Research Article

Journal of CHEMOMETRICS

Received: 6 April 2010, Revised: 4 January 2011, Accepted: 3 February 2011, Published online in Wiley Online Library: 29 April 2011

(wileyonlinelibrary.com) DOI: 10.1002/cem.1393

Dissimilarity representation on functional spectral data for classification

Diana Porro-Muñoz[a,b]*, Isneri Talavera[a], Robert P. W. Duin[b], Noslen Hernández[a] and Mauricio Orozco-Alzate[c]

In chemometrics, spectral data are typically represented by vectors of features in spite of the fact that they are usually plotted as functions of e.g. wavelengths and concentrations. In the representation, this functional information is thereby not reflected. Consequently, some characteristics of the data that can be essential for discrimination between samples of different classes or any other analysis are ignored. Examples are the continuity between measured points and the shape of curves. In the Functional Data Analysis (FDA) approach, the functional characteristics of spectra are taken into account by approximating the data by real valued functions, e.g. splines. Another solution is the Dissimilarity Representation (DR), in which classifiers are trained in a space built by dissimilarities with training examples or prototypes of each class. Functional information may be incorporated in the definition of the dissimilarity measure. In this paper we compare the feature-based representation of chemical spectral data with three other representations: FDA, DR defined on raw data and DR defined on FDA descriptions. We analyze the classification results of these four representations for five data sets of different types, by using different classifiers. We demonstrate the importance of reflecting the functional characteristics of chemical spectral data in their representation, and we show when the presented approaches are more suitable. Copyright © 2011 John Wiley & Sons, Ltd.

Keywords: spectral data; Functional Data Analysis; dissimilarity representation; classification

1. INTRODUCTION

The increasing possibilities of chemometrics raise a growing interest in advanced approaches to an automatic analysis of the collected data. If data sets are small, the accuracy of the results is of concern. One way to improve it is by considering advanced data representations. This paper focusses on finding better representations for the classification of spectral data.

The traditional way of representing spectra is by sampling. The higher the sampling resolution, the more accurate the spectrum is described. However, in the case of the design of a classification system for spectra, this implies a representation in a high-dimensional space. For small training sets of spectra, the resulting classifier will thereby be inaccurate due to the curse of dimensionality or overtraining. Dimension reduction by PCA or PLS is needed, but may not solve the problem fully as they are still based on a statistical analysis of high-dimensional data.

Another way to tackle the problem of small training sets is to improve the original representation at the start. In particular, it may be advantageous to directly include the knowledge that spectra are one-dimensional signals and that neighboring points are connected, i.e. their difference in amplitude is limited. The so-called Functional Data Analysis (FDA) [1,2] uses this structural property of spectra by a functional approximation, e.g. by B-spline basis functions. The dimensionality of the description of a spectrum is thereby reduced from the number of samples to the number of functional parameters.

A recently developed alternative in the field of pattern recognition is the Dissimilarity Representation (DR) [3–6]. This representation was mainly designed for discriminating between different classes of objects (classification), based on the important role that dissimilarities play for this purpose. The fact (or property) that dissimilarities should be smaller for similar objects (same class) and larger for different objects suggests that they could be used as more discriminative features due to their crucial performance in the class constitution. Therefore, in this approach objects are represented by distances as new features, determined by some 'appropriate' dissimilarity measure, to a set of prototype objects usually named the representation set. Classifiers may be then built in the dissimilarity space, where each dimension corresponds to the dissimilarity to an object of the representation set (most representative objects for each class), and then applied to a new object represented the same way. Consequently, the geometry and the structure of a class are determined by the user-defined dissimilarity measure, in which application background information may be expressed. Like the FDA, the DR may make use of the structural data characteristics,

* *Correspondence to: D. Porro-Muñoz, Advanced Technologies Application Center (CENATAV), 7a # 21812 e/ 218 y 222, Rpto. Siboney, Playa, C.P. 12200, La Habana, Cuba.*
E-mail: dporro@cenatav.co.cu

a *D. Porro-Muñoz, I. Talavera, N. Hernández*
Advanced Technologies Application Center (CENATAV), 7a # 21812 e/ 218 y 222, Rpto. Siboney, Playa, C.P. 12200, La Habana, Cuba

b *D. Porro-Muñoz, R. P. W. Duin*
Pattern Recognition Laboratory, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, P.O. Box 5031, 2600GA Delft, The Netherlands

c *M. Orozco-Alzate*
Departamento de Informática y Computación, Universidad Nacional de Colombia Sede Manizales, Kilómetro 7 Vía al Aeropuerto, Campus La Nubia – Bloque Q, Piso 2, Manizales, Colombia

e.g. connectivity between the points and shape of the spectra. Some studies have already been reported on the DR for spectral data [7–9]. It is important to remark that any traditional classifier that operates in feature spaces can also be used in the dissimilarity space.

The FDA and the DR are rather young techniques that have received a good acceptance in chemometrics and pattern recognition, respectively. Both aim to solve the problem of a statistical analysis for high-dimensional data generated by sampling spectra by introducing the possibility of integrating structural knowledge in the representation. Some classical multivariate techniques have been extended for FDA, e.g. Functional Principal Component Analysis PCA [1], Canonical Correlation Analysis [10], Partial Least Squares [11–13] and Linear Discriminant Analysis [12]. If the used models are correct they are expected to perform better than the traditional techniques, as these have to learn (linear) relations from the data. More recently, a number of estimation methods for functional nonparametric classification and regression models have also been introduced. namely k-Nearest Neighbor classifier [14], kernel methods [15–17] such as Support Vector Machine [18,19–21].

DR can be considered as a generalization of the kernel approach studied in machine learning [3,4] as it accepts almost any (dis)similarity measure between objects (spectra), including indefinite ones. Several applications have been studied [5], including spectra [7,9]. Also in chemometrics studies appeared in which similarities between spectra are applied [22–24], but these mainly aim at studying correlation, cluster analysis or visualization. The use of representation has almost not been studied at all.

In this paper, we will compare the DR directly defined on distances between spectra with the FDA approach as well as with their combination, the DR based on the functional description of spectra. By this proposed combination, advantages of both approaches may be combined. By approximating the spectral data with the B-spline basis functions, the structural information in the spectra, e.g. shape, connectivity between measured points, can be incorporated in the dissimilarity measure. According to the basis of the DR, when adding this information about the structure of the objects, the distances between them should be more discriminative features. Therefore, one of the main issues in chemometrics, the small number of objects in high-dimensional spaces can be tackled, as with less but more discriminative data should be enough for the classification task. Hence, nonlinearly separable problems in the feature space can be converted to linear problems in the dissimilarity space.

As a baseline procedure for the comparison, the traditional feature representation in which spectra are represented by their samples is used. The following classifiers are used: the k-Nearest Neighbor (k-NN) rule [25,26], Regularized Linear Discriminant Analysis (RLDA) [26], Soft Independent Modelling of Class Analogy (SIMCA) [27] and the Support Vector Machine (SVM) [28,29]. Section 2 summarizes and defines the foundations of FDA, DR and their combination. Data sets and experimental procedures are presented in Section 3 and results are discussed in Section 4. Finally, our conclusions are presented.

2. THEORY

2.1. Dissimilarity representation

The Dissimilarity Representation (DR) [3,5,6] was originally proposed as a more flexible representation of the objects than the traditional feature-based one. In this approach, which was mainly thought for classification purposes, new features are defined for the objects, such that they are represented by their dissimilarities to a set of representative objects of each class. The fact (or property) that dissimilarities should be smaller for similar objects (same class) and larger for different objects suggests that they could be used as more discriminative features due to their crucial role in the class constitution.

