Submitted to Annales de l'ISUP

FUNCTIONAL SUPERVISED CLASSIFICATION WITH WAVELETS

By Alain Berlinet, Gérard Biau and Laurent Rouvière

Université Montpellier II, Université Paris VI and Université Rennes II

Let X be a random variable taking values in a Hilbert space and let Y be a random label with values in $\{0, 1\}$. Given a collection of classification rules and a learning sample of independent copies of the pair (X, Y), it is shown how to select optimally and consistently a classifier. As a general strategy, the learning sample observations are first expanded on a wavelet basis and the overall infinite dimension is reduced to a finite one via a suitable data-dependent thresholding. Then, a finite-dimensional classification rule is performed on the non-zero coefficients. Both the dimension and the classifier are automatically selected by data-splitting and empirical risk minimization. Applications of this technique to a signal discrimination problem involving speech recordings and simulated data are presented.

1. Introduction.

1.1. *Functional classification.* The problem of classification (or pattern recognition or discrimination) is about guessing or predicting the unknown class of an observation. An observation is usually a collection of numerical measurements represented by a d-dimensional vector. However, in many real-life problems, input observations are in fact (sampled) functions rather than standard high dimensional vectors, and this casts the classification problem into the class of Functional Data Analysis.

The last few years have witnessed important new developments in both the theory and practice of functional classification and related learning problems. Nonparametric techniques have been proved useful for analyzing such functional data, and the literature is growing at a fast pace: Hastie, Buja, and Tibshirani [24] set out the general idea of Functional Discriminant Analysis; Kulkarni and Posner [26] study rates of convergence of k-nearest neighbor regression estimates in general spaces; Hall, Poskitt, and Presnell [23] employ a functional data-analytic method for dimension reduction based on Principal Component Analysis and perform Quadratic Discriminant Analysis on the reduced space, so do Ramsay and Silverman [30, 31]; Ferraty

AMS 2000 subject classifications: Primary 62G10; secondary 62G05

Keywords and phrases: classification, functional data analysis, wavelets, empirical risk minimization, data-splitting, thresholding, shatter coefficient

and Vieu [18, 19] estimate nonparametrically the posterior probability of an incoming curve in a given class; Cardot and Sarda [9] develop functional generalized linear models; Cuevas, Febrero, and Fraiman [12] use depth notions to compute robust distances between curves, whereas Rossi and Villa [32] investigate the use of Support Vector Machines in the context of Functional Data Analysis. For a large discussion and an updated list of references, we refer the reader to the monographs of Ramsay and Silverman [30] and Ferraty and Vieu [20].

Although standard pattern recognition techniques appear to be feasible, the intrinsic infinite-dimensional structure of the observations makes learning suffer from the curse of dimensionality (see Abraham, Biau, and Cadre [1] for a detailed discussion, examples and counterexamples). In practice, before applying any learning technique to modelize real data, a preliminary dimension reduction or model selection step reveals crucial for appropriate smoothing and circumvention of the dimensionality effect. As a matter of fact, filtering is a popular dimension reduction method in signal processing and this is the central approach we take in this paper.

Roughly, filtering reduces the infinite dimension of the data by considering only the first d coefficients of the observations expanded on an appropriate basis. This approach was followed by Kirby and Sirovich [25], Comon [11], Belhumeur, Hepana, and Kriegman [3], Hall, Poskitt, and Presnell [23], or Amato, Antoniadis, and De Feis [2], among others. Given a collection of functions to be classified, Biau, Bunea, and Wegkamp [4] propose to use first Fourier filtering on each signal, and then perform k-nearest neighbor classification in $\mathbb{R}^d$. These authors study finite sample and asymptotic properties of a data-driven procedure that selects simultaneously both the dimension d and the optimal number of neighbors k.

The present paper breaks with three aspects of the methodology described by Biau, Bunea, and Wegkamp [4].

- First a change which can appear as minor but which has major practical implications is in the choice of the basis. As pointed out for example in Amato, Antoniadis, and De Feis [2], wavelet bases offer some significant advantages over other bases. In particular, unlike the traditional Fourier bases, wavelets are localized in both time and frequency. This offers advantages for representing processes that have discontinuities or sharp peaks.
- Second, reordering of the basis using a data-based criterion allows efficient reduction of dimension.
- Finally the classification rule is not restricted to the nearest neighbor

rule as in Biau, Bunea, and Wegkamp [4]. This allows to adapt the rule to the problem under study.

