Shape Descriptors for Classification of Functional Data

Irene Epifanio

Departament de Matemàtiques

Universitat Jaume I

12071-Castelló, Spain.

E-mail: epifanio@uji.es.

February 8, 2008

Abstract

Curve discrimination is an important task in engineering and other sciences. We propose several shape descriptors for classifying functional data, inspired by form analysis from the image analysis field: statistical moments, coefficients of the components of independent component analysis (ICA) and two mathematical morphology descriptors (morphological covariance and spatial size distributions). They are applied to three problems: an artificial problem, a speech recognition problem and a biomechanical application. Shape descriptors are compared with other methods in the literature, obtaining better or similar performances.

Keywords: Curve discrimination; statistical moment; independent component analysis; spatial size distribution; morphological covariance

1 Introduction

Technological advances have permitted the acquisition of functional data or curves. Functional data analysis (FDA) provides statistical procedures for functional observations.

Ramsay and Silverman (2005) give an excellent overview. Ferraty and Vieu (2006) provide a complementary and very interesting view on nonparametric methods for functional data. The field of FDA is quite new and there is still a lot of work to be done, but several applications in different fields have been developed in recent years (Ramsay and Silverman, 2002). Furthermore, the software that the authors used is available on the website for these books.

In this paper, we focus on the curve discrimination problem, that is the objects to classify are functional observations. We have K different types of curve and the aim is to classify a new function as one of the K types. A biomechanical application has partially motivated this paper: the functional assessment of lumbago. We analyze the ground force recordings in the sit-to-stand (STS) movement for three groups: patients, controls and pretenders. The objective is to classify a new observation into one of these three classes.

Curve discrimination arises in many contexts and is an important problem. A clear example is signal discrimination, which has been considered in several papers involving, for instance, the use of high-resolution radar returns for target detection (Hall et al., 2001) or the recognition of speech signals (Hastie et al., 1995; Glendinning and Fleet, 2004). Other interesting applications include medical diagnosis from EEG measurements from multiple scalp sites (Anderson et al., 1998), the automatic classification of rivet defects using eddy currents (Lingvall and Stepinski, 2000), gait analysis (Huang, 2001) and chemometric applications such as the prediction of the fat content of a meat sample based on the near-infrared absorbance spectrum (Ferraty and Vieu, 2003) or a polymer discrimination problem (Naya et al., 2004).

When we have many highly correlated predictors such as those obtained by discretizing a function, classical tools such as Fisher's linear discriminant analysis (LDA) can fail. Different solutions have been proposed. For example, Hastie et al. (1995) overcame this problem by a regularized version of LDA called penalized discriminant analysis (PDA), Marx and Eilers (1999) proposed a generalized linear regression approach and Ferraty and Vieu (2003)

proposed a nonparametric tool, calculating the posterior probability of belonging to a given class of functions by using a consistent kernel estimator. When only fragments of the curves are observed, the technique introduced by James and Hastie (2001) can be used.

Recent advances in functional data classification appear in the following papers. These papers are closely related to the method proposed as they all involve a type of preprocessing (sometimes implicit) of the functional data. Rossi and Conan-Guez (2005) and Conan-Guez and Rossi (2004) proposed the use of neural networks for nonlinear regression and classification of functional data (see also Rossi et al. (2005)); James and Silverman (2005) studied a non linear regression model for functional descriptors (see also James (2002) for some simpler and earlier extensions of the linear functional model by the same author); Biau et al. (2005) studied k-nearest neighbor classifiers for functional data and provided very interesting theoretical results (see also the ongoing work by Berlinet et al. (2005) and related theoretical work by Fromont and Tuleau (2006)); Rossi and Villa (2006) used Support Vector Machines (SVM) for functional data classification; Ferré and Villa (2005) studied a preprocessing approach in which functional data are described via a projection on an optimal basis (in the sense of the inverse regression approach), and subsequently submitted to a neural network for further processing (Ferré and Villa, 2006). Finally, a different approach is introduced by López-Pintado and Romo (2006), where robust procedures based on the concept of depth to classify curves were proposed.

One of the main objectives of the statistics of shapes and forms is to study how they can be described for the purposes of classification or clustering (Stoyan and Stoyan, 1994). According to Stoyan and Stoyan (1994), shapes can be described by three kinds of tools: firstly, set descriptors and mathematical morphology tools; secondly, using landmarks or specific features; and thirdly, employing a function describing the contours. In form analysis, it is usual to analyze binary images and these figures can be well characterized by their contours. Typical boundary descriptors in image analysis are Fourier descriptors and statistical moments (González et al., 2004). Furthermore, as univariate functional data can be

seen as a one-dimensional (1-D) image, we can use 1-D mathematical morphology tools for describing and discriminating according to its shape. As landmark identification is usually done manually, we will not consider that tool in this paper.

In short, we will employ three kinds of shape (or form) descriptors for discriminating: 1) Statistical moments; 2) Coefficients of an appropriate basis function system; and 3) Shape descriptors based on mathematical morphology.

The paper is organized as follows. Shape descriptors are introduced in Section 2, where the methodology is also explained. In Section 3 the proposed methods are applied to three problems: an artificial problem, a speech recognition problem and a biomechanical application. Shape descriptors are compared with some standard methods and more recent techniques. These examples illustrate the good behavior of our descriptors, which provide better or similar performances. Finally, some conclusions are given in Section 4.

2 Methodology: shape descriptors and classification procedures

Let x_1, ..., x_n be a collection of n curves defined on the interval $[a, b]$. It is always assumed that functions satisfy reasonable smoothness conditions and are square integrable functions on this interval. However, in practice the curves are not observed continuously but in a finite set of points.

In supervised curve classification, given a learning sample containing observed curves with known class memberships $\{1, ..., K\}$, our problem is to predict the membership group of new curves. Curves are described by means of different shape features: a feature vector. We calculate these features for curves in the training and test sets. A suitable classifier is then used to assign curves in the test set to a class. If the original membership of each test sample is known, the quality of the classifier can be assessed in terms of the percentage of correct (or incorrect) classifications.

In short, a shape feature vector is associated with each curve (feature extraction stage) and this finite-dimensional vector is employed in the classification stage. We transform each observed curve into an appropriate vector of characteristics that describes our data better. This kind of preprocessing is a powerful method for improving the performance of a learning algorithm, instead of using the raw features (see section on Feature Extraction in Hastie et al. (2001, pp. 126-127)). For the classification stage, we use classical classifiers (Ripley, 1996): quadratic or linear discriminant analysis, depending on which provides better results in each case. Although more sophisticated classifiers could be used in the classification stage, we have chosen those classical classifiers since we are mainly interested in the feature extraction stage. Note that if ideal discriminant features are extracted (each class is represented by a region of the feature space which is well separated from the regions representative of other classes), the task of the classifier should be trivial.

The methods employed to extract the feature vectors describing the shape of the curves are presented in the following sections. Statistical moments are introduced in Section 2.1. Section 2.2 considers the coefficients of an appropriate basis function system, while two mathematical morphology descriptors (morphological covariance and spatial size distributions) are presented in Section 2.3. More details about the appropriate selection of the parameters for each method will be given in Section 3.

2.1 Statistical moments

If our functions were density functions (they could be made positive and normalized to unit area), moments would be directly related to the shape of the functions. It is well-known that the first moment indicates the location, the second moment measures the spread of the curve about the mean, the third moment measures its symmetry with reference to the mean and the fourth moment measures the kurtosis. In image analysis, statistical moments have been widely (and successfully) used for boundary descriptions (González et al., 2004).

Our descriptors are based on this idea, but we do not normalize our functions. If we

had normalized the functions, we would have lost information about the function values, and therefore about the size. With normalization, the scale information is not preserved and only the shape is considered.