It aims at including more information about the characteristics and structure of the objects through the dissimilarity measure, e.g. shape of spectra. There is no general dissimilarity measure for all problems. Hence, the first task in the DR is to select a suitable dissimilarity measure for the problem at hand. The fact that it has to be user-specified is a way for the expert to integrate his knowledge and application [6].

Thus, in this approach, given a set of training objects $\mathbf{X} = x_1, x_2, \ldots x_n$, e.g. spectra, a representation set (a set of prototypes or representative objects for each class) $\mathbf{R}(r_1, \ldots, r_p)$, e.g. the reference spectral sample for each substance that constitutes a class, and a dissimilarity measure, the distance between each object $x_i \in \mathbf{X}$ to each object $r_h \in \mathbf{R}$ will be defined as $\mathrm{d}(x_i, r_h)$. The representation set $\mathbf{R}$ can be a subset of $\mathbf{X}$, $\mathbf{R} \subseteq \mathbf{X}$ or $\mathbf{X}$ itself, being then $\mathbf{D}(\mathbf{X},\mathbf{X})$ a square dissimilarity matrix, or $\mathbf{R}$ and $\mathbf{X}$ can be completely different sets. There are some approaches to select prototypes of the representation set [30]. See reference for further details.

An object from the training set is then represented by a vector of dissimilarities $\mathbf{D}(x_i, \mathbf{R}) = \left[\mathrm{d}(x_i, r_1), \ldots, \mathrm{d}(x_i, r_p)\right]$, which relates it to the prototypes in the representation set. Therefore, in place of the traditional feature matrix $\mathbf{X} \in \mathbb{R}^{n \times q}$, where n runs over the objects and m over the variables, the training set is now represented by the dissimilarity matrix $\mathbf{D}(\mathbf{X}, \mathbf{R})$ of size $n \times p$, which associates all objects from the training set with all objects from the representation set:

$$\mathbf{D} = \begin{pmatrix} \mathrm{d}(x_1, r_1) & \mathrm{d}(x_1, r_2) & \ldots & \mathrm{d}(x_1, r_p) \\ \mathrm{d}(x_2, r_1) & \mathrm{d}(x_2, r_2) & \ldots & \mathrm{d}(x_2, r_p) \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{d}(x_n, r_1) & \mathrm{d}(x_n, r_2) & \ldots & \mathrm{d}(x_n, r_p) \end{pmatrix}$$

We build from this matrix a dissimilarity space $\mathbb{D} \subseteq \mathbb{R}^p$. Objects are represented in this space by the row vectors of the dissimilarity matrix, such that each dimension corresponds to the dissimilarities with one of the representation objects. Using the DR, classifiers are trained in the space of the dissimilarities between objects, instead of the traditional feature space. Consequently, the relationship between all objects in the training and representation sets is used for the classification. If a suitable measure is chosen, the compactness property (objects from the same class should be similar and objects from different classes should be different) of the classes should be more pronounced. Therefore, it should be easier for the classifiers to discriminate between them, such that linear classifiers in dissimilarity space may correspond to nonlinear classifier in feature space. In general, any arbitrary classifier operating on features can be used [4].

Given a test set $\mathbf{Y} = y_1, y_2, \ldots y_g$, these objects are classified in the dissimilarity space, using their distances to the prototypes in $\mathbf{R}$, $\mathbf{D}(\mathbf{Y}, \mathbf{R})$, which is a $g \times p$ matrix.

Some dissimilarity measures have been already proposed for spectral data [7,8]. In recent studies, the shape (ShD) measure was demonstrated to have the capabilitiy for capturing the functional

information. It consists of a sum of the absolute differences between the first Gaussian derivatives of the curves

$$d(x_1, x_2) = \sum_{j=1}^{m} \left| x_{1j}^{\sigma} - x_{2j}^{\sigma} \right|, \; x^{\sigma} = \frac{d}{d_j} G(j, \sigma) * x \qquad (1)$$

The expression of x^{σ} corresponds to the computation of the first Gaussian (that is what G stands for) derivatives of spectra. Thus, a smoothing (blurring) is done by a convolution process ($*$) with a Gaussian filter and σ stands for the smoothing parameter [7]. Good performances have been obtained for chemical spectral data with this measure [7,31].

2.2. Functional Data Analysis

In chemical spectral data as near-infrared, ultraviolet, each spectrum is a function of e.g., wavelengths, concentrations. However, they are usually observed and recorded discretely and so analyzed with multivariate data analysis techniques which consider the spectrum as high-dimensional vectors of different but highly-correlated variables. Therefore, when working with this type of representation, many practical problems can be encountered as the characteristics of the functional nature of the data are not taken into account.

FDA is based on retrieving the intrinsic characteristics of the underlying function from the discrete functional data. Thus, the observations (spectra) can be seen as continuous single entities, instead of sets of different variables. Nevertheless, if the algorithms work on the functional spaces, they can also lead to theoretical and practical difficulties as these have infinite dimensions.

For dealing with the infinite-dimensional problem, FDA methods have been constructed on two general principles: regularization and filtering. The filtering approach is based on using representation methods that allow working in finite dimension. This way of approximation is used here. The first step in FDA is to choose a proper family of basis functions matching best the function(s) to be approximated. Of a variety of bases that exist (Fourier series, polynomial, wavelet and splines), as spectra are generally smooth, it seems that B-splines [32] are more appropriate to approximate them. To make this basis of B-splines $\{\phi_k\}_{k=1}^{K}$ with K the number of basis functions, a number of knots (points) between the start and end wavelengths are defined. A B-spline is run from one knot to another; the different splines can overlap.

Hence, the spectral function $x_i = x_i(\lambda)$ for sample i and wavelengths λ can be described by the linear combination of the basis functions

$$x_i(\lambda) = \sum_{k=1}^{K} c_{ik} \phi_{ik}(\lambda)$$

where $\mathbf{c_i} = [c_{i1}, c_{i2}, \ldots, c_{ik}]$ is the vector of B-spline weights (coefficients) corresponding to each spectrum (object) x_i. These coefficients are computed by minimizing the distance between the observed discrete spectrum x_i at wavelengths λ_j; $\forall j = 1, 2, \ldots, m$ and the fitted curve $x_i\lambda$

$$\mathbf{c_i} = \arg_c \in \mathbb{R}^k \min \sum_{j=1}^{m} \left(x_{ij} - \sum_{k=1}^{K} c_{ik} \phi_{ik}(\lambda_j) \right)^2$$

Filtering can therefore be considered as a preprocessing step in which functional data are consistently transformed into vector data [33]. As we are operating now in a finite-dimensional space, it is possible to work with the coefficients instead of working on the approximating functions. It has been demonstrated that working with these coefficients vectors $\mathbf{c_i}$ is strictly equivalent to working directly on the ϕ_i functions [34].

The function for each spectrum is thus explained by K coefficients, which are represented in a vector $\mathbf{c_i}$, obtaining for the entire data set a matrix $\mathbf{C}(\mathbf{n} \times \mathbf{K})$:

$$\mathbf{C} = \begin{pmatrix} c_{11} & c_{12} & \ldots & c_{1k} \\ c_{21} & c_{22} & \ldots & c_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ c_{n1} & c_{n2} & \ldots & c_{nk} \end{pmatrix}$$

Matrix $\mathbf{C}$ will be taken as the new representation of the data set, and the classifiers or any other data analysis method can use it as input, instead of the original data points. A dimensionality reduction is achieved in this process, which is an important task when working with spectral data. The functional representation by B-splines is already a way of smoothing the curve; other functional processing techniques such as derivation can be done on it. This processing could be beneficial when the analysis of the shape of the function (curvature) is essential for the solution of the problem at hand.

