty Linear Discriminant Analysis (MLDA), and Shrunken Linear Discriminant Analysis (SLDA). All the three regularization techniques compute Linear Discriminant Functions, by default after a preliminary variable selection step, based on alternative estimators of a within-groups covariance matrix that leads to reliable allocation rules in problems where the number of selected variables is close to, or larger than, the number of available observations.

The main purpose of discriminant analysis is to assign an unknown subject to one of K classes on the basis of a multivariate observation $x = (x_1,\ldots,x_J)'$,

where J is the number of variables. The standard LDA procedure does not assume that the populations of the distinct groups are normally distributed, but it assumes implicitly that the true covariance matrices of each class are equal because the same within-class covariance matrix is used for all the classes considered (Thomaz et al., 2006; Wichern and Johnson, 1992). Formally, let S_b be the between-class covariance matrix defined as

$$S_b = \sum_{k=1}^{K} n_k (\bar{x}_k - \bar{x})(\bar{x}_k - \bar{x})^T, \tag{6}$$

and let S_w be the within-class covariance matrix defined as

$$S_w = \sum_{k=1}^{K} (n_k - 1) S_k = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (\bar{x}_{k,i} - \bar{x}_k)(\bar{x}_{k,i} - \bar{x}_k)^T, \tag{7}$$

where $x_{k,i}$ is the J-dimensional pattern i from the k-th class, n_k is the number of training patterns from the k-th class, and K is the total number of classes (or groups) considered. The vector $\bar{x}_k$ and matrix S_k are respectively the unbiased sample mean and sample covariance matrix of the k-th class, while the vector $\bar{x}$ is the overall unbiased sample mean given by

$$\bar{x} = \frac{1}{n} \sum_{k=1}^{K} n_k \bar{x}_k = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} x_{k,i}, \tag{8}$$

where n is the total number of samples $n = n_1 + \cdots + n_K$.

Then, the main objective of LDA is to find a projection matrix (here defined as P_{LDA}) that maximizes the ratio of the determinant of the between-class scatter matrix to the determinant of the within-class scatter matrix (Fisher's criterion). Formally,

$$P_{LDA} = \underset{P}{\operatorname{argmax}} \frac{det\left(P^T S_b P\right)}{det\left(P^T S_w P\right)}. \tag{9}$$

It has been shown (Devijver and Kittler, 1982) that Equation (9) is in fact the solution of the following eigenvector system problem:

$$S_b P - S_w P \Lambda = 0. \tag{10}$$

Note that by multiplying both sides by S_w^{-1}, Equation (10) can be rewritten as

$$\begin{aligned} & S_w^{-1} S_b P - S_w^{-1} S_w P \Lambda = 0 \\ & S_w^{-1} S_b P - P \Lambda = 0 \\ & (S_w^{-1} S_b) P = P \Lambda, \end{aligned} \tag{11}$$

where P and Λ are respectively the eigenvector and eigenvalue matrices of the $S_w^{-1}S_b$ matrix. These eigenvectors are primarily used for dimensionality reduction, as in principal component analysis (PCA) (Rao, 1948).

However, the performance of the standard LDA can be seriously degraded if there are only a limited number of total training observations n compared to the number of dimensions of the feature space J. In this context, in fact the S_w matrix becomes singular. To solve this problem, Yu and Yang (Thomaz et al., 2006; Yu and Yang, 2001) have developed a direct LDA algorithm (called DLDA) for high dimensional data with application to face recognition that diagonalizes simultaneously the two symmetric matrices S_w and S_b. The idea of DLDA is to discard the null space of S_b by diagonalizing S_b first and then diagonalizing S_w.

The following steps describe the DLDA algorithm for calculating the projection matrix P_{DLDA}:

1. diagonalize S_b, that is, calculate the eigenvector matrix V such that $V^T S_b V = \Lambda$;
2. let Y be a sub-matrix with the first m columns of V corresponding to the S_b largest eigenvalues, where $m \leq rank(S_b)$. Calculate the diagonal $m \times m$ sub-matrix of the eigenvalues of Λ as $D_b = Y^T S_b Y$;
3. let $Z = YD_b^{-1/2}$ be a whitening transformation of S_b that reduces its dimensionality from J to m (where $Z^T S_b Z = I$). Diagonalize $Z^T S_w Z$, that is, compute U and D_w such that $U^T(Z^T S_w Z)U = D_w$;
4. calculate the projection matrix as $P_{DLDA} = D_w^{-1/2} U^T Z^T$.