Let x be a curve defined on the interval $[0, 1]$. If defined on $[a, b]$, this can be easily affine-transformed to the unit interval. In practice, the curve is observed in m equi-spaced points $\{x(t_k); k = 1, ..., m\}$, with $t_1 = 1/m$ and $t_m = 1$. Although x is not a density curve, we can extrapolate moment definitions on this curve and, abusing the language, we also call them moments.

The non-central moment of order r of x can be computed as:

$$m_r = \sum_{k=1}^{m} t_k^r \cdot x(t_k). \tag{1}$$

Note that as the curve is not normalized, the moment of order zero can be useful. Furthermore, the unit interval has been considered because otherwise the values could increase greatly when exponentiated, as the curve is not normalized. Also note that Eq. 1 can be seen as an approximation of $\int_0^1 t^r x(t)dt$. In this paper, the truncated monomial basis t^r is used because of its simplicity, interpretability and the good results obtained in pattern recognition. Another basis with other properties can be chosen. For example, Legendre moments are well established orthogonal moments. See Shutler (2002) for a good compendium on statistical moments.

The curve can be described by a vector of its moments, which is used in the classification stage. The number of moments used, that is to say the length of the feature vector, is selected by Monte Carlo cross-validation (MC CV) (Burman, 1989), i.e. repeated random splits of the data set into a training and a validation set. The data are partitioned 50 times into disjoint train and validation subsets, where the validation subset is 62.5% or 50% of the overall data, depending on the specific example. The key distinction between MC CV and v-fold cross-validation is that in MC CV the different validation subsets are chosen randomly and need not be disjoint. We compute the results for a different number

of moments and we choose the length for which the best results are obtained. Afterwards, performance is assessed by another application of MC CV on newly generated data.

2.2 Coefficients of independent component analysis (ICA) components

In shape analysis, it is quite common to select the Fourier basis and use the basis coefficients as shape descriptors since there are good numerical methods and programs for its calculation and a non-mathematical interpretation of some coefficients is frequently possible (Stoyan and Stoyan, 1994).

The Fourier basis is especially useful for an extremely stable function, with no strong local features and the same curvature order everywhere (Ramsay and Silverman, 2005), which could make them inappropriate for our functional data. We consider a more sophisticated basis system, adapted to our data: ICA. A good reference on ICA is Hyvärinen et al. (2001). Let us briefly recall the definition of ICA.

Assume that we observe n linear mixtures $x_1(t)$, ..., $x_n(t)$ of n independent components $s_j(t)$,

$$x_i(t) = \sum_{j=1}^{n} a_{ij} s_j(t), \text{ for all } i. \tag{2}$$

In practice, we have discretized curves $(\{x_i(t_k); k = 1, ..., m\})$, therefore we can consider the $m \times n$ data matrix $\mathbf{X} = \{x_i(t_k)\}$ to be a linear combination of independent components, i.e. $\mathbf{X} = \mathbf{SA}$, where columns of $\mathbf{S}$ contain the independent components and $\mathbf{A}$ is a linear mixing matrix. ICA attempts to un-mix the data by estimating an un-mixing matrix $\mathbf{W}$ where $\mathbf{XW} = \mathbf{S}$.

Under this generative model the measured signals in $\mathbf{X}$ will tend to be more Gaussian than the source components (in $\mathbf{S}$) due to the Central Limit Theorem. Thus, in order to extract the independent components/sources, we search for an un-mixing matrix $\mathbf{W}$ that maximizes the nongaussianity of the sources.

We could use principal component analysis (PCA) as in Hall et al. (2001) and Ferraty and

Vieu (2003). However, ICA has recently been shown to be a potentially powerful method for analyzing signals (Back and Weigend, 1997; Hastie et al., 2001; Hyvärinen et al., 2001). In comparison with PCA, ICA allows the underlying structure of the data to be more readily observed. PCA is a purely second-order statistical method, whereas ICA requires the use of higher-order statistics. Therefore, ICA can be seen as an extension to PCA.

Furthermore, ICA can be considered a variant of projection pursuit (Hastie et al., 2001). Projection pursuit is a technique for finding interesting projections of multidimensional data. ICA has also been successfully used for feature extraction and image analysis (Hyvärinen et al., 2001). In fact, the basis functions obtained when ICA is applied to images are spatially localized and selective for orientation and spatial frequency (scale) and are therefore similar to basis functions of multiscale wavelet representations (Olshausen and Field, 1996; Bell and Sejnowski, 1997).

We compute ICA for functions in the training set. The coefficients in this base (S) for functions in the test set can be easily obtained by least squares fitting (Ramsay and Silverman, 2005). If $\mathbf{y} = \{y(t_k)\}_{k=1}^m$ is a discretized function in the test set, its coefficients are: $(\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T\mathbf{y}$. These coefficients constitute the feature vector used in the classification stage.

Note that we assume that all functions are observed at the same points, which is a normal situation. In any case, this is not a restrictive issue since we can always fit a basis and estimate the functions at the desired points.

When applying ICA to real data, before the application of the ICA algorithm, it is useful to preprocess the data (Hyvärinen et al., 2001). When dealing with signals, as in our case, filtering (smoothing) of the data is useful for reducing noise. Another very useful thing to do is to previously reduce the dimension of the data by PCA, thus reducing noise and preventing overlearning. Therefore we compute the PCA first, retaining a certain number of components, and then estimate the same number of independent components as the PCA reduced dimension.

The number of independent components used is also selected by MC CV.

2.3 Mathematical morphology descriptors

Mathematical morphology is a theory for the analysis of spatial structures which aims at analyzing the shape and form of objects. Many powerful techniques have been developed with this in mind and a good reference in this field is Soille (2003). Although image analysis is the main field of application for this theory, it can also be applied to 1-D continuous functions or, in practice, 1-D discretized functions. For instance, morphological tools were previously applied to radar signals (Rivest, 2006).

In this paper we use two kinds of morphological descriptors, useful for extracting shape and size characteristics, which were originally successfully employed in texture classification. These descriptors are morphological covariance (Soille, 2003) and spatial size distributions (Ayala and Domingo, 2001). Before recalling their definitions, we introduce some concepts.

2.3.1 Preliminaries

Let B be a subset of the 1-D Euclidean space $\mathbb{R}$, called the structuring element. Let $\check{B}$ be the reflection of B with respect to the origin ($\check{B} = \{-b : b \in B\}$), and let f be a 1-D function.

The erosion of f by B is $[\varepsilon_B(f)](t) = \inf_{b \in B} f(t+b)$, the dilation of f by B is defined by $[\delta_B(f)](t) = \sup_{b \in B} f(t+b)$ and finally, the opening of f by B is $\gamma_B(f) = \delta_{\check{B}}[\varepsilon_B(f)]$.

These definitions are illustrated in Fig. 1, where a discrete version is considered, since we usually observe discretized curves. If the curve is viewed as a topographic relief, the value of the curve at a certain point would be its elevation. Erosion will remove any 'hills' or 'peaks' of the original curve which are smaller than the structuring element. However, it will also cause 'valley' widening which will result in sharper 'peaks'. Dilation will remove 'valleys' that cannot contain the structuring element and 'peaks' will be widened. Finally, opening will cut 'peaks' that cannot contain the structuring element. See Soille (2003) for

more illustrative examples.

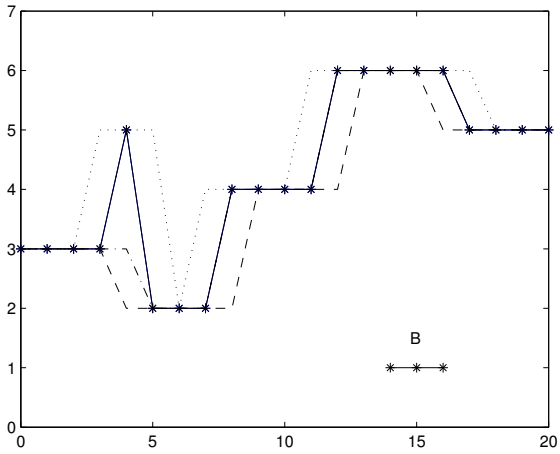

Figure 1: Erosion $(--)$, dilation $(\cdot)$ and opening $(-\cdot-)$ by B on a 1-D discrete signal $(-)$ superimposed. B is a line segment of three points, (-1,0,1).