2.3. Dissimilarity representation and Functional Data Analysis

Based on the advantages that both of the previous approaches show for the representation of spectral data, we propose to compute the DR from the functional representation of the data (DR-FDA), instead of computing it from the feature-based one. By using the functional approximation, we are highlighting the structural (in terms of how the composition of the substances is reflected in the shape of the spectra) information of the spectra. Hence, we have a more faithful representation of the real spectrum than by using the feature-based one. Moreover, by using B-spline basis, the interpretability of the data is still maintained. Because the new variables (coefficients) depend only on some spectral regions, a range of these original variables can be associated with each new variable. Therefore, the spectral regions responsible for the discrimination can still be depicted from the functional representation. A method to achieve this association was recently introduced [33].

Moreover, if spectra are represented by dissimilarities, the use of a suitable dissimilarity measure allows emphasizing important details of the particular problem, which cannot be easily treated by the simple feature representation, e.g. shape and continuity of the measured points of the spectrum. Any knowledge on the problem, e.g. discriminative spectral regions, or on the spectral background can also be included in the measure. Besides, by using the dissimilarities as features, we are considering the relationship between all objects (structure of the classes) as information. This is very important for the discrimination between the classes, and even more when the number of samples is very small (a typical problem in chemometrics), as less but more discriminative data should be enough for the classification task. Thus, nonlinearly separable problems in the feature space can be converted to linear problems in the dissimilarity space. Furthermore, by using the DR for chemical spectral data, the problem of high-dimensionality spaces is also eradicated. The highest dimension that could have the data now is the number of objects in the training set, which are usually much less than variables (spectral bands) in these types of data sets. Therefore, it can be less computationally expensive.

In consequence, it is to be expected that if the dissimilarities between a set of spectra $\mathbf{D}(\mathbf{X}, \mathbf{R})$ are derived from their functional representation, i.e. the vector of coefficients $\mathbf{c}$ obtained from the approximation of each spectrum by B-spline basis functions; the classification results may improve. A better description of the objects than feature-based will be used, where patterns in the structure of the spectra can be more exploited. Example of a simple dissimilarity measure that can be computed on the functional data is the Manhattan (L1-norm) distance:

$$\mathrm{d}(x_1, x_2) = \sum_{k=1}^{K} |c_{1k} - c_{2k}| \tag{2}$$

The Manhattan distance is one of the most commonly used in many research areas, and particulary for the comparison of spectral data [8,9,23,31]. This measure views the spectrum as a high-dimensional feature vector, making just a band-to-band comparison, therefore neglecting the connectivity between the measured points of the spectra. However, the functional information from the spectra can be taken into account if the distance is computed from their functional representation by B-splines approximation.

In the following section, the feature, functional and DR approaches will be compared to the proposed one on different chemical spectral data sets. Four classifiers are used to show the efficacy of this approach compared to the others.

3. MATERIALS AND METHODS

A comparative study is carried out between the two representations presented above (DR and FDA) and the proposed approach (DR-FDA), using the feature representation as a baseline comparison. The performance of four classifiers will be evaluated on these four representations of five chemical spectral data sets.

3.1. Data sets

The first data set, named Tecator (Figure 1), originates from the food industry [35]. It consists of 215 near infrared absorbance spectra of meat samples, recorded on a Tecator Infratec Food and Feed Analyzer. Each observation consists in a 100 channel absorbance spectrum in the 850–1050 nm wavelength range. It is associated with a content description of meat sample, obtained by analytic chemistry.

The classification problem consists in separating 77 meat samples with a high fat content (more than 20%), from 138 samples with a low fat content (less than 20%). Original spectra are preprocessed; each spectrum is reduced to zero mean and unit variance.

The second data set (Figure 1) is composed of near infrared (NIR) transmittance spectra of pharmaceutical tablets [36]. It consists of 310 spectra and 404 variables in a range of wavelengths from 7400 to 10 500 cm^{-1}. Four different (classes) dosages of nominal content of active substance are analyzed: class A (5 mg), B (10 mg), C (15 mg) and D (20 mg) per tablet. There are 70 objects in class A and 80 in each of the other classes. As reported in Reference [36], a Multiplicative Scatter Correction (MSC) [37] was used as a preprocessing method. The MSC transforms the spectrum x to z, such that $z(j) = (x(j) - a)/b$, with a the intercept and b the slope of a least-squares regression of the values $x(j)$, on the corresponding values $r(j)$ for a reference spectrum. Usually this reference spectrum is the mean of all the available spectra [38].

The third data set is a real-world data set, which was obtained from a cooperation with the Oil Industry in Cuba. It consists of 80 fuel samples of Fourier Transform Infrared (FT-IR) transmittance spectra (Figure 2) in a wavelength range of 600–4000 cm^{-1}. A base line correction and smoothing were performed on the data. The classification problem consists in determining the fuel type of the samples: regular gasoline (16 samples), especial gasoline (15 samples), regular diesel (16 samples), naphtha (16 samples), turbo diesel (9 samples) and kerosene (8 samples).

The fourth data set is another fuel real-world data set of 101 samples measured at 127 wavelengths in a range of 275–220 nm, but this time measures have been taken by an Ultra-Violet Visible (UV) spectrophotometer (see Figure 2). The classification problem consists also in determining the fuel type of the samples: regular gasoline (23 samples), especial gasoline (21 samples), regular diesel (22 samples), naphtha (18 samples) and turbo diesel (17 samples).

The last data set consists of 101 NIR spectra of four different common pharmaceutical excipients (classes), with 27, 14, 17 and 13 samples in each class, respectively, measured at 700 wavelengths (see Figure 3). The goal is to develop a classification model to identify to which of the pharmaceutical excipients belongs to a new sample. This is an example data set from the chemometrics software Pirouette [39]. For both of the previous data sets, original spectra are preprocessed such that each spectrum is reduced to zero mean and unit variance.

3.2. Software and optimization

The experiments were all performed in Matlab. For the case of FDA the FDAFuns [40] toolbox was used, and PRTools toolbox [41]

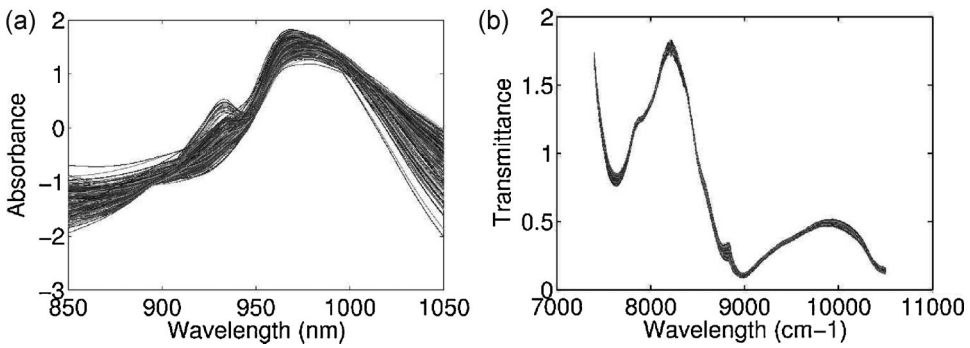

Figure 1. Data sets (a) Tecator and (b) Tablets.