Throughout the manuscript, we will adopt the point of view of automatic pattern recognition described, to a large extent, in Devroye [15]. In this setup, one uses a validation sequence to select the best rule from a rich class of discrimination rules defined in terms of a training sequence. For the clarity of the paper, all important concepts and inequalities regarding this classification paradigm are summarized in the next subsection. In Section 2, we outline our method, state its finite sample performance and prove consistency of the classification rule. Section 3 offers some experimental results both on real-life and simulated data.

1.2. *Automatic pattern recognition.* This section gives a brief exposition and sets up terminology of automatic pattern recognition. For a detailed introduction, the reader is referred to Devroye [15].

To model the automatic learning problem, we introduce a probabilistic setting. Denote by $\mathcal{F}$ some abstract separable Hilbert space, and keep in mind that the choice $\mathcal{F} = L_2([0,1])$ (that is, the space of all square integrable functions on $[0,1]$) will be a leading example throughout the paper. The data consists of a sequence of $n+m$ i.i.d. $\mathcal{F} \times \{0,1\}$-valued random variables $(X_1, Y_1), \ldots, (X_{n+m}, Y_{n+m})$. The X_i's are the *observations*, and the $Y_i's$ are the *labels*[1]. Note that we artificially split the data into two independent sequences, one of length n, and one of length m: we call the n sequence the *training sequence*, and the m sequence the *validation sequence*. A discrimination rule is a (measurable) function $g : \mathcal{F} \times (\mathcal{F} \times \{0,1\})^{n+m} \to \{0,1\}$. It classifies a new observation $x \in \mathcal{F}$ as coming from class $g\,(x, (X_1, Y_1), \ldots, (X_{n+m}, Y_{n+m}))$. We will write $g(x)$ for the sake of convenience.

The probability of error of a given rule g is

$$L_{n+m}(g) = \mathbf{P}\{g(X) \neq Y | (X_1, Y_1), \ldots, (X_{n+m}, Y_{n+m})\},$$

where (X, Y) is independent of the data sequence and is distributed as (X_1, Y_1). Although we would like $L_{n+m}(g)$ to be small, we know (see e.g. Devroye, Györfi, and Lugosi [16], Theorem 2.1, page 10), that $L_{n+m}(g)$ cannot be smaller than the Bayes probability of error

$$L^* = \inf_{s:\mathcal{F}\to\{0,1\}} \mathbf{P}\{s(X) \neq Y\}.$$

[1] In this study we restrict our attention to binary classification. The reason is simplicity and that the binary problem already captures many of the main features of more general problems.

In the learning process, we aim at constructing rules with probability of error as close as possible to L^*. To do this, we employ the training sequence to design a class of data-dependent discrimination rules and we use the validation sequence as an impartial judge in the selection process. More precisely, we denote by $\mathbf{D}_n$ a (possibly infinite) collection of functions $g : \mathcal{F} \times (\mathcal{F} \times \{0,1\})^n \to \{0,1\}$, from which a particular function $\hat{g}$ is selected by minimizing the *empirical risk* based upon the validation sequence:

$$\hat{L}_{n,m}(\hat{g}) = \frac{1}{m} \sum_{i=n+1}^{n+m} \mathbf{1}_{[\hat{g}(X_i) \neq Y_i]} = \min_{g \in \mathbf{D}_n} \frac{1}{m} \sum_{i=n+1}^{n+m} \mathbf{1}_{[g(X_i) \neq Y_i]}.$$

At this point, observe that in the formulation above, for $x \in \mathcal{F}$,

$$g(x) = g(x, (X_1, Y_1), \ldots, (X_n, Y_n))$$

and

$$\hat{g}(x) = \hat{g}(x, (X_1, Y_1), \ldots, (X_{n+m}, Y_{n+m})),$$

i.e., the discriminators g themselves are based upon the training sequence only, whereas the chosen classifier $\hat{g}$ depends on the *entire data set*, as the rest of the data is used for selecting the classifiers.

Since, conditionally to the training sequence, $\hat{L}_{n,m}(g)$ is an unbiased estimate of $L_n(g)$, we expect that $L_{n+m}(\hat{g})$ is close to $\inf_{g \in \mathbf{D}_n} L_n(g)$. This is captured in the following inequality (see Devroye, Györfi, and Lugosi [16], Lemma 8.2, page 126):

$$L_{n+m}(\hat{g}) - \inf_{g \in \mathbf{D}_n} L_n(g) \leq 2 \sup_{g \in \mathbf{D}_n} \left| \hat{L}_{n,m}(g) - L_n(g) \right|. \tag{1.1}$$