Note that by replacing the between-class covariance matrix S_b with total covariance matrix S_T ($S_T = S_b + S_w$), the first two steps of the algorithm become exactly the PCA dimensionality reduction technique (Yu and Yang, 2001).

Two other approaches commonly used to avoid both the critical singularity and instability issues of the within-class covariance matrix S_w are SLDA and the MLDA (Thomaz et al., 2006). Firstly, it is important to note that the within-class covariance matrix S_w is essentially the standard pooled covariance matrix S_p multiplied by the scalar $(n-K)$. Then,

$$S_w = \sum_{k=1}^{K} (n_k - 1) S_k = (n-K) S_p. \tag{12}$$

From this property, the key idea of some regularization proposals of LDA (Campbell, 1980; Guo et al., 2006; Peck and Van Ness, 1982) is to replace the pooled covariance matrix S_p of the within-class covariance matrix S_w with the following

convex combination:

$$\hat{S}_p(\gamma) = (1-\gamma)S_p + \gamma\bar{\lambda}I, \tag{13}$$

where $\gamma \in [0,1]$ is the shrinkage parameter, which can be selected to maximize the leave-one-out classification accuracy (Cawley and Talbot, 2003), I is the identity matrix, and $\bar{\lambda} = J^{-1}\sum_{j=1}^{J}\lambda_j$ is the average eigenvalue, which can be written as $J^{-1}trace(S_p)$. This regularization approach, called SLDA, would have the effect of decreasing the larger eigenvalues and increasing the smaller ones, thereby counteracting the biasing inherent in eigenvalue sample-based estimation (Hastie et al., 1995; Thomaz et al., 2006).

In contrast, in the MLDA method a multiple of the identity matrix determined by selecting the largest dispersions regarding the S_p average eigenvalue is used. In particular, if we replace the pooled covariance matrix S_p of the covariance matrix S_w (shown in Equation (12)) with a covariance estimate of the form $\hat{S}_p(\delta) = S_p + \delta I$ (where $\delta \geq 0$ is an identity matrix multiplier), then the eigen-decomposition of a combination of the covariance matrix S_p and the $J \times J$ identity matrix I can be written as

$$\begin{aligned}\hat{S}_p(\delta) &= S_p + \delta I \\ &= \sum_{j=1}^{r}\lambda_j\phi_j(\phi_j)^T + \delta\sum_{j=1}^{J}\phi_j(\phi_j)^T \\ &= \sum_{j=1}^{r}(\lambda_j+\delta)\phi_j(\phi_j)^T + \sum_{j=1}^{J}\delta\phi_j(\phi_j)^T,\end{aligned} \tag{14}$$

where r is the rank of S_p (note that $r \leq J$), λ_j is the j-th eigenvalue of S_p, ϕ_j is the j-th corresponding eigenvector, and δ is the identity matrix multiplier previously defined. In fact, in Equation (14) the identity matrix is defined as $I = \sum_{j=1}^{J}\phi_j(\phi_j)^T$. Now, given the convex combination shown in Equation (13), the eigen-decomposition can be written as

$$\begin{aligned}\hat{S}_p(\gamma) &= (1-\gamma)S_p + \gamma\bar{\lambda}I \\ &= (1-\gamma)\sum_{j=1}^{r}\lambda_j\phi_j(\phi_j)^T + \gamma\sum_{j=1}^{J}\bar{\lambda}\phi_j(\phi_j)^T.\end{aligned} \tag{15}$$

The steps of the MLDA algorithm are shown follows:

1. Find the Φ eigenvectors matrix and Λ eigenvalues matrix ff S_p, where $S_p =$

$(n-K)S_w$ (from Equation (12));
2. Calculate S_p average eigenvalues as $J^{-1}trace(S_p)$;
3. Construct a new matrix of eigenvalues based on the following largest dispersion values :

$$\Lambda^* = diag\left[max(\lambda_1, \bar{\lambda}), \ldots, max(\lambda_J, \bar{\lambda})\right];$$

4. Define the revised within-class covariance matrix:

$$S_w^* = (n-K)S_p^* = (n-K)(\Phi\Lambda^*\Phi^T).$$

Then, the MLDA approach is based on replacing S_w with S_w^* in the Fisherâs criterion formula described in Equation (9).