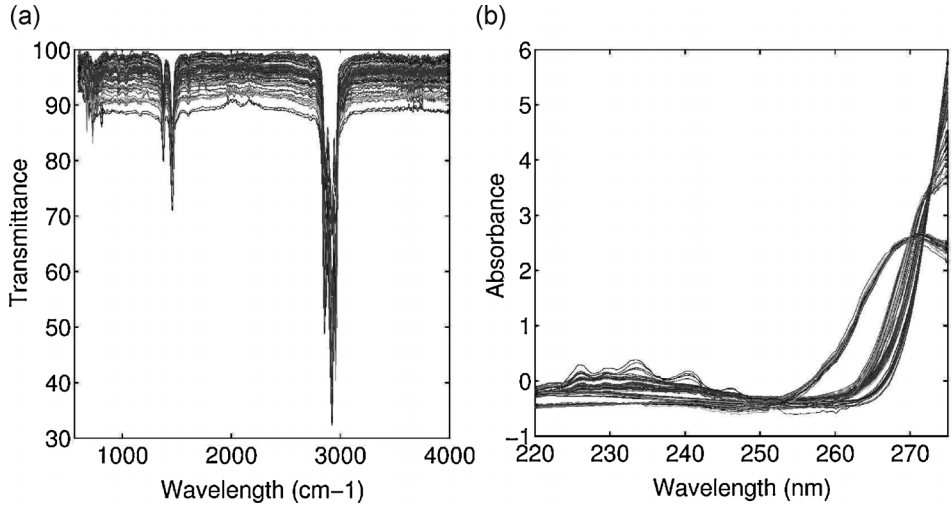

Figure 2. Fuel data set from (a) FT-IR and (b) UV-VIS.

for the DR and classification of the data. The experiments have been designed in the following way:

(1) A comparison is made between the performance of classifiers on the four representations of each dataset: feature-based (spectral), functional (FDA), DR and the proposed combination (DR-FDA).
(2) For the classification of the data we used four classifiers: *k*-Nearest Neighbor (*k*-NN), Support Vector Machine (SVM), Soft Independent Modeling of Class Analogy (SIMCA) and Regularized Linear Discriminant Classifier (RLDC). To solve the multiclass problems of some of the data sets, the one-versus-all classification scheme is applied. For the *k*-NN classifier, a leave-one-out optimization for k is computed. An optimal number of $k = 1$ was obtained for all data sets; thus, a 1-NN classifier is applied. For the SVM classifier, the Gaussian and linear kernel were applied in the five data sets, to show their performances on their different representations. The optimal regularization parameter C and Gaussian kernel width parameter σ were tuned in a grid search based on a *k*-fold cross-validation procedure. In the case of the Linear kernel, the regularization parameter was optimized in a *k*-fold cross-validation. The Linear Discriminant Classifier (LDC) assumes that the classes are described by multi-normal distributions with the same/different covariance matrices. Since for $n \times n$ DRs the estimated covariance matrix **S** is singular, its inverse cannot be determined. Therefore, its regularized version is used instead (RLDC). Regularization takes care that the inverse operation is possible by emphasizing the diagonal values (variances) of the matrix **S** with reference to the off-diagonal elements (covariances) [3,4]. To find the regularization parameters of RLDC, an automatic regularization (optimization over training set by cross-validation) process was done.
(3) For the functional representation, each spectra was represented by a *l*th order B-spline approximation with K basis functions. The optimal values for the number of B-spline bases and the order of the splines were chosen by a leave-one-out cross-validation, using the error in the approximation of the curve as evaluation criteria [34]. In the comparison with all representations, the results for the FDA are reported for the performance of classifiers on the functional representation, and when the second derivative is applied on it.
(4) For the DR, the shape distance was applied (Equation 1). For DR-FDA we used the Manhattan distance on the functional representation as defined in Equation (2). The results shown in this case are those computed on the functional representation version (see above) for which the classifiers performed better. The entire set of samples was used as a representation set for all data.
(5) For all data sets, a *k*-folds cross-validation procedure was repeated 10 times, such that all objects are used for training and test at some moment. Consequently, the information of all samples is taken into account for the modeling of the problem. Classifiers' performances are evaluated in terms of the Average Classification Error (ACE), and the standard deviation from the different repetitions is taken into account.

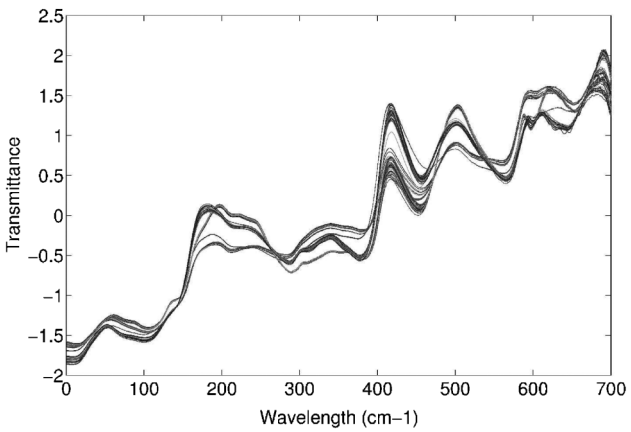

Figure 3. Pharmaceutical excipients data set.

4. RESULTS AND DISCUSSION

4.1. Tecator data set

The classification results (averaged classification error) for the different representations of Tecator data set are shown in Table I.

Table I. Averaged cross-validation error in % (with standard deviation) for Tecator dataset for different classifiers

Classifiers	Representations				
	Feature	DR	FDA	FDA (+ 2nd der)	DR-FDA
1-NN	4.11 (0.2)	3.72 (0.2)	2.97 (0.2)	2.15 (0.1)	1.41 (0.1)
RLDA	6.82 (0.5)	1.72 (0.01)	3.02 (0.3)	2.74 (0.3)	0.98 (0.1)
SVM (Gaussian kernel)	2.56 (0.2)	2.65 (0.2)	1.21 (0.3)	0.93 (0.2)	0.6 (0.04)
SVM (Linear kernel)	3.51 (0.2)	2.88 (0.1)	1.8 (0.2)	1.21 (0.05)	0.47 (0)
SIMCA	5.77 (0.01)	3.3 (0.2)	2.6 (0.07)	2.31 (0.3)	1.21 (0.04)

The data set was split into different training and test sets in a 10-fold cross-validation repeated 10 times. For the functional approach, the leave-one-out error calculation leads to the selection of an optimal basis of 48 B-splines of order 5.

As stated above, in these data, the samples of the two classes differ in their fat content, which is reflected in changes in the shape of the spectra. We can observe remarkable differences in curvature of the spectra between the samples of the two classes (fat < 20 and fat > 20). High fat content spectra (fat > 20) have sometimes two local maxima instead of one (fat < 20) (Figure 4). As it is reported in the literature [19], we computed the second derivative of the functional data to highlight these differences.

In the case of the derivative-based distance, shape (ShD), the second derivative was also applied. The smoothing parameter σ was optimized in a 10-fold cross-validation procedure repeated 10 times. The best results were achieved with $\sigma = 2$.

In Table I, it is shown that classifiers perform better on the functional data than on their original (feature) representation. The good performance of most classifiers on the functional space is due to the fact that, from the functional point of view, there is a great amount of information to obtain when shape changes are present in the curve. Hence, the FDA by B-splines is capable of using the information embedded in the curvature of the spectrum. The use of the second derivatives emphasizes the peaks in the curve, therefore making it easier to see the differences [19]. Nevertheless, it seems that it is not enough to discriminate between both classes.

For the DR, the results with the shape dissimilarity measure are very good compared to the results on the original feature-based data. This dissimilarity measure takes into account the shape information (functional) that can be obtained from the derivatives. Compared to FDA, the results are usually worst (taking the standard deviation into account). This can be due to, in cases like this, the use of the B-splines and second derivatives afterwards are more capable of extracting the functional information than the shape dissimilarity measure.