Thus, upper bounds for $\sup_{g \in \mathbf{D}_n} |\hat{L}_{n,m}(g) - L_n(g)|$ provide us with upper bounds for the suboptimality of $\hat{g}$ within $\mathbf{D}_n$. When the class of rules $\mathbf{D}_n$ is finite with cardinality bounded by N_n, upper bounds can be obtained via a direct application of Hoeffding's inequality:

$$\mathbf{E}_n \left\{ \sup_{g \in \mathbf{D}_n} \left| \hat{L}_{n,m}(g) - L_n(g) \right| \right\} \leq \sqrt{\frac{\log(2N_n)}{2m}} + \frac{1}{\sqrt{8m \log(2N_n)}}, \tag{1.2}$$

where the notation $\mathbf{E}_n$ means the expectation conditional on the training sequence of length n. The inequality above is useless when $N_n = \infty$. It is here that we can apply the inequality of Vapnik and Chervonenkis [34] or one of its modifications. We first need some more notation. For fixed training sequence $(x_1, y_1), \ldots, (x_n, y_n)$, denote by $\mathbf{C}_n$ the collection of all sets

$$\mathbf{C}_n = \Big\{ \{x \in \mathcal{F} : g(x) = 1\} : g \in \mathbf{D}_n \Big\},$$

and define the shatter coefficient as

$$\mathbb{S}_{\mathbf{C}_n}(m) = \max_{(x_1,\ldots,x_m)\in\mathcal{F}^m} \text{Card } \{\{x_1,\ldots,x_m\} \cap C : C \in \mathbf{C}_n\}.$$

Then

$$\mathbf{E}_n\left\{ \sup_{g\in\mathbf{D}_n} \left|\hat{L}_{n,m}(g) - L_n(g)\right|\right\}$$

$$\leq \sqrt{\frac{8\log\left(4\mathbb{S}_{\mathbf{C}_n}(2m)\right)}{m}} + \frac{1}{\sqrt{(m/2)\log\left(4\mathbb{S}_{\mathbf{C}_n}(2m)\right)}}. \tag{1.3}$$

For more information and improvements on these inequalities, we refer the reader to the monograph of Devroye, Györfi, and Lugosi [16], and to the comprehensive surveys of Boucheron, Bousquet, and Lugosi [5, 6].

2. Dimension reduction for classification. The theory of wavelets has recently undergone a rapid development with exciting implications for nonparametric estimation. Wavelets are functions that can cut up a signal into different frequency components with a resolution matching its scale. Unlike the traditional Fourier bases, wavelet bases offer a degree of localization in space as well as frequency. This enables development of simple function estimates that respond effectively to discontinuities and spatially varying degree of oscillations in a signal, even when the observations are contaminated by noise. The books of Daubechies [14] and Meyer [27] give detailed expositions of the mathematical aspects of wavelets.

As for now, to avoid useless technical notation, we will suppose that the feature space $\mathcal{F}$ is equal to the Hilbert space $L_2([0,1])$, and we will sometimes refer to the observations X_i as "the curves". Extension to more general separable Hilbert spaces is easy. We recall that $L_2([0,1])$ can be approximated by a *multiresolution analysis*, i.e., a ladder of closed subspaces

$$V_0 \subset V_1 \subset \ldots \subset L_2([0,1])$$

whose union is dense in $L_2([0,1])$, and where each V_j is spanned by 2^j orthonormal scaling functions $\phi_{j,k}, k = 0,\ldots,2^j-1$. At each resolution level $j \geq 0$, the orthonormal complement W_j between V_j and V_{j+1} is generated by 2^j orthonormal wavelets $\psi_{j,k}, k = 0,\ldots,2^j-1$ obtained by translations and dilatations of a function ψ (called *mother wavelet*):

$$\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k).$$

As an illustration, Figure 1 displays Daubechies' mother wavelets with $p = 1, 2, 4, 6, 8$ and 10 vanishing moments (Daubechies [14]). Note that the case $p = 1$ corresponds to the Haar basis (Haar [22]).