2.3.2 Morphological covariance

The morphological covariance of a function f is defined by

$$\int_a^b \varepsilon_{P_{2,\mathbf{v}}}(f)(t)dt, \tag{3}$$

where $P_{2,\mathbf{v}}$ is a pair of points separated by a vector $\mathbf{v}$, i.e. $\{0, v\}$. It could be interpreted as the autocorrelation function by looking at the evolution of the integral value for an increasing vector length.

The estimated morphological covariance for a set of vectors is the feature vector used in the classification stage. As in practical situations curves are not observed continuously but at design points, in order to compute the morphological covariance, integrals are numerically approximated (the trapezoidal rule can be used (Press et al., 1992)).

An illustrative example of the morphological covariance of a 1-D signal f (upper panel)

is provided in Fig. 2 (lower-left panel), together with the sample autocorrelation function of the signal (lower-right panel). The original signal $f = sin(20\pi t) + sin(10\pi t) + 2$ has been discretized at 1001 points (t = 0, 0.001, 0.002, ..., 1). Note that the morphological covariance of a 1-D signal is simply $\int min\{f(t), f(t+v)\}dt$, when the distance between the pair of points is v. In addition, the morphological covariance could be used for automatically determining the period of periodic functions, since peaks of the morphological covariance indicate the period, as can be easily shown from the morphological covariance definition. Although both morphological covariance and autocorrelation functions compare the signal and its translation, they are not always similar. For example, if $f_1 = sin(20\pi t) + 20t$ and $f_2 = 20t$ in the interval $(0, 1)$, both autocorrelation functions are very similar, while morphological covariance is able to discriminate between them.

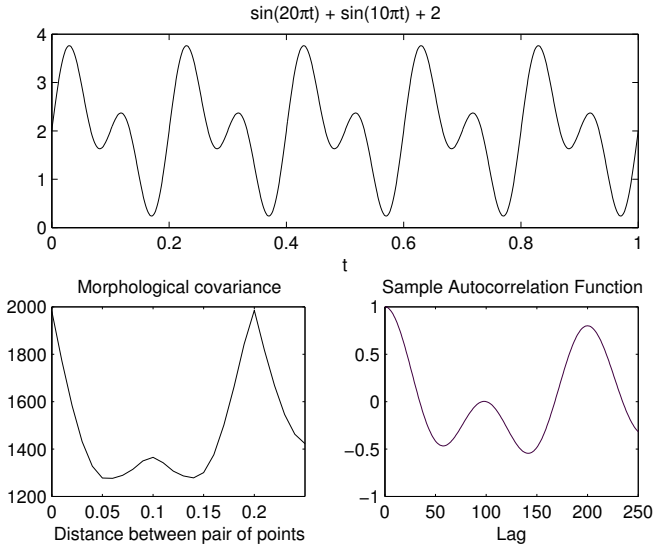

Figure 2: The upper panel shows a function. Its morphological covariance and sample autocorrelation function are displayed in the lower-left and lower-right panels, respectively.

The grid of values of $\mathbf{v}$ in (3) where the morphological covariance is estimated is also determined by MC CV. One possible means of determining these values is to represent the morphological covariance for the average or a subset of each group and to inspect

these graphics in order to visually detect zones where the differences among groups are more noticeable. Morphological covariance could be estimated in these zones. But a more suitable and objective means is to consider different starting points (for example, from 1 to 5) and to try out several increments (for example, from 2 to 10).

2.3.3 Spatial size distributions

Spatial size distributions were proposed by Ayala and Domingo (2001). They combine a granulometric analysis of the function with the auto-correlation function; the usual granulometric size distribution arises as a particular case of this formulation. Granulometries are an axiomatization of the sieving process. A size distribution of a powder can be obtained by passing it through progressively finer sieves and by measuring the mass retained by each sieve. Matheron (1975) formalized this process. A particularly useful example in the applications is a family of openings by αB of increasing size α, where $\alpha B = \{\alpha b : b \in B\}$ stands for the set homothetic to B with a positive scale α.

The complete development of the spatial size distributions can be found in Ayala and Domingo (2001). As we just estimate them for the particular case of using openings, we only introduce this particular definition and not the general one. The spatial size distribution S of a function f at point (α, β) is:

$$S(\alpha, \beta) = \frac{\int_{\beta U} \int_a^b f(t)f(t+h) - \gamma_{B(\alpha)}(f)(t)\gamma_{B(\alpha)}(f)(t+h)dtdh}{(\int_a^b f(t)dt)^2}, \tag{4}$$

where U is a closed interval containing the origin, $B(\alpha)$ is the structuring element of size α, and β, $\alpha \geq 0$. If $B(\alpha)$ was a closed interval, S would be a cumulative distribution function. However, if $B(\alpha)$ is not compact and convex, we do not obtain a cumulative distribution function, but we can still use the estimates as descriptive characteristics. In this paper, the best results are obtained with $B(\alpha) = P_{2,\alpha} = \{0, \alpha\}$. Therefore, results will be presented for this structuring element. Again, in practice curves are discretized, so integrals are numerically approximated. Our design points are equally spaced by s, therefore U is chosen

as $[-s, s]$. In the particular case that β equals zero, we consider the following value for $(\alpha, 0)$:

$$\frac{\int_a^b f(t)^2 - \gamma_{B(\alpha)}(f)(t)^2 dt}{(\int_a^b f(t)dt)^2}.$$

Fig. 3 shows two discretized curves at 1001 points (t from 0 to 100 in increments of 0.1). Both curves are a sum of four identical Normal probability density functions, but located at different points. We have computed S for both functions, using line segments of increasing length. For illustrative purposes, we have also calculated the corresponding discretized joint densities ($s(\alpha, \beta) = S(\alpha, \beta)$ - $S(\alpha, \beta-1)$ - $S(\alpha-1, \beta) + S(\alpha-1, \beta-1)$) since they are easier to interpret visually. Fig. 4 displays the corresponding joint densities. The first marginal of the spatial size distribution is the size component, while the second marginal is concerned with the spatial arrangement. The peak observed in the first marginal (α) indicates the size (variance) of the Normal functions, which is identical for the two curves. However, the locations of the different Normal functions are dissimilar for the two curves, and this behavior is captured by the second component (β), where the peaks indicate the separation between the Normal functions. In the first case, separations are multiples of 20, while in the second case we have a peak for each separation between the Normal functions: 10 (third and fourth Normal), 20 (second and third Normal), 30 (second and fourth Normal) and 40 (first and second Normal). See Ayala and Domingo (2001) for more illustrative examples.

The estimates of S at several points constitute the feature vector used in the classification stage. The discrete set of points where S is estimated is also determined by MC CV. Analogously to the morphological covariance, a suitable means is to consider different starting points (for example, from 1 to 5 for α and $\{0, 5, 10, 15\}$ for β) and to try out several increments (for example, from 2 to 10 for α and in increments of 5 for β).