In this data set, we can see that linear classifiers perform a bit better on the functional space, but not good enough. Results with nonlinear classifiers are still better. This might be because classes are nonlinearly separable. In Figure 5, the scores of all samples from a PCA of two principal components from each representation are plotted. In all cases more than 95% is retained in these two principal components. These are not demonstrative plots, but they show in someway how the structure of the classes can be according to each representation. It can also be seen that all classifiers, even the linear classifiers, perform better when computing the dissimilarities on the functional representation of the data (see Equation 2), corroborating our hypothesis. These results are also better than that obtained in the literature [19]. The functional information that is not captured by the measure itself is obtained by the FDA, and thus included in the DR. The relationship between all objects is also considered when analyzing them all-against-all in the dissimilarity space, which is very important for the discrimination between the classes. These results could be even improved by some other expert knowledge introduced into the dissimilarity measure.

4.2. Fuel data set by FT-IR and UV-VIS

With these two data sets, we are tackling the same problem of discriminating between types of fuel, but using two different instrumental techniques. Besides, they are not based on the same

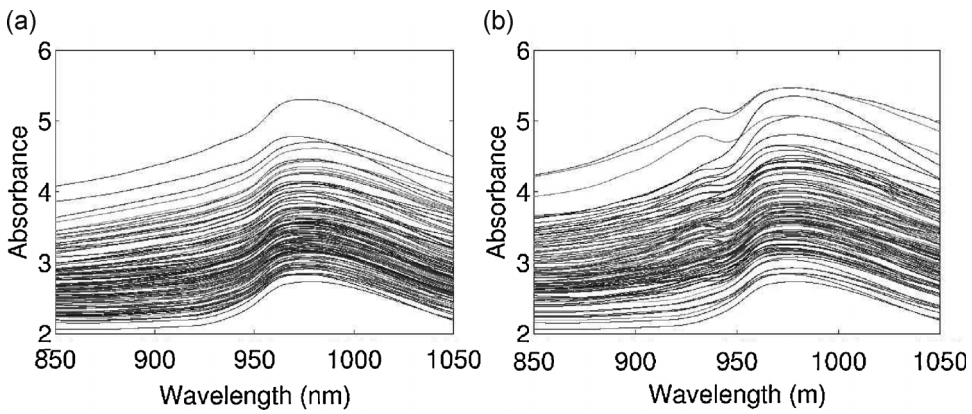

Figure 4. Data sets (a) Tecator (fat < 20) and (b) Tecator (fat > 20).

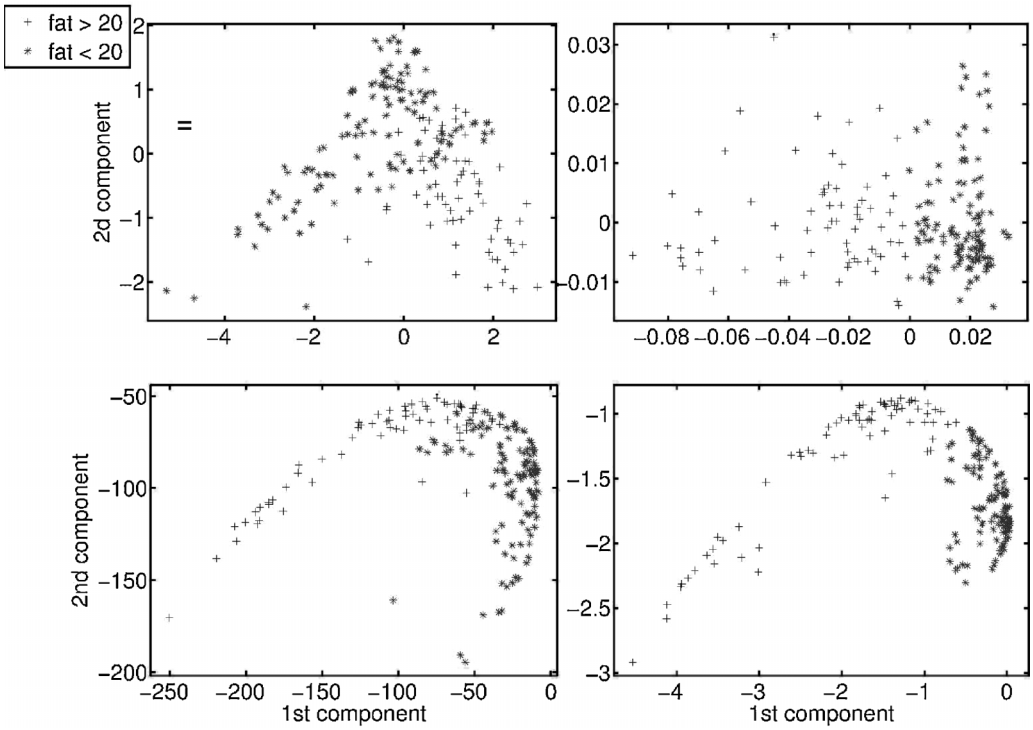

Figure 5. PCA of two components for the four representations of Tecator data set: feature (top-left), FDA (top-right), DR (bottom-left) and DR-FDA (bottom-right).

samples, and in the FT-IR data set, one more class is analyzed. In this case, the samples of the classes differ in the substances by which they are composed (see Figure 6), and therefore they differ in shape (although sometimes it is difficult to determine for all of them) in some parts of their spectrum.

The data sets were split into different training and test sets in a 8-fold and 10-fold cross-validation for the FT-IR and UV-VIS, respectively; these splits were repeated 10 times. For the functional approach, the leave-one-out error calculation leads to the selection of an optimal basis of 850 B-splines of order 4 for the FT-IR data set and 30 B-splines of order 4 for the UV-VIS. We also computed for both of them the second derivative of the functional data, to highlight the curvature differences. In the case of the derivative-based distance, shape, the second derivative was applied. The smoothing parameter σ was optimized in a 5-times 10-fold cross-validation procedure and the best results were achieved with $\sigma = 3$ also for both cases.

It can be observed from Tables II and III that the pattern in the behavior of the results is very similar to that obtained for Tecator data set. For both of them we are in the same situation. The difference between the samples of the classes is mainly in the curvature of the spectra.

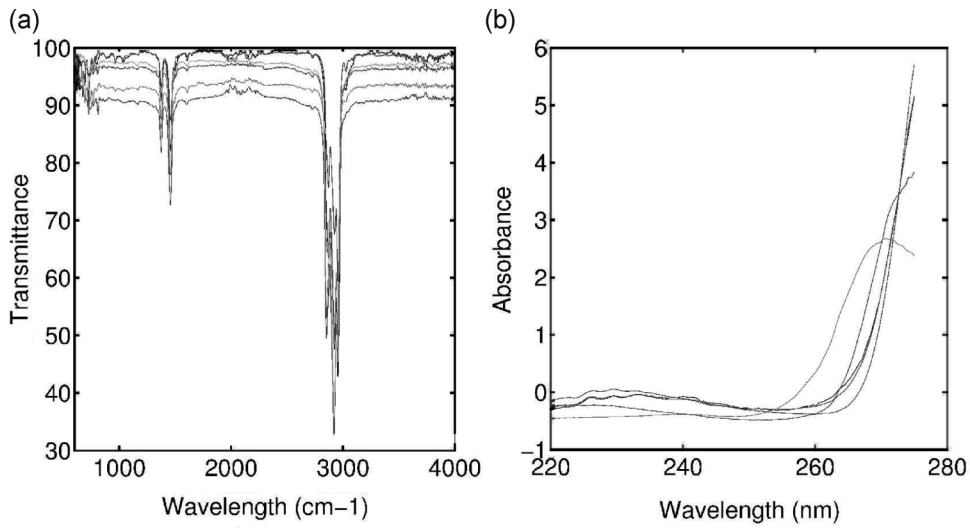

Figure 6. One sample of each of the classes of Fuel (a) FT-IR and (b) UV-VIS data sets.