3. PARTIAL LEAST SQUARES DISCRIMINANT ANALYSIS (PLS-DA)

Multivariate regression methods like principal component regression (PCR) and partial least squares regression (PLS-R) enjoy large popularity in a wide range of fields and are mostly used in situations where there are many, possibly correlated, predictor variables and relatively few samples, a situation that is common, especially in chemistry, where developments in spectroscopy since the seventies have revolutionized chemical analysis (Pérez-Enciso and Tenenhaus, 2003; Wehrens and Mevik, 2007). In fact, the origin of PLSR lies in chemistry (Martens, 2001; Wehrens and Mevik, 2007; Wold, 2001).

In practice, there are not many differences between the use of PCR and PLS-R; in most situations, the methods achieve similar prediction accuracies. Note that with the same number of latent variables, PLS-R will cover more of the variation in Y and PCR will cover more of the variation in X. (Wehrens and Mevik, 2007).

Partial least squares discriminant Analysis (PLS-DA) is a variant of PLS-R that can be used when the response variable Y is categorical. Under certain circumstances, PLS-DA provides the same results as the classical approach of Euclidean distance to centroids (EDC) (Davies and Bouldin, 1979) and under other circumstances, the same as that of linear discriminant analysis (LDA) (Izenman, 2013). However, in different contexts this technique is specially suited to deal with models with many more predictors than observations and with multicollinearity, two of the main problems encountered when analyzing hyperspectral detection data (Pérez-Enciso and Tenenhaus, 2003).

3.1. MODEL AND ALGORITHM

PLS-DA is derived from PLS-R, where the response vector Y assumes discrete values. In the usual multiple linear regression model (MLR) approach we have

$$Y = XB + F, \tag{16}$$

where X is the $n \times J$ data matrix, B is the $J \times 1$ regression coefficients matrix, F is the $n \times 1$ error vector, and Y is the $n \times 1$ response variable vector. In this approach, the least squares solution is given by $B = (X^T X)^{-1} X^T Y$.

In many cases, the problem is the singularity of the $X^T X$ matrix (e.g., when there are multicollinearity problems in the data or the number of predictors is larger than the number of observations). Both PLS-R and PLS-DA solve this problem by decomposing the data matrix X into P orthogonal scores T $(n \times P)$ and loadings matrix P $(J \times P)$, and the response vector Y into P orthogonal scores T $(n \times P)$ and loadings matrix Q $(1 \times P)$. Then, let E and F be the $n \times J$ and $n \times 1$ error matrices associated with the data matrix X and response vector Y, respectively. There are two fundamental equations in the PLS-DA model:

$$\begin{aligned} X &= TP^T + E \\ Y &= TQ^T + F. \end{aligned} \tag{17}$$

Now, if we define a $J \times P$ weights matrix W, we can write the scores matrix as

$$T = XW(P^T W)^{-1}, \tag{18}$$

and by substituting it into the PLS-DA model, we obtain

$$Y = XW(P^T W)^{-1} Q^T + F, \tag{19}$$

where the regression coefficient vector B is given by

$$\hat{B} = W(P^T W)^{-1} Q^T. \tag{20}$$

In this way, an unknown sample value of Y can be predicted by $\hat{Y} = X\hat{B}$, i.e. $\hat{Y} = XW(P^T W)^{-1} Q^T$. The PLS-DA algorithm estimates the matrices W, T, P, and Q through the following steps (Brereton and Lloyd, 2014).

Algorithm 1 Partial Least Squares

1: Fixed P, initialize the residuals matrices $E_0 = X$ and $F_0 = Y$;
2: **for** $p = 1$ to P **do**
3: Calculate PLS weights vector
$W_p = E_0^T F_0$;
4: Calculate and normalize scores vector
$T_p = E_0 W_p (W_p^T E_0^T E_0 W_p)^{-1/2}$;
5: Calculate the X loadings vector
$P_p = E_0^T T_p$;
6: Calculate Y loading
$Q_p = F_0^T T_p$;
7: Update the X residuals vector
$E_0 = E_0 - T_p P_p^T$;
8: Update the Y residuals vector
$F_0 = F_0 - T_p Q_p^T$;
9: **end for**
10: Obtain output matrices W, T, P, Q.