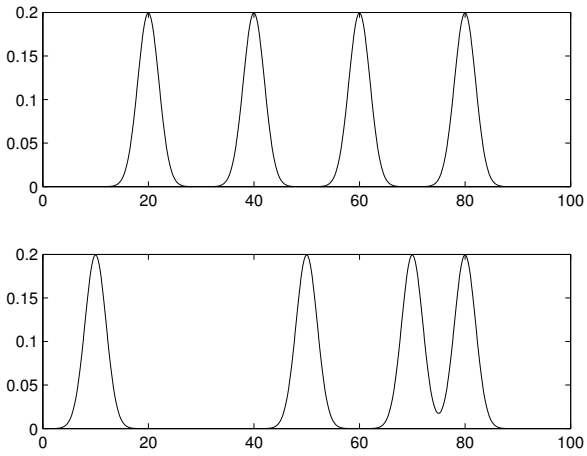

Figure 3: Two discretized curves (see the text for details).

3 Experimental results

In this section the proposed descriptors are compared with several existing alternative approaches through three different problems. The first two problems were considered in Ferraty and Vieu (2003), where they performed a comparative study with some standard methods and more recent techniques. The first problem is a functional version of well-known simulated data, namely waveform data (Breiman et al., 1984). The second problem deals with phoneme classification. Furthermore, we study a third problem which partially motivated this work: a biomechanical application.

This paper has a strong computational component. The work is done by using MatLab (for computing morphological descriptors), SPSS and the free software R (R Development Core Team, 2007). Some of the libraries of R used are: *fda* (Ramsay and Wickham, 2007), *mda* (Hastie and Tibshirani, 2006) and *pls* (Wehrens and Mevik, 2006). There are several algorithms for obtaining ICA. We use the FastICA algorithm available for MatLab and R (`http://www.cis.hut.fi/projects/ica/fastica/`), with the default parameters (Hyvärinen, 1999).

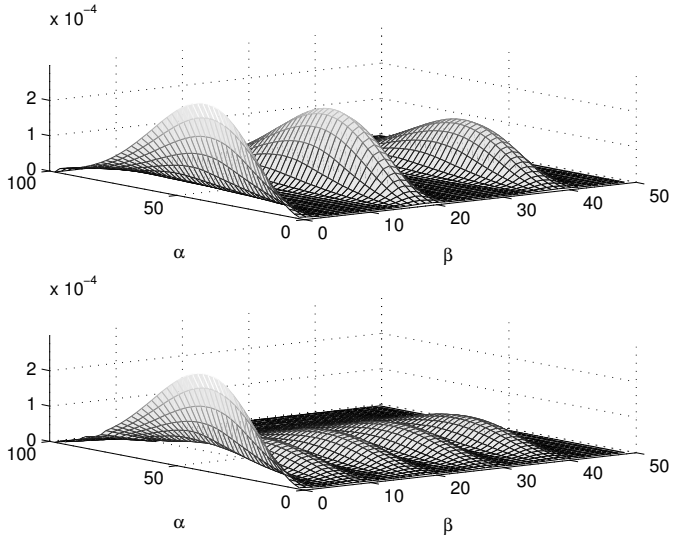

Figure 4: The joint spatial size densities of curves in Fig. 3.

3.1 Waveform data

The original waveform data introduced by Breiman et al. (1984) and used more recently by Hastie et al. (1994), were treated as functional data by Ferraty and Vieu (2003). They simulated a digitized curve x at 101 points ($t = 1, 1.2, 1.4, ..., 21$) such that

- $x(t) = uh_1(t) + (1-u)h_2(t) + \epsilon(t)$ for class 1,
- $x(t) = uh_1(t) + (1-u)h_3(t) + \epsilon(t)$ for class 2, and
- $x(t) = uh_2(t) + (1-u)h_3(t) + \epsilon(t)$ for class 3,

where u is uniform on $[0, 1]$, $\epsilon(t)$ are standard normal variables, and the h_i are the shifted triangular waveforms: $h_1(t) = max(6 - |t - 11|, 0)$, $h_2(t) = h_1(t - 4)$ and $h_3(t) = h_1(t + 4)$.

The sample sizes are the same as in Ferraty and Vieu (2003). Each training sample contains 150 observations in each class, whereas each validation (or test) sample has 750 observations, 250 in each class, in the case where we are in the model selection phase (or in the final model assessing phase). Note that we generate 100 different full datasets, 50 for

MC CV selecting the best parameters, and another 50 datasets for assessing the error, that is to say 120000 curves. The same datasets are used for the different methods. A sample of five curves for each of the three classes appears in Fig. 5.

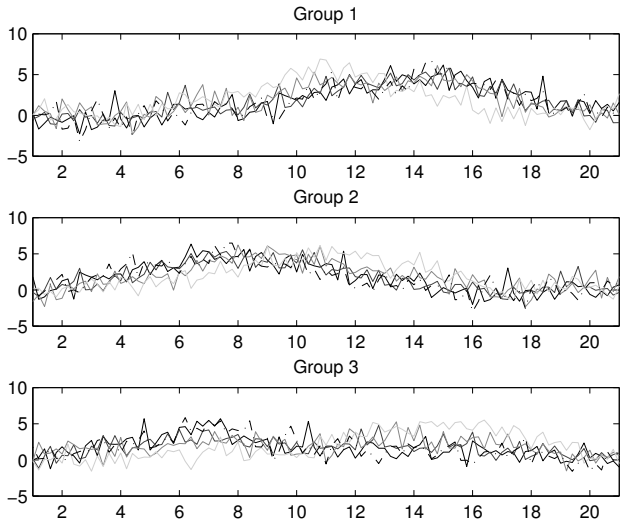

Figure 5: Five waveform curves for each class.

Ferraty and Vieu (2003) performed a wide comparative study, computing firstly the results for the classification and regression tree procedure by Breiman et al. (1984) (CART); secondly, the flexible discriminant analysis proposed by Hastie et al. (1994) (FDA) using three regression methods; thirdly, the multivariate partial least-squares regression adapted to the classification setting (Martens and Naes, 1989) (MPLSR); and fourthly, the penalized discriminant analysis introduced by Hastie et al. (1995) using the classical ridge penalty (PDA/Ridge). They also considered the nonparametric method using kernel and semi-metrics based on PCA introduced by themselves (Ferraty and Vieu, 2003), which also corresponds to the method proposed by Hall et al. (2001) (NPCD/PCA). Finally, they used the nonparametric method including the MPLSR method in its semi-metric, also introduced by themselves (Ferraty and Vieu, 2003) (NPCD/MPLSR).

As the data are somewhat noisy, it could be worth smoothing the data before calculating the proposed descriptors. Shape descriptors are computed for the raw data and for the smoothed data with a different smoothing parameter λ (we consider $\lambda = 0.01$, 1, 10, 100 and 1000). Specifically, data are smoothed by roughness penalized spline smoothing (Ramsay and Silverman, 2005). We use 32 basis functions of B-splines of order 6, knots are equally spaced, and the roughness penalty is the integrated squared second derivative. Finally, we choose the λ which provides the best results by MC CV. All combinations of smoothing and parameters for each method are considered by MC CV in each model selection phase for the different descriptors.

For the descriptors based on statistical moments, the best MC CV results are achieved by first smoothing the data with $\lambda = 1000$ (although the results did not change very much, less than 0.5% if we worked with the raw data), taking as the feature vector the first five moments (of order 0, 1, 2, 3 and 4) and then using quadratic discrimination.

Fig. 6 shows the average test error rates with quadratic discrimination over 50 new randomly built samples as a function of the number of moments and λ.