Table II. Averaged cross-validation error in % (with standard deviation) for Fuel (FT-IR) data set for different classifiers

Classifiers	Representations				
	Feature	DR	FDA	FDA (+ 2nd der)	DR-FDA
1-NN	32.4 (0.8)	14.3 (0.7)	18.1 (0.8)	10.4 (0.7)	7.5 (0.8)
RLDA	16.1 (0.5)	13.6 (0.2)	14 (0.5)	11.8 (0.6)	9.5 (0.4)
SVM (Gaussian kernel)	11 (0.5)	8.8 (0.3)	12.8 (0.8)	9.6 (0.5)	6.3 (0.3)
SVM (Linear kernel)	17.8 (0.8)	7.4 (0.4)	14.9 (0.5)	12 (0.8)	4.5 (0.2)
SIMCA	22.5 (1.1)	15.6 (0.6)	14.9 (1.1)	10.7 (1)	8.7 (1)

Table III. Averaged cross-validation error in % (with standard deviation) for Fuel (UV-VIS) data set for different classifiers

Classifiers	Representations				
	Feature	DR	FDA	FDA (+ 2nd der)	DR-FDA
1-NN	13.1 (0.3)	19.8 (0.4)	14.6 (0.4)	11.4 (0.2)	10.7 (0.3)
RLDA	21.4 (1)	13.3 (0.8)	14.9 (0.7)	10.8 (0.4)	7 (0.6)
SVM (Gaussian kernel)	13.1 (0.2)	11.4 (0.3)	16.8 (0.1)	9.4 (0.2)	8.1 (0.2)
SVM (Linear kernel)	15 (0.5)	12.2 (0.5)	14.1 (0.6)	12.5 (0.1)	8.7 (0.2)
SIMCA	21.2 (0.6)	19.5 (0.4)	20.6 (0.4)	17.5 (0.7)	14.3 (0.3)

It can be noticed that in both of these data sets the results with the functional data and the DR outperform those obtained for the feature-based representation of the data. It is also remarkable how the classifiers' accuracy improves even more when the DR is computed on the second derivative of the functional representation of the data. However, the results are a bit worst for the UV-VIS spectra due to the characteristics of the instrumental techniques. It seems that the information obtained in the FT-IR spectra is more discriminative specially for regular and especial gasoline. It is worth noticing the advantage of the DR in this case, where we have six (FT-IR) and five (UV-VIS) classes and a few samples for each of them. If the dissimilarities have captured more structure from the data, it should be sufficient to discriminate with few data. The FT-IR data set is also the case, where in the functional representation a dimensionality reduction was achieved, but still it was high for the number of samples available. From this approximation more reduction could not be possible, otherwise important information could be lost in the smoothing process; besides, it was the result from the optimization process. Thus, the advantage of the DR in these cases is highlighted as classifiers are built in a more balanced space and do not have to deal with the high-dimensionality problems.

4.3. Pharmaceutical excipient data set

A sample of each of the four pharmaceutical excipient classes is shown in Figure 7. This is another example where the objects from different classes have differences in shape in some parts of their spectra. Thus, it would be important to take into account this information in their representation, such that it would be possible to discriminate better between them.

The data set was split into different training and test sets in a 10-fold cross-validation procedure. For the functional approach, the leave-one-out error calculation leads to the selection of an optimal basis of 150 B-splines of order 4. We also computed for both of them the second derivative of the functional data to highlight the curvature. In the case of the derivative-based distance, shape, the second derivative was applied. The optimal value for the smoothing parameter was $\sigma = 1$.

As it is shown in Table IV for all classifiers, the combination of DR with FDA outperforms the other representations in general. In this case, although there is an improvement, it is not as remarkable as in the previous data sets. The classification problem seems to be not so difficult as it can be observed in the

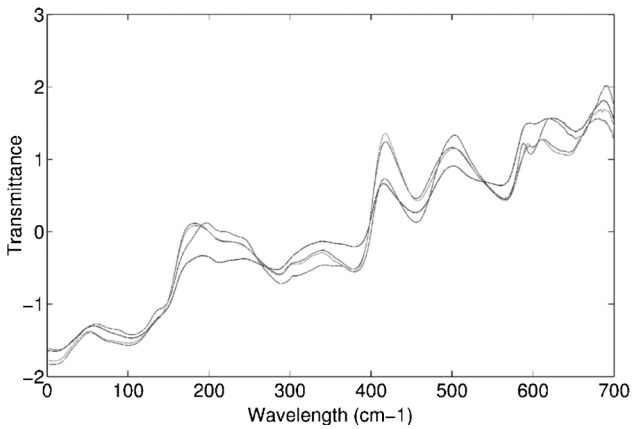

Figure 7. One sample of each of the classes of pharmaceutical excipient data set.

Table IV. Averaged cross-validation error in % (with standard deviation) for pharmaceutical excipient data set for different classifiers

Classifiers	Representations				
	Feature	DR	FDA	FDA (+ 2nd der)	DR-FDA
1-NN	3.9 (0.05)	2.7 (0.3)	3.2 (0.3)	2.1 (0.5)	1.4 (0.1)
RLDA	2.9 (0.6)	1.7 (0.4)	2.5 (0.3)	1.4 (0.1)	1 (0.2)
SVM (Gaussian kernel)	2.8 (0.05)	1.8 (0.5)	2.1 (0.3)	1.5 (0.1)	1.21 (0.2)
SVM (Linear kernel)	3.1 (0.1)	2.1 (0.8)	2.8 (0.2)	1.9 (0.2)	1 (0.3)
SIMCA	4.2 (0.6)	3.7 (0.1)	4.7 (0.04)	3.8 (0.4)	2.7 (0.7)

results. Thus, it depends on the application how much the users are willing to give up in some efficiency for a bit more of efficacy.

From the studies of all these data sets, it can be seen that when the spectra of different classes are characterized by having differences in their curvature, this is discriminative information that should be taken into account. These differences between the classes are not always so visible or clear, such that the pattern for each of them could be extracted easily. Thus, the use of mathematical operators like the second derivative can emphasize the curvature of the spectrum; therefore, the shape difference is more highlighted to be further used by the classifiers.

In the results shown for the previous data sets, the SVM shows the better results on all the representations. This could be due to the fact that these data sets are mostly nonlinear. This classifier has shown a high flexibility to confront complex data sets. Such is the case of spectral chemical data sets where the number of samples is small and is high dimensionality and the classes are completely unbalanced with respect to the number of samples that belong to each of them. Nevertheless, it should be noticed that in most cases, when classifiers are built on the DR from FDA, the linear classifiers, i.e. RLDA and SVM (linear kernel), have better or the same performance as the nonlinear ones, showing again the feasibility or one of the advantages of this approach. Nonlinearly separable problems in the feature space can be converted to linear problems in the dissimilarity space. All the previously discussed corroborates the idea that the proposed combination can be optimal in cases like this, where the DR is generated from the functional data extracted by the approximation with B-splines.