4. APPLICATION TO REAL DATA

In this section we show an application of the method to real data. In particular, we compare the results obtained by partial least squares discriminant analysis (PLS-DA) and the other classification techniques discussed in Section 2.

4.1. DATASET

The dataset consists of 162 drupes of olives harvested in 2010 belonging to three different cultivars (response variable): 54 *Dolce di Andria* (low phenolic concentration), 54 *Moraiolo* (high phenolic concentration), and 54 *Nocellara Etnea* (medium phenolic concentration). Spectral detection is performed using a portable NIR device (diffuse reflectance mode) in the 1100–2300 nm wavelength range, with 2 nm wavelength increments (601 observed variables) (Bellincontro et al., 2012).

4.2. PRINCIPAL RESULTS

In order to evaluate the prediction capability of the model, the entire data set has been randomly divided into a *training set* composed of 111 balanced observations (i.e., about 70% of the entire sample, with each class composed of 37 elements), and a *test set* (drawn from the sample) composed of 51 observations balanced across the three cultivars (i.e., about 30% of the entire sample and each class composed by 17 elements) (Guyon et al., 1998).

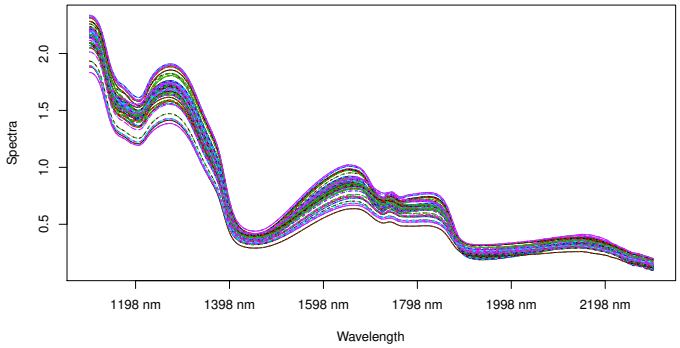

Figure 1: Representation of spectral detections performed on the 1100–2300 nm wavelength range

The first step of the analysis consists in selecting the optimal number of components P, i.e., the number of latent scores to consider for representing the original variable space. For this purpose, the latent subspace must explain the largest possible proportion of the total variance to guarantee the best model estimation. Table 1 shows the proportion of the total variance explained by the first five components identified by PLS-DA.

Table 1: Cumulative proportion of the total variance explained by the first five components (percent values)

	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5
Exp.Variance	61.152	35.589	0.892	0.982	1.167
Cum. Sum	61.152	96.741	97.633	98.615	99.782

The table shows that the first two components explain about 97% of the total variance, and only the first two latent scores have a significant contribution. Thus, it seems that the best latent subspace is represented by the plane composed of the first two identified components. However, in order to guarantee the best model estimate, it is also useful to understand its prediction quality with regard to the different subspace dimensions. In other words, the selection of the optimal number of components must be related to some criterion that ensures the maximum prediction quality of the estimated model. In this paper, we propose the maximization of the chi-squared test applied on the comparison between the real

training partition and the predicted training partition (Rao and Scott, 1981). Figure 2 represents the chi-squared values for different numbers of components (i.e., from 2 to 10 selected components).

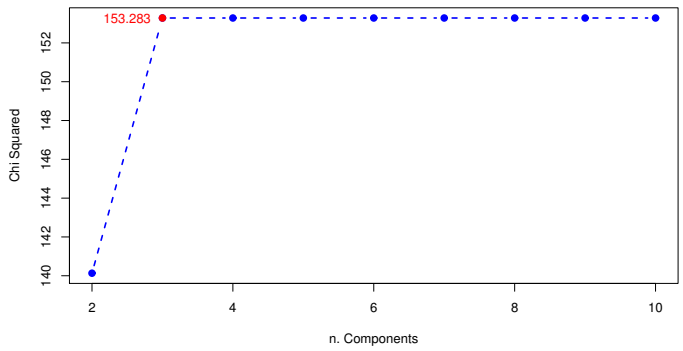

Figure 2: Chi-squared values with respect to different choices of components number

In the scree-plot shown in Figure 2, the chi-squared criterion suggests $P = 3$ as the optimal number of components, where the maximum value of the chi-squared test is equal to 153.28. Then, we can select three components to estimate the model, but we can use the plane composed of the first two latent scores to represent the estimated groups (i.e., using 97% of the total information in the data).