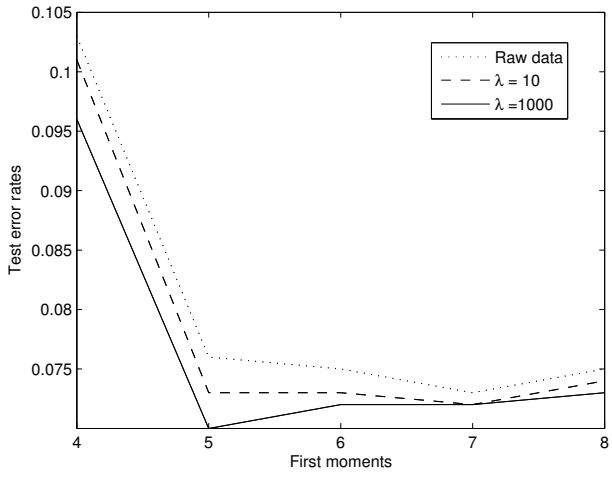

Figure 6: Waveform data as in Ferraty and Vieu (2003): average rates of misclassification over 50 samples as a function of the number of moments and smoothing parameter λ.

For the descriptors based on ICA coefficients, we work with the raw data and the dimension is reduced to three by PCA. Afterwards, the ICA coefficients for the three components are computed and used with the quadratic classifier. The performance of test error rates over 50 random samples with quadratic and linear classifiers and a different number of components is given in Table 1. The first three principal components (of centered curves) and independent components for a training set are shown in Fig. 7 for illustration. Note that if a reconstruction of a curve based on these PCA components had to be done, the location should have to be added. The second (third) ICA component is similar to the first (second) PCA component. In order to aid their interpretation, the independent components are subtracted from the overall mean function. The first of these functions resembles h_3, the generating function, whereas the second and third functions resemble the mean functions for the second and third group, respectively.

Table 1: Waveform data as in Ferraty and Vieu (2003): average rates of misclassification with quadratic and linear classifiers over 50 samples as a function of the number of ICA components.

No. components	2	3	4	5	6	7
Quadratic	0.262	0.065	0.066	0.067	0.069	0.070
Linear	0.323	0.070	0.071	0.073	0.074	0.077

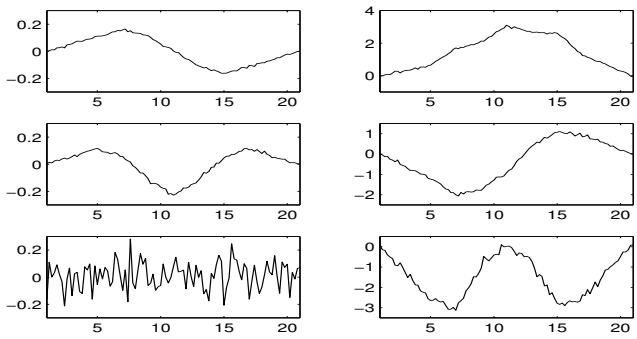

Figure 7: Waveform data as in Ferraty and Vieu (2003): the principal component (left hand panel) and independent component (right hand panel) solutions.

Before using morphological descriptors, the data are smoothed by $\lambda = 100$. The morphological covariance is estimated at 6 points from 50 to 65, in increments of 3 (the zone where the differences among groups are more noticeable). Then the quadratic classifier is employed. For the S, values of α are taken from 3 to 75 in increments of 9, and β varies from 0 to 45 in increments of 15, giving a maximum of 36 characteristics. This time, the linear classifier provides better results.

Table 2 shows the test error rates averaged over 50 simulations, together with the standard deviation for our methods and the most appropriate approaches considered in Ferraty and Vieu (2003), with parameters (in parenthesis) selected to give the smallest misclassification error as in Ferraty and Vieu (2003). The number in brackets for our methods indicates the number of characteristics.

Table 2: Waveform data as in Ferraty and Vieu (2003): test error rates averaged over 50 simulations. Parameter values given in parentheses are discussed in the text.

Method	Test error rates	Std. dev.
Morph. Covariance (6ch)	0.097	0.013
PDA/Ridge ($\lambda = 2000$)	0.087	0.016
NPCD/PCA ($q_{opt} = 3$)	0.079	0.012
NPCD/MPLSR ($q_{opt} = 3$)	0.072	0.012
S (36ch)	0.071	0.012
Moments (5ch)	0.070	0.011
ICA (3ch)	0.065	0.011

Our shape descriptors (excepting the morphological covariance) provide better or similar results, with small standard deviations. We must highlight that ICA descriptors greatly decrease the error rate to 6.5%. In this problem, morphological covariance does not achieve the same discriminatory power as others. Nevertheless, it obtains much better results than FDA or CART (see Ferraty and Vieu (2003)).

We also consider another procedure for assessing performance: in each realization of train/validate/test data, a distinct value of the tuning parameters using only that set is chosen, and is used to assess the performance of the model over the test set. Results are nearly the same: 0.066 for ICA and 0.071 for moments.

3.1.1 Comparisons with recent methods

We compare our descriptors with two recent methods that use the same data: the one by Rossi and Conan-Guez (2005) and the one by Ferré and Villa (2005).

Firstly, the results from Rossi and Conan-Guez (2005) can be directly compared with those shown in Table 2, since the experimental setup is identical. They obtained 0.098 with a classical one hidden layer perceptron, 0.065 with their functional multilayer perceptron approach and 0.072 with an alternative implementation of their approach.

Secondly, although Breiman's waves are also used in Ferré and Villa (2005), the experimental setup is different. We reproduce it in order to compare our descriptors. Training and test samples of size 1500 are also used. Only results for our best descriptors in this example (moments and ICA) are calculated. MC CV is used with raw data for determining the best parameters. Note that curves are discretized at 21 points and equal priors are used (so for each dataset there are roughly 500 observations in each class), which causes misclassification rates to be higher. Table 3 gives a description of the performance of test error rates over 50 random samples for the first seven moments and three ICA components with the linear classifier, together with the best result obtained by functional SIR (Sliced Inverse Regression).

Table 3: Waveform data as in Ferré and Villa (2005): percentages of misclassification over 50 samples (mean, median, standard deviation, first quartile and minimum, respectively).

Method	Mean	Median	Std. dev.	1^{st} Q.	Min.
SIR (Ferré and Villa, 2005) %	15.92	15.93	0.55	15.60	14.73
Moments (7ch) %	14.43	14.23	0.98	13.73	12.27
ICA (3ch) %	13.95	13.97	0.91	13.27	12.2

3.2 Phoneme data

This is the well-known speech recognition problem described by Hastie et al. (1995). The dataset was formed by selecting five phonemes for classification based on digitized speech from the TIMIT database (TIMIT Acoustic-Phonetic Continuous Speech Corpus, NTIS,

US Dept of Commerce). The phonemes were transcribed as follows: "sh" as in "she", "dcl" as in "dark", "iy" as the vowel in "she", "aa" as the vowel in "dark", and "ao" as the first vowel in "water". A total of 4509 speech frames of 32 msec duration were selected. A log-periodogram was computed from each speech frame, which is one of several widely used methods for casting speech data in a form suitable for speech recognition. Thus the data used in what follows consist of 4509 log-periodograms of length 150 (we retain only the first 150 frequencies as in Ferraty and Vieu (2003)). Fig. 8 displays the first five log-periodograms for each class of phoneme.

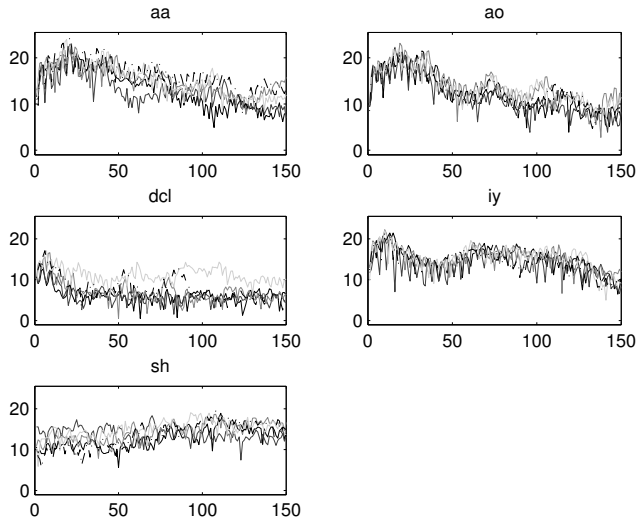

Figure 8: Five log-periodograms for each class.