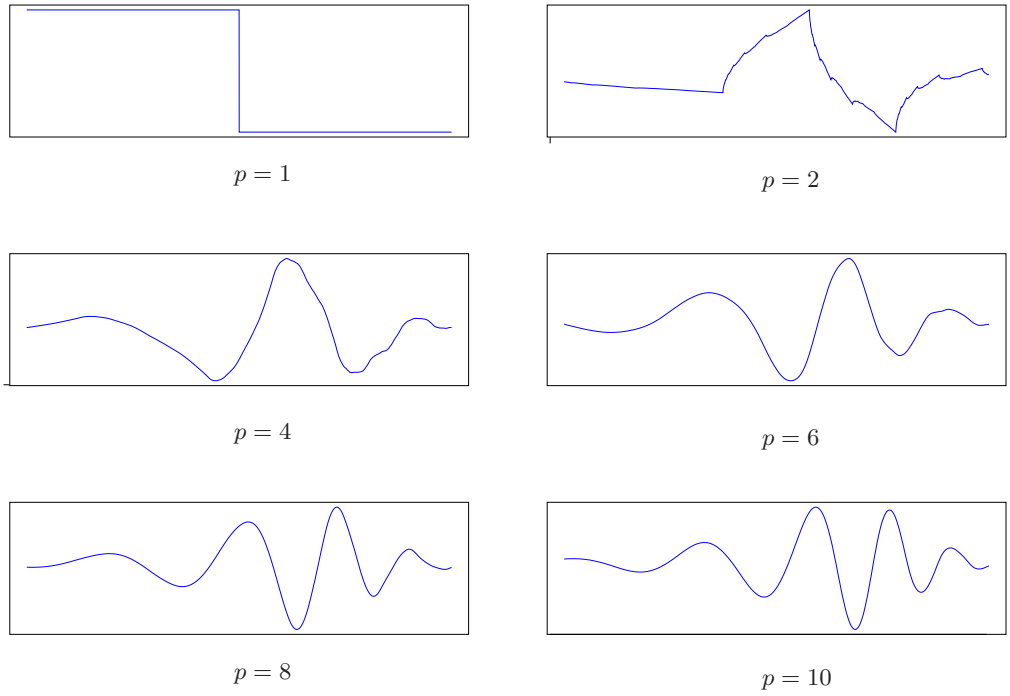

FIG 1. *Mother wavelets for Daubechies' compactly supported wavelets with p vanishing moments ($p = 1, 2, 4, 6, 8$ and 10).*

Thus, the family

$$\bigcup_{j\geq 0} \{\psi_{j,k}\}_{k=0,\dots,2^j-1}$$

completed by $\{\phi_{0,0}\}$ forms an orthonormal basis of $L_2([0,1])$. As a consequence, any observation X in $L_2([0,1])$ reads

$$X(t) = \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} \zeta_{j,k}\psi_{j,k}(t) + \eta\phi_{0,0}(t), \quad t \in [0,1], \tag{2.1}$$

where

$$\zeta_{j,k} = \int_0^1 X(t)\psi_{j,k}(t)\mathrm{d}t \quad \text{and} \quad \eta = \int_0^1 X(t)\phi_{0,0}(t)\mathrm{d}t,$$

and the consistency (2.1) is understood in $L_2([0,1])$. We are now ready to introduce our classification algorithm and discuss its consistency properties. Using the notation of Subsection 1.2, we suppose that the data consists of a sequence of $n+m$ i.i.d. $L_2([0,1]) \times \{0,1\}$-valued random observations $(X_1, Y_1), \ldots, (X_{n+m}, Y_{n+m})$. Given a multiresolution analysis of $L_2([0,1])$, each observation X_i is expressed as a series expansion

$$X_i(t) = \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} \zeta_{j,k}^i \psi_{j,k}(t) + \eta^i \phi_{0,0}(t), \quad t \in [0,1]. \tag{2.2}$$

It will be convenient to reindex the sequence

$$\{\phi_{0,0}, \psi_{0,0}, \psi_{1,0}, \psi_{1,1}, \psi_{2,0}, \psi_{2,1}, \psi_{2,2}, \psi_{2,3}, \psi_{3,0}, \ldots\}$$

into $\{\psi_1, \psi_2, \ldots\}$. With this scheme, equality (2.2) may be rewritten as

$$X_i(t) = \sum_{j=1}^{\infty} X_{ij} \psi_j(t), \quad t \in [0,1], \tag{2.3}$$

with the random coefficients

$$X_{ij} = \int_0^1 X_i(t) \psi_j(t)\, \mathrm{d}t.$$

Denote by $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots)$ the sequence of coefficients associated with X_i. In our quest of dimension reduction, we first fix in (2.2) a maximum (large) resolution level J ($J \geq 0$, possibly function of n) so that

$$X_i(t) \approx \sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} \zeta_{j,k}^i \psi_{j,k}(t) + \eta^i \phi_{0,0}(t), \quad t \in [0,1]$$

or equivalently, using (2.3),

$$X_i(t) \approx \sum_{j=1}^{2^J} X_{ij} \psi_j(t), \quad t \in [0,1].$$

At this point, we could try to use these finite-dimensional approximations of the observations, and let the data select optimally one of the $2^{2^J} - 1$ non-empty subbases of $\{\psi_1, \ldots, \psi_{2^J}\}$. By doing so, we would be faced with catastrophic performance bounds and unreasonable computing time. To circumvent this problem, we suggest the following procedure.