Figure 3 shows the loadings distributions and the squared of the loadings distributions of the three Xs' latent scores, measured on all the observed variables (i.e., on the 1100–2300 nm wavelength range).

By observing the behavior of the loadings, we can say that the wavelengths from about 1100 nm to about 1500 nm have a high negative contribution to the first two components, while they have a positive contribution to the third component; the wavelengths from about 1500 nm to about 1900 nm have a negative contribution to all three components, with the largest contribution to the first component; finally, the wavelengths from about 1900 nm to about 2300 nm have a positive contribution to both the first and the third component, while they have a negative contribution to the second component.

Now, we compare the classification results obtained by the PLS-DA procedure with results obtained by other classifiers, including K-nearest neighbor (KNN), support vector machine (SVM), diagonal linear discriminant analysis

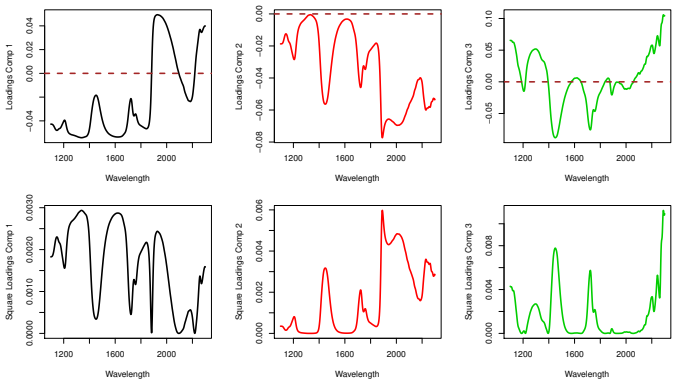

Figure 3: The loadings distributions (top) and squared loadings distributions (bottom) of the three latent scores measured on all the observed variables

(DLDA), maximum uncertainty linear discriminant analysis (MLDA), and shrunken linear discriminant analysis (SLDA). For the measurement of the model prediction quality, we have used *mis classification rate* (MIS), *adjusted Rand Index* (ARI) (Hubert and Arabie, 1985), and the *chi-squared* test (χ^2). The three measures have been computed on the comparison between the real data partition and the predicted partition.

Formally, let Table 2 (here called T) be the $K \times K$ confusion matrix where the real data partition and the predicted partition have been compared, $MIS = 1 - n^{-1}\left[\sum_{r=1}^{R}\sum_{=1}^{C} n_{rc}\right]$, while $ARI = \frac{\sum_{r=1}^{R}\sum_{=1}^{C}\binom{n_{rc}}{2} - \binom{n}{2}^{-1}\sum_{r=1}^{R}\binom{n_{r\cdot}}{2}\sum_{c=1}^{C}\binom{n_{\cdot c}}{2}}{\frac{1}{2}\left[\sum_{r=1}^{R}\binom{n_{r\cdot}}{2}+\sum_{c=1}^{C}\binom{n_{\cdot c}}{2}\right] - \binom{n}{2}^{-1}\sum_{r=1}^{R}\binom{n_{r\cdot}}{2}\sum_{c=1}^{C}\binom{n_{\cdot c}}{2}}$.

Table 2: An example of a confusion matrix between the real data partition and the predicted partition

		Predicted partition			
		P_1	$\cdots$	P_C	
Real partition	R_1	n_{11}	$\cdots$	n_{1C}	$n_{1\cdot}$
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	R_R	n_{R1}	$\cdots$	n_{RC}	$n_{R\cdot}$
		$n_{\cdot 1}$	$\cdots$	$n_{\cdot C}$	n

Table 3 shows the results for the quality of the model predictions obtained on the training set and the test set.

Table 3: Model prediction quality computed on the training set and the test set

	Training set			*Test set*		
	MIS	ARI	χ^2	MIS	ARI	χ^2
PLS-DA	0.002	0.880	153.283	0.008	0.710	77.182
KNN	0.027	0.755	151.744	0.157	0.625	65.294
SVM	0.072	0.797	152.688	0.137	0.615	69.750
DLDA	0.241	0.368	101.599	0.255	0.351	46.714
MLDA	0.078	0.734	149.577	0.010	0.699	72.311
SLDA	0.005	0.712	150.456	0.011	0.702	75.899

From the results, we can see that PLS-DA has the best performance on both the training set and the test set. This result is confirmed by the representation of the predicted partition on the first two Xs' latent scores (i.e., on about 97% of the total data variance) as shown in Figures 4 and 5 (training set and the test set, respectively). In fact, we can see that, with respect to the other studied methodologies, PLS-DA identifies more homogeneous and better-separated classes.