4.4. Tablet data set

In these data, the spectra of the samples of the different classes are very similar; they barely variate in the intensity of one peak at 8830 cm^{-1} (corresponding to 1132 nm) (Figure 8). This peak corresponds to the only visually characteristic band of the active substance, which is identified as the second overtone of the aromatic C–H stretch. It is partially overlapping with the peak at 8200 cm^{-1}(1220 nm), originating from the primary excipient, microcrystalline-cellulose [36]. It seems that we are in the presence of another nonlinearly separable classes problem.

The data set was split into different training and test sets in a 10-fold cross-validation procedure. For the functional approach, the leave-one-out error calculation leads to the selection of an optimal basis of 150 B-splines of order 4. In the case of the derivative-based distance, shape, the first derivative was applied. The smoothing parameter σ was optimized in a 5-times 10-fold cross-validation and the best results were achieved with $\sigma = 2$.

The results are shown in Table V. In this data set, it is to be expected that the results with the functional representation from the B-splines approximation do not make much of a difference. This could be explained by the fact that only a small amount of information can be extracted from these data, from the functional point of view. The second derivative does not give any information of curvature either.

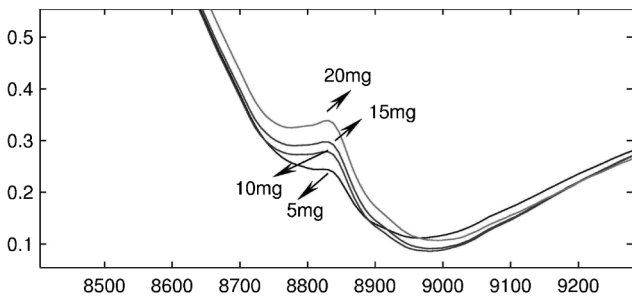

Figure 8. Nominal content of active substance (mg) for the classes.

Table V. Averaged cross-validation error in % (with standard deviation) for Tablet data set for different classifiers

Classifiers	Representations				
	Feature	DR	FDA	FDA (+ 2nd der)	DR-FDA
1-NN	16.7 (0.1)	8.7 (0.2)	15.1 (0.1)	22.9 (0.5)	14.9 (0.5)
RLDA	25.6 (0.6)	7.7 (0.4)	20.8 (0.3)	24.8 (0.4)	10.87 (0.2)
SVM (Gaussian kernel)	14.1(0.3)	6.7 (0.3)	10.7 (0.2)	14.5 (0.3)	9.6 (0.3)
SVM (Linear kernel)	20.1 (0.5)	10.5 (0.2)	15.6 (0.3)	17.8 (0.4)	11.2 (0.2)
SIMCA	23.8 (0)	14.8 (0.4)	24.2 (0.1)	28.6 (0.05)	16.2 (0.1)

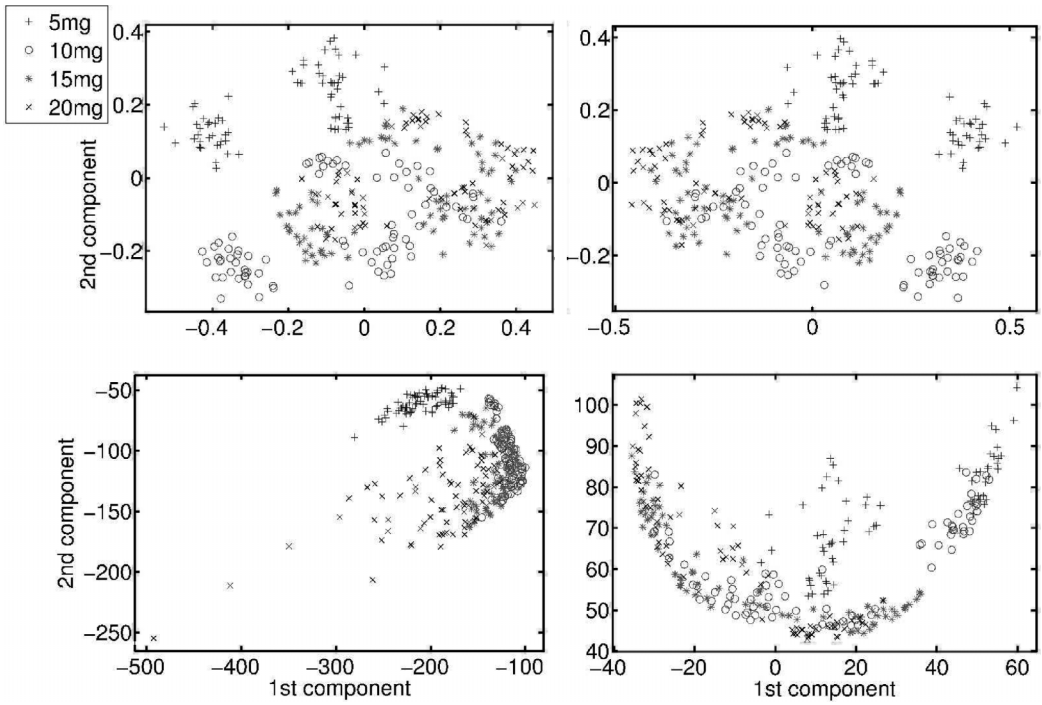

Figure 9. PCA of two components for the four representations of Tablet data set: feature (top-left), FDA (top-right), ShD (bottom-left) and DR-FDA (bottom-right).

Indeed, it can be observed from the above table that the results with the functional representation are very similar to that obtained on the original spectra (considering the standard deviation). With the DR, good results are obtained in general, even more than for FDA based on B-splines. This shows that this measure is also capable of detecting the intensity changes between the different curves, even when they are so slight as in this case. In Figure 9 we again try to show with two-components PCA (more than 95% of variance is retained for all representations) that the classes could be slightly more separable in this space than in the others.

However, if we analyze the performance of classifiers on the dissimilarity space obtained from the functional representation, the behavior is reasonable. Most classifiers perform better on the dissimilarity space generated from the original feature-based data than that of the functional representation. It is understandable that in this case, the scarce functional information extracted by the B-splines does not benefit the DR This lack of information in the functional representation can be caused by the loss of some information when using only the coefficients resulting from the smoothing process with projection of the function in the B-spline basis.

5. CONCLUSIONS

We presented three alternative ways to improve the representation of chemical spectral data. The first makes use of the physical knowledge of the spectral background of the data, by modeling their relations in a DR The second makes use of the spectral connectivity by approximating the spectra by spline functions (FDA). In the third, we propose to compute the DR on the functional representation of the data (DR-FDA). Therefore, the functional information of spectra is taken into account and we can make use of the advantages of both approaches. Comparisons were made by classifying five chemical spectral data sets, expressed by their feature and the three other representations. We can conclude that: in chemical spectral data sets where changes in the shape of the spectra of different classes are present, e.g. Tecator and Fuel data sets, both FDA and the DR outperform the results on the feature space. In the comparison of these types of data by their dissimilarities, better results are obtained with measures that take the functional information into account. Such is the case of the shape dissimilarity measure and the proposed combination of the DR on the functional data. Nevertheless, the latter is shown to be the best option in this case. In data sets where the differences between the samples are referred only to intensity changes, e.g. Tablet, the shape dissimilarity is capable of improving the results obtained on the feature space. However, FDA (with B-splines) is not able to extract the functional information from these types of data. Therefore, the computation of the DR on the functional data does not improve, since it is influenced by the errors of the functional approach.

Acknowledgements

The authors acknowledge financial support from the FET program within the EU FP7, under the SIMBAD project (contract 213250). They also thank the project Cálculo científico para catacterización e identificación en problemas dinámicos (code Hermes 10722) granted by Universidad Nacional de Colombia.