The experimental setup is analogous to the one in Ferraty and Vieu (2003) in order to compare the results. We randomly build 100 samples of size 2000: 50 for obtaining the best parameters and 50 for assessing the error by MC CV. Each sample is divided into a training sample with 750 log-periodograms (150 per class) and a validation (or test) sample, in the case where we are in the model selection phase (or in the final model assessing phase), which contains 1250 log-periodograms, 250 per class.

As seen in Fig. 8, phoneme data are noisy. Therefore, shape descriptors are applied to

both raw data and smoothed data by roughness penalized spline smoothing, following the same procedure explained in Section 3.1.

With regard to statistical moments, the first fifteen moments with raw data give the best results with quadratic discrimination. Regarding ICA coefficients, the complete dimension after smoothing with $\lambda = 0.01$ is considered. Thus, the feature vector of length 31 is utilized with the linear classifier. This basis resembles wavelets, which are used in speech recognition problems as they provide a representation localized in time and frequency. The morphological covariance is estimated from the raw data, with time gaps running from 3 to 148 in increments of 5, giving a maximum of 30 characteristics. The quadratic classifier is selected, the same as for S, which is computed after smoothing the data with $\lambda = 10$. S is estimated at α from 1 to 141 in increments of 10 and β from 0 to 135 in increments of 45. This gives a maximum of 60 characteristics.

Fig. 9 displays boxplots with test error rates of the 50 test samples for our descriptors and the best standard in this case: PDA/Ridge with shrinkage penalty coefficient $\lambda = 1000$. The performance of moments and ICA descriptors is similar to the best standard. Table 4 presents these results, together with the standards NPCD/PCA ($q_{opt} = 6$) and NPCD/MPLSR ($q_{opt} = 7$).

Table 4: Phoneme data as in Ferraty and Vieu (2003): average percentages of misclassification over 50 simulations.

	Moments	PDA/Ridge	ICA	NPCD/MPLSR	S	Morph. Cov.	NPCD/PCA
%	8.56	8.58	8.84	8.99	10.2	10.2	10.75

A double layer of MC CV (Ripley, 1996, section 2.7.) is also considered: for each one of the 50 samples used for parameter selection, 1250 of the non-selected log-periodograms (250 per class) are chosen to assess the performance, using the respective training sample. The results are similar to the previous ones, with means of 8.43%, 8.45% and 8.82% for PDA/Ridge, moments and ICA descriptors, respectively.

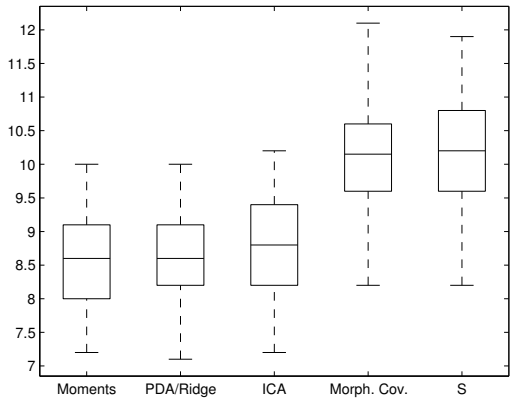

Figure 9: Boxplots of the percentages (%) of misclassification over 50 test samples for phoneme data as in Ferraty and Vieu (2003).

3.2.1 Comparisons with recent methods

We compare our descriptors with three recent methods that use the same data, although the length of the log-periodograms is 256: the one by Ferré and Villa (2006), the one by Conan-Guez and Rossi (2004), and Rossi and Conan-Guez (2005), and finally the one by Rossi and Villa (2006). We reproduce the different experimental setups, and only calculate results for our best descriptors in this example (moments and ICA).

According to the experimental design in Ferré and Villa (2006), we randomly build 50 samples, where both the learning and test samples contain 1735 log-periodograms (347 for each class). The same parameters used previously are considered with the linear classifier. Table 5 gives a description of the performances. SIR-NNr stands for the method introduced in Ferré and Villa (2006) (the functional SIR regularized by penalization, which precedes a neural network), and SIR-NNp stands for a classical SIR (as presented in Ferré and Yao (2003)) as preprocessing of a neural network.

With regard to the work by Conan-Guez and Rossi (2004), and Rossi and Conan-Guez (2005), the training and test sets are fixed and preestablished. They contain 3340 and 1169 log-periodograms respectively. We compute the results for a different number of moments

Table 5: Phoneme data as in Ferré and Villa (2006): percentages of misclassification over 50 samples (mean, median, standard deviation, first quartile and minimum, respectively).

Method	Mean	Median	Std. dev.	1^{st} Q.	Min.
SIR-NNr %	8.21	8.16	0.56	7.90	6.74
SIR-NNp %	8.38	8.24	0.59	7.95	7.20
Moments (15) %	8.46	8.39	0.56	8.18	6.63
ICA (31) %	8.63	8.56	0.53	8.24	7.61
PDA/Ridge %	8.95	8.99	0.54	8.70	7.20
NPCD/PCA %	9.78	9.68	0.65	9.34	8.30

(with raw data, as before) and independent components (with smoothed data with $\lambda = 0.01$, as before) with quadratic discrimination. The classification error rates obtained by a multi-layer perceptron and the functional multi-layer perceptron approach are 8.30% and 7.70% respectively (Conan-Guez and Rossi, 2004). Table 6 shows our results. For selecting a particular model, we split the training set into two parts (proportionally to the original sizes): a new training set with 2474 samples and a validation set with 866 samples. The validation set is used to estimate prediction error for model selection and then, having chosen the final model, we estimate the prediction error on the original test set using the model selected with the original training set. Using this methodology, the first eleven moments and 30 ICA components are chosen, with percentages of misclassification of 7.53% and 7.70%, respectively.

Table 6: Percentages of misclassification over the test set, as a function of the number of moments (M) and ICA components (I), for phoneme data as in Conan-Guez and Rossi (2004), labelled by CGR, and Rossi and Villa (2006), labelled by RV.

M	10		11		12		13		14		15	
CGR %	7.70		7.53		7.70		7.44		7.36		7.87	
RV %	18.68		18.22		19.36		18.91		18.68		20.05	
I	20	21	22	23	24	25	26	27	28	29	30	31
CGR %	8.13	7.44	7.87	8.13	8.13	8.55	8.04	7.70	7.53	7.70	7.70	8.13
RV %	17.77	17.77	18.0	18.91	18.91	19.13	19.13	19.59	18.45	20.05	19.59	18.68

In the paper by Rossi and Villa (2006), the training and test sets are again fixed and preestablished, with sizes of 1278 and 439 respectively. However, they restricted themselves to classifying "aa" against "ao" because this is the most difficult sub-problem in the

database. We compute the results for a different number of moments and independent components, as previously, with the quadratic and linear classifier, respectively. The error rates on the test set for the methods reported in Rossi and Villa (2006) are: 22.10% for Functional Gaussian SVM, 19.36% for Functional linear SVM and 20.05% for Linear SVM. Table 6 summarizes our results. Using the same methodology as before for selecting a particular model, the first eleven moments and 31 ICA components are chosen, with percentages of misclassification of 18.22% and 18.68%, respectively.

3.3 Biomechanical data

As mentioned earlier, a biomechanical application has partially motivated this paper: the functional assessment of lower back pain. The sit-to-stand (STS) movement is of interest in functional evaluations related to balance control, lower extremity dysfunctions or lower back pain.

Three groups, interesting from a clinical point of view, are considered: first, healthy volunteers; second, back pain patients and third, subjects with previous back pain episodes who are pretending to suffer from them still.