First, for each $d = 1, \ldots, 2^J$, we assume to be given beforehand a (possibly infinite) collection $\mathbf{D}_n^{(d)}$ of rules $g^{(d)} : \mathbb{R}^d \times (\mathbb{R}^d \times \{0,1\})^n \to \{0,1\}$ working in $\mathbb{R}^d$ and using the n d-dimensional training data as input. We will denote by $\mathbb{S}_{\mathbf{C}_n^{(d)}}(m)$ the corresponding shatter coefficients (see Subsection 1.2) and, with a slight abuse of notation, by $\mathbb{S}_{\mathbf{C}_n}^{(J)}(m)$ the shatter coefficient corresponding to the collection $\cup_{d=1}^{2^J} \mathbf{D}_n^{(d)}$ of all rules embedded in $\mathbb{R}^{2^J}$. Observe that

$$\mathbb{S}_{\mathbf{C}_n}^{(J)}(m) \leq \sum_{d=1}^{2^J} \mathbb{S}_{\mathbf{C}_n^{(d)}}(m).$$

Second, we let the n training data reorder the first 2^J basis functions $\{\psi_1, \ldots, \psi_{2^J}\}$ into $\left\{\psi_{j_1}, \ldots, \psi_{j_{2^J}}\right\}$ via the scheme

$$\sum_{i=1}^{n} X_{ij_1}^2 \geq \sum_{i=1}^{n} X_{ij_2}^2 \geq \ldots \geq \sum_{i=1}^{n} X_{ij_{2^J}}^2. \tag{2.4}$$

In other words, we just let the training sample decide by itself which basis functions carry the most significant information.

We finish the procedure by a **third** selection step: pick the *effective* dimension $d \leq 2^J$ and a classification rule $g^{(d)}$ in $\mathbf{D}_n^{(d)}$ by approximating each X_i by $\mathbf{X}_i^{(d)} = (X_{ij_1}, \ldots, X_{ij_d})$.

The dimension d and the classifier $g^{(d)}$ are simultaneously selected using the data-splitting device described in Subsection 1.2. Precisely, we choose both d and $g^{(d)}$ optimally by minimizing the empirical probability of error based on the independent validation set, that is

$$\left(\hat{d}, \hat{g}^{(\hat{d})}\right) \in \underset{d=1,\ldots,2^J, g \in \mathbf{D}_n^{(d)}}{\operatorname{argmin}} \left[\frac{1}{m} \sum_{i=n+1}^{n+m} \mathbf{1}_{[g^{(d)}(\mathbf{X}_i^{(d)}) \neq Y_i]}\right]. \tag{2.5}$$

Note that the second step of our algorithm is somewhat related to wavelet shrinkage, that is, certain wavelet coefficients are reduced to zero. Wavelet shrinkage and thresholding methods constitute a powerful way to carry out signal analysis, especially when the underlying process has sparse wavelet representation. They are computationally fast and automatically adapt to spatial and frequency inhomogeneities of the signal. A review of the advantages of wavelet shrinkage appears in Donoho, Johnstone, Kerkyacharian, and Picard [17]. In our functional classification context, the preprocessing step (2.4) allows to shrink *globally* all learning data. This point is crucial,

as individual shrinkages would lead to different significant bases for each function in the data set.

Apart from being conceptually simple, this method leads to the classifier $\hat{g}(\mathbf{x}) = \hat{g}^{(\hat{d})}(\mathbf{x}^{(\hat{d})})$ with a probability of misclassification

$$L_{n+m}(\hat{g}) = \mathbf{P}\{\hat{g}(\mathbf{X}) \neq Y \mid (\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_{n+m}, Y_{n+m})\},$$

where, for a generic X, $\mathbf{X}^{(d)} = (\mathbf{X}_{j_1}, \ldots, \mathbf{X}_{j_d})$ denotes the first d coefficients reordered via the scheme (2.4). The selected rule $\hat{g}$ satisfies the following optimal inequality, whose proof is clear from (1.1) and (1.3):

THEOREM 2.1.