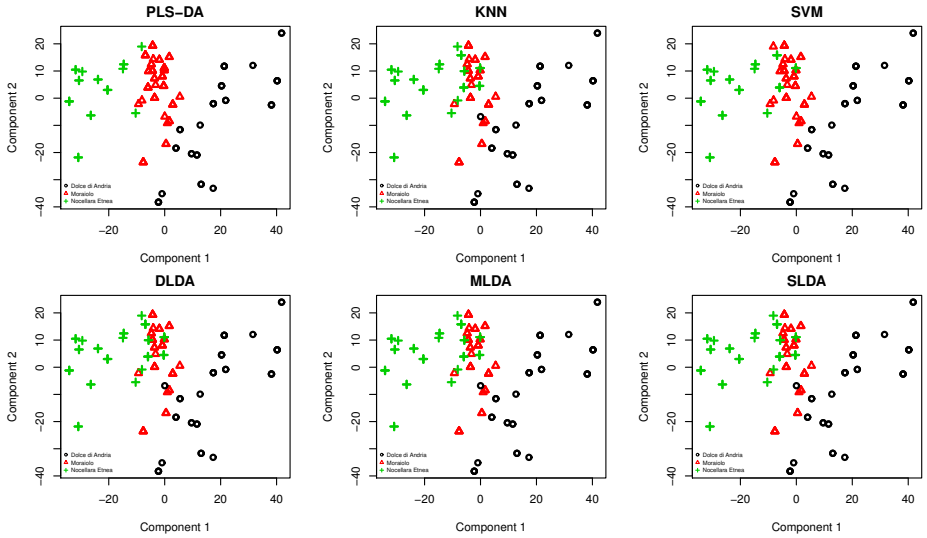

Figure 4: Representation of the predicted partition on the first two latent scores (training set)

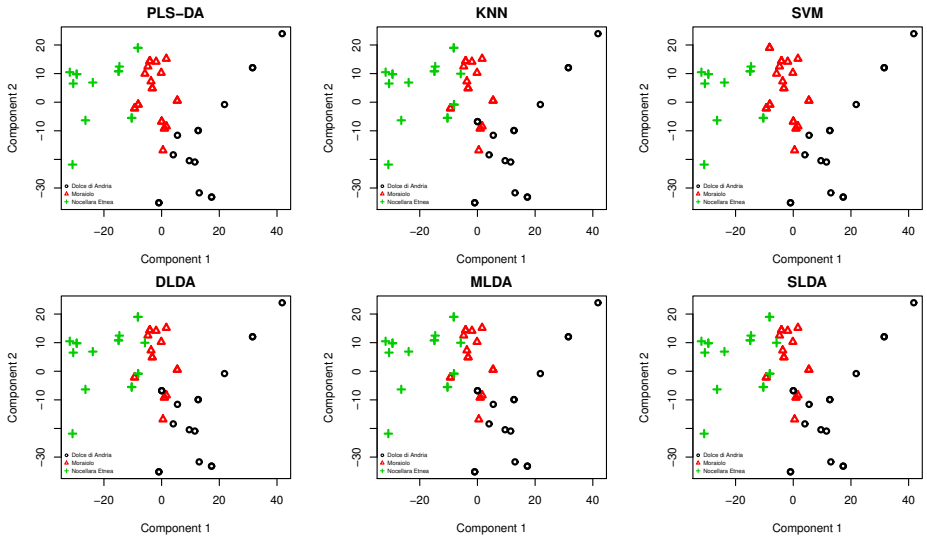

Figure 5: Representation of the predicted partition on the first two latent scores (test set)

5. CONCLUDING REMARKS

Data acquired via spectroscopic detection represent a hard challenge for researchers, who face two crucial problems: data dimensionality larger than the number of observations, and high correlation levels among the variables. In this paper, partial least squares discriminant analysis (PLS-DA) modeling was proposed as a method to classify hyperspectral data. The results obtained on real data show that PLS-DA identifies classes that are more homogeneous and better-separated than other commonly used methods, such as non-parametric classifiers and other discriminant functions.