The controls consisted of 59 healthy volunteers between the ages of 23 and 65 with no history of locomotor disturbance. The patients group consisted of 42 subjects between the ages of 21 and 62 recruited from the Rehabilitation Department of the Arnau de Vilanova Hospital. All of them had a history of chronic, nonspecific lower back pain. Two criteria were used to select the subjects: a nonzero score on the Jensen pain scale (Jensen et al., 1989) and a low or moderate score on the Oswestry lower back pain disability questionnaire (Fairbank et al., 1980). Patients suffering an acute pain episode were rejected for ethical reasons. Finally, a group of subjects pretending to have current back-pain problems was included in the study. This group consisted of 13 volunteers between the ages of 25 and 56 having suffered a previous lower back pain episode at least two years before the experiment, but with no current symptoms. They were instructed to remember the disability associated

with back pain and to sham limitations in their STS movement, as if they were trying to trick the examiner, with the aim of obtaining possible benefit from their insurance company.

During the experiment each subject performed five trials of STS movement. The starting position was sitting with their arms across their chest on a height-adjustable chair. The subjects were instructed to rise until reaching a relaxed standing position.

Two force platforms ($Dinascan - IBV$) were used to register ground reaction forces on both feet during the experiments. These recordings were added in order to obtain the resulting force. Normalization with the subjects' weights was done and ground reaction force was expressed as a percentage of that corporal weight.

Before testing the different methods, a smoothing (60 cubic B-spline basis functions are considered) and registration process for the different replications of a given subject are carried out, applying the R function *registerfd* included in the package *fda*, using the minimum eigenvalue of a cross-product matrix as the continuous registration criterion and the mean of the replications as the target function (Ramsay and Silverman, 2005). Our final data are the means of the five registered observations for each individual (note that 114 registration processes are carried out). Fig. 10 shows these functions for the individuals in each group. Only the first three seconds of the movement are displayed. We can see that this is not an easy problem.

As the sample size is small (114), we use a leave-one-out strategy to evaluate the performance of the methods. The lowest number of errors achieved for the standard methods PDA/Ridge ($\lambda = 1000$), MPLSR (7 components) and NPCD/PCA ($q = 7$) is 15 (13.16%), and 16 (14.04%) for NPCD/MPLSR ($q = 5$). Regarding moments, we use the first five, giving 16 errors (14.04%) with quadratic discrimination. For ICA coefficients, twelve coefficients are selected, providing 12 errors (10.53%) with linear discrimination. The same number of errors is obtained by S with linear discrimination for eighteen points: α from 5 to 40 in increments of 7 and β from 5 to 105 in increments of 50. Points where morphological covariance is estimated go from 3 to 28 in increments of 5, giving a maximum of 6 charac-

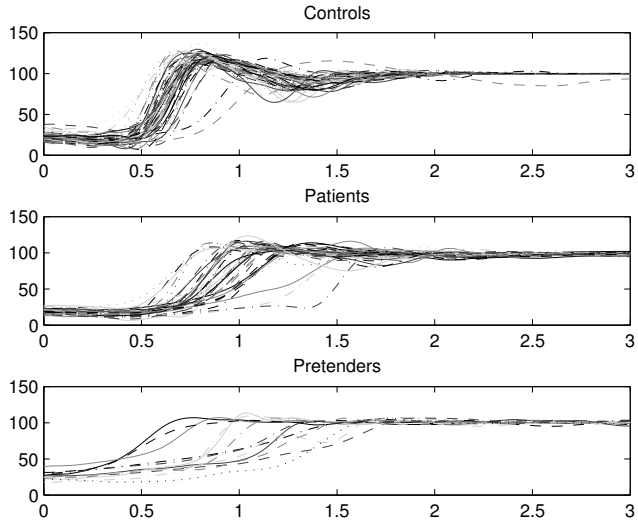

Figure 10: Individual observations for the three groups in the biomechanical application.

teristics. The error obtained with the linear classifier is 14 (12.28%). Therefore, the best results are achieved by the ICA coefficients and S, followed by morphological covariance. Note that data are not smoothed by a roughness penalty as in the previous examples, since these data are not noisy.

Table 7 shows the confusion matrices for ICA coefficients (the first number in each cell) and S (the second number in each cell), where we can identify different behavior. We obtain a nearly perfect classification for the pretenders group for the ICA coefficients, whereas S classifies the control group nearly perfectly. If we use both sets of descriptors with the linear classifier, the number of errors is reduced to 10 (8.77%), with nearly correct classification for the control and pretenders groups, as the third numbers in each cell of Table 7 indicate. However, if we consider several descriptors simultaneously, we need to proceed with caution, since this could overfit the training sample and generalize new samples poorly (Hastie et al., 2001).

Table 7: Biomechanical problem: confusion matrices (in absolute counts) for ICA coefficients, S and their combination. Entries are obtained using leave-one-out cross-validation.

ICA coefficients / S / Combination of ICA and S			
	Control	Patients	Pretenders
Control	55 / 58 / 57	2 / 1 / 1	2 / 0 / 1
Patients	5 / 5 / 2	36 / 36 / 36	1 / 1 / 4
Pretenders	0 / 0 / 0	2 / 5 / 2	11 / 8 / 11

4 Conclusions

We have tackled the curve discrimination problem from the point of view of shape descriptors. Three different kinds of shape descriptors have been studied: statistical moments, coefficients of ICA components and two mathematical morphology descriptors, morphological covariance and spatial size distributions. They have shown their efficiency in three different problems, where very good results have been achieved using few characteristics.

We can consider which shape descriptor could be the best. As nearly always, there is no single answer. ICA coefficients have given the best results for the waveform data (moments and S have also performed better than the standard methods). However, moments are the best for the phoneme data, the ICA approach being quite similar. For the biomechanical application, the ICA approach and S have been the best. Furthermore, morphological covariance has also given better results than the standard methods.

In short, depending on the problem, some shape descriptors could be more appropriate than others, although good results are obtained in general, better than or similar to those from existing techniques.

The comparisons with recent methods in different scenarios confirm that our descriptors are competitive. Depending on the examples considered, better or equal performance is achieved. For instance, similar performance is obtained if the functional multilayer perceptron of Rossi and Conan-Guez is considered, with both waveform data and phoneme data, although our descriptors provide slightly better results for phoneme data. Regarding the papers by Ferré and Villa with SIR, we obtain similar or slightly worse performances with

phoneme data, but much better performances with waveform data. Finally, the comparison with Rossi and Villa's SVM for functional data reveals that our descriptors improve on their results.

We must also highlight that shape descriptors can be computed easily since there are libraries available for computing them (see Hyvärinen (1999) for the computational efficiency of the FastICA algorithm and Soille (2003) for appropriate algorithms for computing morphological transformations). The moments can even be computed in very few code lines. They are clearly the cheapest method in terms of computational cost. Although considerable cross-validation is considered, this is computationally simple as all the generation of features is fast, and the learning algorithms themselves are also quick.

Future work would be to study other shape descriptors and to extend these ideas to the bivariate (or multivariate) functional case. An easy option in the ICA approach could be to concatenate the observations of the two functions into a single long vector, as is done for computing bivariate functional PCA (Ramsay and Silverman, 2005). For moments and morphological descriptors, an easy option could also be to consider each function separately, i.e. a marginal approach. For morphological descriptors, we would have to tackle the same problems (and alternatives) as morphology for multichannel images, where we have vector valued pixels (Soille, 2003).

On the other hand, our prime objective in this paper was to introduce new shape descriptors that could be used to represent functional data prior to classification. An additional point to study would be to analyze the results if other classifiers were used, such as neural networks, k-nearest neighbors, support vector classifiers, etc., or a combination of them.