$$\mathbf{E}\{L_{n+m}(\hat{g})\} - L^* \leq L^*_{2^J} - L^* + \mathbf{E}\Big\{ \inf_{d=1,\ldots,2^J, g^{(d)} \in \mathbf{D}_n^{(d)}} L_n(g^{(d)}) \Big\} - L^*_{2^J}$$
$$+ 2\mathbf{E}\left\{ \sqrt{\frac{8 \log\left(4\mathbb{S}^{(J)}_{\mathbf{C}_n}(2m)\right)}{m}} + \frac{1}{\sqrt{(m/2)\log\left(4\mathbb{S}^{(J)}_{\mathbf{C}_n}(2m)\right)}} \right\}.$$

Here

$$L^*_{2^J} = \inf_{s:\mathbb{R}^{2^J} \to \{0,1\}} \mathbf{P}\{s(\mathbf{X}^{(2^J)}) \neq Y\}$$

stands for the Bayes probability of error when the feature space is $\mathbb{R}^{2^J}$.

We may view the first term, $L^*_{2^J} - L^*$, on the right of the inequality as an approximation term – the price to pay for using a finite-dimensional approximation. This term converges to zero by Lemma 2.1 below, which is a special case of Theorem 32.3, page 567 in Devroye, Györfi, and Lugosi [16].

LEMMA 2.1. *We have*

$$L^*_{2^J} - L^* \to 0 \quad \text{as } J \to \infty.$$

The second term, $\mathbf{E}\{\inf_{d=1,\ldots,2^J, g^{(d)} \in \mathbf{D}_n^{(d)}} L_n(g^{(d)})\} - L^*_{2^J}$, can be handled by standard results on classification. Let us first recall the definition of a *consistent* rule: a rule g is consistent for a class of distributions $\mathcal{D}$ if $\mathbf{E}\{L_n(g)\} \to L^*$ as $n \to \infty$ for all distributions $(X, Y) \in \mathcal{D}$.

COROLLARY 2.1. *Let $\mathcal{D}$ be a class of distributions. For fixed $J \geq 0$, assume that we can pick from each $\mathbf{D}_n^{(2^J)}$, $n \geq 1$, one $g_n^{(2^J)}$ such that the*

sequence $(g_n^{(2^J)})_{n\geq 1}$ *is consistent for* $\mathcal{D}$. *If*

$$\lim_{n\to\infty} m = \infty, \quad \textit{and, for each } J, \quad \lim_{n\to\infty} \mathbf{E}\left\{\frac{\log \mathbb{S}_{\mathbf{C}_n}^{(J)}(2m)}{m}\right\} = 0,$$

then the automatic rule $\hat{g}$ *defined in (2.5) is consistent for* $\mathcal{D}$ *in the sense*

$$\lim_{J\to\infty} \lim_{n\to\infty} \mathbf{E}\{L_{n+m}(\hat{g})\} = L^*.$$

Proof The proof uses Theorem 2.1, Lemma 2.1, and the upper bound

$$\mathbf{E}\left\{\inf_{d=1,\ldots,2^J, g^{(d)}\in\mathbf{D}_n^{(d)}} L_n(g^{(d)})\right\} - L_{2^J}^* \leq \mathbf{E}\{L_n(g_n^{(2^J)})\} - L_{2^J}^*.$$

■

This consistency result is especially valuable since few approximation results have been established for functional classification. Corollary 2.1 shows that a consistent rule is selected if, for each fixed $J \geq 0$, the sequence of $\mathbf{D}_n^{(2^J)}$'s contains a consistent rule, even if we do not know which functions from $\mathbf{D}_n^{(2^J)}$ lead to consistency. If we are only concerned with consistency, Corollary 2.1 reassures us that nothing is lost as long as we take m much larger than $\log \mathbf{E}\{\mathbb{S}_{\mathbf{C}_n^{(J)}}(2m)\}$. Often, this reduces to a very weak condition on the size m of the validation set. Note also that it is usually possible to find upper bounds on the random variable $\mathbb{S}_{\mathbf{C}_n^{(J)}}(2m)$ that depend on n, m and J, but not on the actual values of the random variables $(X_1, Y_1), \ldots, (X_n, Y_n)$. In this case, the bound is distribution-free, and the problem is purely combinatorial: count $\mathbb{S}_{\mathbf{C}_n^{(J)}}(2m)$. Examples are now presented.

Example 1: k-NN rules. In the k-nearest neighbor rule (k-NN), a majority vote decision is made over the labels based upon the k nearest neighbors of x in the training set. This procedure is among the most popular nonparametric methods used in statistical pattern recognition with over 900 research articles published on the method since 1981 alone! Dasarathy [13] has provided a comprehensive collection of around 140 key papers.