REFERENCES

1. Silverman B. Smoothed functional principal components analysis by choice of norm. *Ann. Stat.* 1996; **24**(1): 1–24.
2. Ramsay JO, Silverman BW. *Functional Data Analysis*. (2nd edn.) Springer Series in Statistics. Springer-Verlag: New York, USA, 1997.
3. Pekalska E, Duin RPW. *The Dissimilarity Representation For Pattern Recognition. Foundations and Applications*. World Scientific: Toh Tuck Link, Singapore, 2005.
4. Pekalska E, Duin RPW. Classifiers for dissimilarity-based pattern recognition. In: *International Conference on Pattern Recognition*: Barcelona, Spain, 2000; 12–16.
5. Pekalska E, Duin RPW. Dissimilarity representations allow for building good classifiers. *Pattern Recognit. Lett.* 2002; **23**(8): 943–956.
6. Duin RPW, Pekalska E, Paclik P, Tax DMJ. The dissimilarity representation, a basis for domain based pattern recognition? *Representations in Pattern Recognition, invited talk (refereed) IAPR Workshop*, Cambridge 2004; 43–56.
7. Paclik P, Duin RPW. Dissimilarity-based classification of spectra: computational issues. *Real Time Imaging* 2003; **9**(4): 237–244.
8. Paclik P, Duin RPW. Classifying spectral data using relational representation. *Spectral Imaging Workshop*, Graz, Austria, 2003.
9. Orozco-Alzate M, García ME, Duin RPW, Castellanos CG. Dissimilarity-based classification of seismic signals at Nevado del Ruiz Volcano. *Earth Sci. Res. J.* 2006; **10**(2): 57–65.
10. Leurgans SE, Moyeed RA, Silverman B. Canonical correlation analysis when the data are curves. *J. Roy. Stat. Soc. Ser. B* 1993; **55**: 725–740.
11. Aguilera AM, Ocaña FA, Valderrama MJ. An approximated Principal Component Prediction model for continuous time stochastic processes. *Appl. Stoch. Model. Data Anal.* 1997; **13**: 61–72.
12. Cardot H, Ferraty F, Sarda P. Functional linear model. *Stat. Probab. Lett.* 1999; **45**(1): 11–22.
13. Preda C, Saporta G. PLS regression on stochastic processes. *Comput. Stat. Data Anal.* 2005; **48**(1): 149–158.
14. Cérou F, Guyader A. Nearest neighbor classification in infinite dimension. *ESAIM: P&S* 2006; **10**: 340–355.
15. Abraham C, Biau G, Cadre B. On the kernel rule for function classification. *Ann. Inst. Stat. Math.* 2006; **58**(3): 619–633.
16. Biau G, Bunea F, Wegkamp MH. Functional classification in hilbert spaces. *IEEE Trans. Inform. Theory* 2005; **51**(6): 2163–2172.
17. Ferraty F, Vieu P. *Nonparametric Functional Data Analysis: Theory and Practice (Springer Series in Statistics)*. Springer-Verlag: New York Inc., 2006.
18. Preda C. Regression models for functional data by reproducing kernel Hilbert spaces methods. *J. Stat. Plan. Infer.* 2007; **137**(3): 829–840.
19. Rossi F, Villa N. Support vector machine for functional data classification. *Neurocomputing* 2006; **69**(7–9): 730–742.
20. Hernández N, Biscay RJ, Talavera I. Support vector regression methods for functional data. In *CIARP*, (vol. 4756), LNCS, Springer: Viña del Mar, Chile, 2007; 564–573.
21. Hernández N, Talavera I, Biscay RJ, Porro D, Ferreira MC. Support vector regression for functional data in multivariate calibration problems. *Anal. Chim. Acta* 2009; **642**(1–2): 110–116.
22. Xu Y, Gong F, Dixon S, Brereton R, Soini H, Novotny M, Oberzaucher E, Grammar K, Penn D. Application of dissimilarity indices, principal coordinates analysis and rank tests to peak tables in metabolomics of the gas chromatography mass spectrometry of human sweat. *Anal. Chem.* 2007; **79**(15): 5633–5641.
23. Komsta L, Skibinski R, Grech-Baran M, Galaszkiewicz A. Multivariate comparison of drugs UV spectra by hierarchical cluster analysis-comparison of different dissimilarity functions. In: *Annales Universitaits Marie Curie-Sklodowska*, vol. 20: Medical University of Lublin, Lublin, Polonia, 2007; 97–105.
24. Varmuza K, Karlovits M, Demuth W. Spectral similarity versus structural similarity: infrared spectroscopy. *Anal. Chim. Acta* 2003; **490**(1–2): 313–324.
25. Duda RO, Hart PE, Stork DG. *Pattern Classification*. Wiley: New York, 2001.
26. Fukunaga K. Introduction to Statistical Pattern Recognition (2nd edn.) *Computer Science and Scientific Computing*. Academic Press Professional, Inc., San Diego, USA, 1990.
27. Wold S, Sjostrom M. SIMCA: A method for analyzing chemical data in terms of similarity and analogy. *Chemometrics Theory and Application* 1977; **52**: 243–282.
28. Schölkopf B. Support Vector Learning. *PhD thesis*: Munich, Germany, 1997.
29. Vapnik V. *Statistical Learning Theory*. John Wiley & Sons, Inc.: New York, USA, 1998.
30. Pekalska E, Duin RPW. Prototype selection for finding efficient representations of dissimilarity data. In: *International Conference on Pattern Recognition*, (vol. 3), Quebec, Canada, 2002; 37–40.
31. Porro-Muñoz D, Talavera I, Duin RPW, Hernández N. The representation of chemical spectral data for classification. In: *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Proceedings of the 14th Iberoamerican Congress on Pattern Recognition CIARP 2009*, vol. 5856, LNCS, Springer: Guadalajara, Mexico, 2009; 513–520.
32. Wahba G. *Spline Models for Observational Data*. SIAM [Society for Industrial and Applied Mathematics]: Philadelphia, USA, 1990.
33. Rossi F, Francois D, Wertz V, Meurens M, Verleysen M. Fast selection of spectral variables with b-spline compression. *Chemom. Intell. Lab. Syst.* 2007; **86**(2): 208–218.
34. Rossi F, Delannayc N, Conan-Gueza B, Verleysen M. Representation of functional data in neural networks. *Neurocomputing* 2005; **64**(2): 183–210.
35. Thodberg HH. Tecator dataset. Danish Meat Research Institute. http:lib.stat.cmu.edu/datasets/tecator/ 1995.
36. Dyrby M, Engelsen S, Nørgaard L, Bruhn M, Lundsberg Nielsen L. Chemometric quantitation of the active substance in a pharmaceutical tablet using Near Infrared (nir) Transmittance and nir ft Raman spectra. *Appl. Spectrosc.* 2002; **56**(5): 579–585.
37. Geladi P, MacDougall D, Martens H. Linearization and scatter-correction for near-infrared reflectance spectra of meat. *Appl. Spectrosc.* 1985; **39**(3): 491–500.
38. Fearn T, Riccioli C, Garrido-Varo A, Guerrero-Ginel JE. On the geometry of SNV and MSC. *Chemom. Intell. Lab. Syst.* 2009; **96**: 22–26.
39. Pirouette software. Infometrix. http://www.infometrix.com.
40. Ramsay JO. Fdafuns toolbox.ftp://ego.psych.mcgill.ca/pub/ramsay/FDAfuns/ 2005.
41. Duin RPW, Juszczak P, de Ridder D, Paclik P, Pekalska E, Tax DMJ. PRTools: a Matlab toolbox for pattern recognition. http://www.prtools.org/download.html/ 2004.