Acknowledgements

This work has been partially supported by Grants CICYT MTM2005-08689-C02-02 and TIN2006-10134. The author thanks G. Ayala for attracting her attention to functional data analysis, and M.A. García for his support. The author would like also to thank the

anonymous associate editor and both reviewers for their very constructive suggestions which led to an improvement in this paper.

References

Anderson, C. W., Stolz, E. A., and Shamsunder, S. (1998), "Multivariate auto-regressive models for classification of spontaneous electroencephalographic signals during mental tasks," *IEEE Transactions on Biomedical Engineering*, 45, 277–286.

Ayala, G. and Domingo, J. (2001), "Spatial size distributions: Application to shape and texture analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23, 1430–1442.

Back, A. D. and Weigend, A. S. (1997), "A first application of independent component analysis to extracting structure from stock returns," *International Journal of Neural Systems*, 8, 473–484.

Bell, A. J. and Sejnowski, T. J. (1997), "The independent components of natural scenes are edge filters," *Vision Research*, 37, 3327–3338.

Berlinet, A., Biau, G., and Rouvière, L. (2005), "Functional classification with wavelets," Submitted. Available at http://www.math.univ-montp2.fr/∼biau/bbr3.ps.

Biau, G., Bunea, F., and Wegkamp, M. (2005), "Functional classification in Hilbert spaces," *IEEE Transactions on Information Theory*, 51, 2163–2172.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984), *Classification and Regression Trees*, Wadsworth.

Burman, P. (1989), "A comparative study of ordinary cross-validation, v-fold cross-validation, and the repated learning-testing methods," *Biometrika*, 76, 503–514.

Conan-Guez, B. and Rossi, F. (2004), "Phoneme discrimination with functional multilayer perceptron," in *Classification, Clustering, and Data Mining Applications (Proceedings of IFCS 2004)*, pp. 157–165.

Fairbank, J. C. T., Davies, J. B., Mbaot, J. C., and O'Brien, J. P. (1980), "The oswestry low back pain disability questionnaire," *Physiotherapy*, 66, 71–78.

Ferraty, F. and Vieu, P. (2003), "Curves discrimination: a nonparametric functional approach," *Computational Statistics and Data Analysis*, 44, 161–173.

— (2006), *Nonparametric Functional Data Analysis: Theory and Practice*, Springer.

Ferré, L. and Villa, N. (2005), "Discrimination de courbes par régression inverse fonctionnelle," *Revue de Statistique Appliquée*, LIII, 39–57.

— (2006), "Multilayer perceptron with functional inputs: an inverse regression approach," *Scandinavian Journal of Statistics*, 33, 807–823.

Ferré, L. and Yao, A. (2003), "Functional sliced inverse regression analysis," *Statistics*, 37, 475–488.

Fromont, M. and Tuleau, C. (2006), "Functional classification with margin conditions," in *Proceedings of the 19th Annual Conference on Learning Theory (COLT06)*, pp. 94–108.

Glendinning, R. H. and Fleet, S. L. (2004), "Classifying non-uniformly sampled vector-valued curves," *Pattern Recognition*, 37, 1999–2008.

González, R., Woods, R., and Eddins, S. (2004), *Digital image processing using MATLAB*, Prentice Hall.

Hall, P., Poskitt, D., and Presnell, B. (2001), "A functional data-analytic approach to signal discrimination," *Technometrics*, 43, 1–9.

Hastie, T., Buja, A., and Tibshirani, R. (1995), "Penalized discriminant analysis," *Annals of Statistics*, 23, 73–102.

Hastie, T. and Tibshirani, R. (2006), *mda: Mixture and flexible discriminant analysis. R port by Leisch, F., Hornik, K. and Ripley, B. D.*

Hastie, T., Tibshirani, R., and Buja, A. (1994), "Flexible discriminant analysis by optimal scoring," *Journal of the American Statistical Association*, 89, 1255–1270.

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning. Data mining, inference and prediction*, Springer-Verlag.

Huang, P. S. (2001), "Automatic gait recognition via statistical approaches for extended template features," *IEEE Transactions on Systems, Man and Cybernetics B*, 31, 818–824.

Hyvärinen, A. (1999), "Fast and robust fixed-point algorithms for indepent component analysis," *IEEE Transactions on Neural Networks*, 10, 626–634.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001), *Independent component analysis*, New York: Wiley.

James, G. and Hastie, T. (2001), "Functional linear discriminant analysis for irregularly sampled curves," *Journal of the Royal Statistical Society, Series B*, 533–550.

James, G. M. (2002), "Generalized linear models with functional predictor variables," *Journal of the Royal Statistical Society, Series B*, 411–432.

James, G. M. and Silverman, B. (2005), "Functional adaptive model estimation," *Journal of the American Statistical Association*, 565–576.

Jensen, M. P., Karoly, P., O'Riordan, E. F., Bland, F., and Burns, R. S. (1989), "The subjective experience of acute pain. An assessment of the utility of 10 indices," *Clinical Journal of Pain*, 5, 153–159.

Lingvall, F. and Stepinski, T. (2000), "Automatic detection and classifying defects during eddy current inspection of riveted lap-joints," *Independent Nondestructive Testing and Evaluation*, 33, 47–55.

López-Pintado, S. and Romo, J. (2006), "Depth-based classification for functional data," in *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, vol. 72, pp. 103–120.

Martens, H. and Naes, T. (1989), *Multivariate Calibration*, Wiley.

Marx, B. and Eilers, P. (1999), "Generalized linear regression on sampled signals and curves: a P-spline approach," *Technometrics*, 41, 1–13.

Matheron, G. (1975), *Random Sets and Integral Geometry*, London: Wiley.

Naya, S., Cao, R., and Artiaga, R. (2004), "Nonparametric regression with functional data for polymer classification," in *Proceedings in Computational Statistics*, pp. 1569–1576.

Olshausen, B. A. and Field, D. (1996), "Emergence of simple-cell receptive field properties by a learning a sparse code for natural images," *Nature*, 381, 607–609.

Press, W. H., Flannery, B. P., Teulosky, S. A., and Vetterling, W. T. (1992), *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge: Cambridge University Press.

R Development Core Team (2007), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, URL `http://www.R-project.org`. ISBN 3-900051-07-0.

Ramsay, J. and Silverman, B. W. (2002), *Applied Functional Data Analysis*, Springer.

Ramsay, J. O. and Silverman, B. W. (2005), *Functional Data Analysis*, Springer.

Ramsay, J. O. and Wickham, H. (2007), *fda: Functional Data Analysis*, URL `http://www.functionaldata.org`.

Ripley, B. D. (1996), *Pattern recognition and neural networks*, Cambridge University Press.

Rivest, J. (2006), "Granulometries and pattern spectra for radar signals," *Signal Processing*, 86, 1094–1103.

Rossi, F. and Conan-Guez, B. (2005), "Functional multi-layer perceptron: a nonlinear tool for functional data analysis," *Neural Networks*, 18, 45–60.

Rossi, F., Delannay, N., Conan-Guez, B., and Verleysen, M. (2005), "Representation of functional data in neural networks," *Neurocomputing*, 64, 183–210.

Rossi, F. and Villa, N. (2006), "Support vector machine for functional data classification," *Neurocomputing*, 69, 730–742.

Shutler, J. (2002), "Statistical moments," In CV online: On-Line Compendium of Computer Vision. Available: http://homepages.inf.ed.ac.uk/rbf/CVonline/.

Soille, P. (2003), *Morphological Image Analysis. Principles and Applications*, 2nd ed., Springer-Verlag.

Stoyan, D. and Stoyan, H. (1994), *Fractals, Random Shapes and Point Fields. Methods of Geometrical Statistics*, Wiley.

Wehrens, R. and Mevik, B.-H. (2006), *pls: Partial Least Squares Regression (PLSR) and Principal Component Regression (PCR)*, URL `http://mevik.net/work/software/pls.html`.