If $\mathbf{D}_n^{(d)}$ contains all NN-rules (all values of k) in dimension d, then $\mathbf{D}_n^{(d)}$ increases with n and depends very much on the training set. A trivial bound in this case is

$$\mathbb{S}_{\mathbf{C}_n^{(d)}}(2m) \leq n$$

because there are only n members in $\mathbf{D}_n^{(d)}$. Consequently,

$$\mathbb{S}_{\mathbf{C}_n}^{(J)}(2m) \leq 2^J n.$$

Stone [33] proved the striking result that k-NN classifiers are universally consistent if $X \in \mathbb{R}^d$, provided $k \to \infty$ and $k/n \to 0$. Therefore, we see that our strategy leads to a consistent rule whenever $J/m \to 0$ and $\log n/m \to 0$ as $n \to \infty$. Thus, we can take m equal to a small fraction of n without loosing consistency. Consistent classifiers can also be obtained by other local averaging methods as long as $\mathcal{F} = \mathbb{R}^d$, see e.g. Devroye, Györfi, and Lugosi [16]. On the other hand, the story is radically different in general spaces $\mathcal{F}$. Abraham, Biau, and Cadre [1] present counterexamples indicating that the moving window rule (Devroye, Györfi, and Lugosi [16], Chapter 10) is not consistent for general $\mathcal{F}$, and they argue that restrictions on the space $\mathcal{F}$ (in terms of metric covering numbers) and on the regression function $\eta(x) = \mathbb{E}\{Y|X = x\}$ cannot be given up. By adapting the arguments in Abraham, Biau, and Cadre [1], it can be shown that the k-NN classifier is consistent, provided η is continuous on the separable Hilbert space $L_2([0,1])$, $k \to \infty$ and $k/n \to 0$ (see Cérou and Guyader [10]).

Example 2: Binary tree classifiers. Classification trees partition $\mathbb{R}^d$ into regions, often hyperrectangles parallel to the axes. Among these, the most important are the binary classification trees, since they have just two children per node and are thus easiest to manipulate and update. Many strategies have been proposed for constructing the binary decision tree (in which each internal node corresponds to a cut, and each terminal node corresponds to a set in the partition). For examples and list of references, we refer the reader to Devroye, Györfi, and Lugosi [16], Chapter 20.

If we consider for example all binary trees in which each internal node corresponds to a split perpendicular to one of the axes, then

$$\mathbb{S}_{\mathbf{C}_n^{(d)}}(2m) \leq \big(1 + d(n+2m)\big)^k,$$

where k is the maximum number of consecutive orthogonal cuts (or internal nodes). Therefore,

$$\mathbb{S}_{\mathbf{C}_n}^{(J)}(2m) \leq \sum_{d=1}^{2^J} \big(1 + d(n+2m)\big)^k \leq 2^J \big(1 + 2^J(n+2m)\big)^k.$$

3. Applications. In this section, we propose to illustrate the performance of the wavelet classification algorithm presented in Section 2. The method has been tested using:

- The six wavelets bases generated by Daubechies' mother wavelets depicted in Figure 1. For $p = 1, 2, 4, 6, 8, 10$, we denote by daub$p$ the

wavelet basis generated by the mother wavelet with p vanishing moments.

- Three collections of rules $\mathbf{D}_n^{(d)}$ performing in the finite-dimensional space $\mathbb{R}^d$. We will use the following acronyms:
  - → W-NN when $\mathbf{D}_n^{(d)}$ consists of all d-dimensional nearest-neighbor classifiers.
  - → W-QDA when $\mathbf{D}_n^{(d)}$ consists of the Quadratic Discriminant Analysis rule (Devroye, Györfi, and Lugosi [16], Chapter 13) performed in dimension d. We only considered dimensions d which do not exceed the minimum number of training data in each group.
  - → W-CART when $\mathbf{D}_n^{(d)}$ corresponds to the classification and regression tree procedure of Breiman, Friedman, Olshen, and Stone [8].