Moreover, we think that PLS-DA is a very important tool in terms of dimensionality reduction, as it can maximize the total variance of data using just a few components (i.e., the Xs' latent scores). In fact, the PLS-DA components enable a good graphical representation of the partition, which is not possible with other approaches.

In future studies, the use of PLS for unsupervised classification could be a useful tool when both the number and structure of the groups are unknown.

References

Agrawal, R., Gehrke, J., Gunopulos, D., and Raghavan, P. (1998). *Automatic subspace clustering of high dimensional data for data mining applications*, vol. 27. ACM.

Balabin, R.M., Safieva, R.Z., and Lomakina, E.I. (2010). Gasoline classification using near infrared (nir) spectroscopy data: Comparison of multivariate techniques. In *Analytica Chimica Acta*, 671 (1-2): 27–35.

Bellincontro, A., Taticchi, A., Servili, M., Esposto, S., Farinelli, D., and Mencarelli, F. (2012). Feasible application of a portable nir-aotf tool for on-field prediction of phenolic compounds during the ripening of olives for oil production. In *Journal of agricultural and food chemistry*, 60 (10): 2665–2673.

Brereton, R.G. and Lloyd, G.R. (2014). Partial least squares discriminant analysis: taking the magic away. In *Journal of Chemometrics*, 28 (4): 213–225.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Campbell, N.A. (1980). Shrunken estimators in discriminant and canonical variate analysis. In *Applied Statistics*, 5–24.

Cawley, G.C. and Talbot, N.L. (2003). Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers. In *Pattern Recognition*, 36 (11): 2585–2592.

Cayuela, J.A. and Camino, M.d.C.P. (2010). Prediction of quality of intact olives by near infrared spectroscopy. In *European journal of lipid science and technology*, 112 (11): 1209–1217.

Clemmensen, L., Hastie, T., Witten, D., and Ersbøll, B. (2011). Sparse discriminant analysis. In *Technometrics*, 53 (4): 406–413.

Cortes, C. and Vapnik, V. (1995). Machine learning. In *Support vector networks*, 20: 273–297.

Davies, D.L. and Bouldin, D.W. (1979). A cluster separation measure. In *IEEE transactions on pattern analysis and machine intelligence*, (2): 224–227.

Devijver, P.A. and Kittler, J. (1982). *Pattern recognition: A statistical approach*. Prentice hall.

Ding, B. and Gentleman, R. (2005). Classification using generalized partial least squares. In *Journal of Computational and Graphical Statistics*, 14 (2): 280–298.

Dudoit, S., Fridlyand, J., and Speed, T.P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. In *Journal of the American statistical association*, 97 (457): 77–87.

Fisher, T.J. and Sun, X. (2011). Improved stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix. In *Computational Statistics & Data Analysis*, 55 (5): 1909–1918.

Fix, E. and Hodges, J.L. (1989). Discriminatory analysis. nonparametric discrimination: consistency properties. In *International Statistical Review/Revue Internationale de Statistique*, 57 (3): 238–247.

Gallardo, L., Osorio, E., and Sanchez, J. (2005). Application of near infrared spectroscopy (nirs) for the real-time determination of moisture and fat contents in olive pastes and wastes of oil extraction. In *Alimentación Equipos y Tecnologia*, 24 (206): 85–89.

Galvan, V., Gorostiza, O.F., Banwait, S., Ataie, M., Logvinova, A.V., Sitaraman, S., Carlson, E., Sagi, S.A., Chevallier, N., Jin, K., et al. (2006). Reversal of alzheimer's-like pathology and behavior in human app transgenic mice by mutation of asp664. In *Proceedings of the National Academy of Sciences*, 103 (18): 7130–7135.

Garcia, J.M., Seller, S., and Perez-Camino, M.C. (1996). Influence of fruit ripening on olive oil quality. In *Journal of agricultural and food chemistry*, 44 (11): 3516–3520.

Guo, Y., Hastie, T., and Tibshirani, R. (2006). Regularized linear discriminant analysis and its application in microarrays. In *Biostatistics*, 8 (1): 86–100.

Guyon, I., Makhoul, J., Schwartz, R., and Vapnik, V. (1998). What size test set gives good error rate estimates? In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20 (1): 52–64.

Hardy, O.J., Maggia, L., Bandou, E., Breyne, P., Caron, H., CHEVALLIER, M.H., Doligez, A., Dutech, C., Kremer, A., LATOUCHE-HALLÉ, C., et al. (2006). Fine-scale genetic structure and gene dispersal inferences in 10 neotropical tree species. In *Molecular ecology*, 15 (2): 559–571.

Hastie, T., Buja, A., and Tibshirani, R. (1995). Penalized discriminant analysis. In *The Annals of Statistics*, 73–102.

Hubert, L. and Arabie, P. (1985). Comparing partitions. In *Journal of classification*, 2 (1): 193–218.

Izenman, A.J. (2013). Linear discriminant analysis. In *Modern multivariate statistical techniques*, 237–280. Springer.

Jimenez, L.O. and Landgrebe, D.A. (1998). Supervised classification in high-dimensional space: geometrical, statistical, and asymptotical properties of multivariate data. In *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 28 (1): 39–54.

Joachims, T. (2005). A support vector method for multivariate performance measures. In *Proceedings of the 22nd international conference on Machine learning*, 377–384. ACM.

Kemsley, E. (1996). Discriminant analysis of high-dimensional data: a comparison of principal components analysis and partial least squares data reduction methods. In *Chemometrics and intelligent laboratory systems*, 33 (1): 47–61.

Kriegel, H.P., Kröger, P., and Zimek, A. (2009). Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. In *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3 (1): 1.

León, L., Garrido-Varo, A., and Downey, G. (2004). Parent and harvest year effects on near-infrared reflectance spectroscopic analysis of olive (olea europaea l.) fruit traits. In *Journal of agricultural and food chemistry*, 52 (16): 4957–4962.

Martens, H. (2001). Reliable and relevant modelling of real world data: a personal account of the development of pls regression. In *Chemometrics and intelligent laboratory systems*, 58 (2): 85–95.

Misaki, M., Kim, Y., Bandettini, P.A., and Kriegeskorte, N. (2010). Comparison of multivariate classifiers and response normalizations for pattern-information fmri. In *Neuroimage*, 53 (1): 103–118.

Peck, R. and Van Ness, J. (1982). The use of shrinkage estimators in linear discriminant analysis. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (5): 530–537.

Pérez-Enciso, M. and Tenenhaus, M. (2003). Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (pls-da) approach. In *Human genetics*, 112 (5-6): 581–592.

Rao, C.R. (1948). The utilization of multiple measurements in problems of biological classification. In *Journal of the Royal Statistical Society. Series B (Methodological)*, 10 (2): 159–203.

Rao, J.N. and Scott, A.J. (1981). The analysis of categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence in two-way tables. In *Journal of the American statistical association*, 76 (374): 221–230.

Suykens, J.A. and Vandewalle, J. (1999). Least squares support vector machine classifiers. In *Neural processing letters*, 9 (3): 293–300.

Thomaz, C.E., Kitani, E.C., and Gillies, D.F. (2006). A maximum uncertainty lda-based approach for limited sample size problems-with application to face recognition. In *Journal of the Brazilian Computer Society*, 12 (2): 7–18.

Tran, T.N., Wehrens, R., and Buydens, L.M. (2006). Knn-kernel density-based clustering for high-dimensional multivariate data. In *Computational Statistics & Data Analysis*, 51 (2): 513–525.

Wang, Z. and Xue, X. (2014). Multi-class support vector machine. In *Support Vector Machines Applications*, 23–48. Springer.

Wehrens, R. and Mevik, B.H. (2007). The pls package: principal component and partial least squares regression in r. In *Journal of Statistical Software*, 18 (2).

Weston, J. and Watkins, C. (1998). Multi-class support vector machines. *Tech. rep.*, Citeseer.

Wichern, D.W. and Johnson, R.A. (1992). *Applied multivariate statistical analysis*, vol. 4. Prentice Hall New Jersey.

Wold, S. (2001). Personal memories of the early pls development. In *Chemometrics and Intelligent Laboratory Systems*, 58 (2): 83–84.

Yu, H. and Yang, J. (2001). A direct lda algorithm for high-dimensional dataâwith application to face recognition. In *Pattern recognition*, 34 (10): 2067–2070.

Zhang, H., Berg, A.C., Maire, M., and Malik, J. (2006). Svm-knn: Discriminative nearest neighbor classification for visual category recognition. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2, 2126–2136. IEEE.