In addition, our functional classification methodology is compared with four alternative approaches:

- F-NN refers to the Fourier filtering approach combined with the k-NN rule studied in Biau, Bunea, and Wegkamp [4]. In this method, the k-NN discrimination rule is performed on the first d coefficients of a Fourier series expansion of each curve. The effective dimension d and the number of neighbors k are selected by minimizing the empirical probability of error based on the validation sequence plus an additive penalty term $\lambda_d/\sqrt{m}$ which avoids overfitting. We choose the penalty term as suggested by the authors, namely $\lambda_d = 0$ for $d \leq n$ and $\lambda_d = \infty$ for $d > n$.
- NN-Direct denotes the k-nearest neigbor rule directly applied to the observations $X_1, \ldots, X_n$ without any preliminary dimension reduction step. As for the Fourier method described above, the optimal number of neighbors is selected using data-splitting and empirical risk.
- MPLSR refers to the Multivariate Partial Least Square Regression for functional classification. This approach is studied in detail in Preda and Saporta [28] and in Preda, Saporta, and Lévéder [29]. The number of PLS components is selected by minimizing the empirical probability of error based on the validation sequence.
- RF corresponds to the Random Forest algorithm of Breiman [7]. A random forest is a collection of tree predictors, where each tree is constructed from a bootstrap sample drawn with replacement from the training data. Instead of determining the optimal split over all possible splits on all covariates, a subset of the covariates, drawn at random, is used.

For the free parameters of W-CART and RF, we used the default values of the R-packages *tree* and *randomForest* (these packages are available at the url http://lib.stat.cmu.edu/R/CRAN/). Our codes are available by request.

3.1. *Speech recognition.* We first tested the different methods on a speech recognition problem. We study a part of TIMIT database which was investigated in Hastie, Buja, and Tibshirani [24]. The data are log-periodograms corresponding to recording phonemes of 32 ms duration. We are concerned with the discrimination of five speech frames corresponding to five phonemes transcribed as follows: "sh" as in "she" (872 items), "dcl" as in "dark" (757 items), "iy" as the vowel in "she" (1163 items), "aa" as the vowel in "dark" (695 items) and "a0" as the first vowel in "water" (1022 items). The database is a multispeaker database. Each speaker is recorded at a 16 kHz sampling rate and we retain only the first 256 frequencies (see Figure 2). Thus the data consists of 4509 series of length 256 with known class word membership.

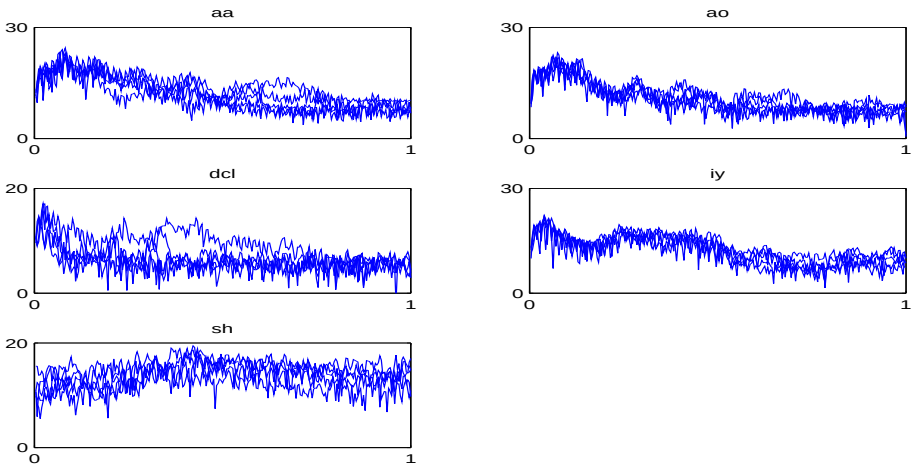

FIG 2. *A sample of 5 log-periodograms, one in each class.*

We decided to retain 250 observations for training and 250 observations for validation. Since the curves are sampled at $256 = 2^8$ equidistant points, we fix the maximum resolution level J at 8. The error rate (**ER**) of the elected rule $\hat{g}$ for classifying new observations is unknown, but it can be estimated consistently using the rest of the data $(X_{501}, Y_{501}), \ldots, (X_{4509}, Y_{4509})$, via the formula

$$\mathbf{ER} = \frac{1}{4009} \sum_{i=501}^{4509} \mathbf{1}_{[\hat{g}(X_i) \neq Y_i]}.$$

Table 1 and Table 2 display the estimated error rates for the different methods. All results are averaged over 100 random partitions of the data. Figure 3 shows the boxplots of the selected dimensions for wavelet and Fourier algorithms.

Method \ Basis	daub1	daub2	daub4	daub6	daub8	daub10
W-NN	0.111	0.110	0.112	0.114	0.113	0.112
W-QDA	0.097	0.102	0.108	0.108	0.113	0.115
W-CART	0.112	0.130	0.162	0.159	0.163	0.185

TABLE 1
Estimated error rates for wavelet filtering methods.

Method	ER
F-NN	0.137
NN-Direct	0.113
MPLSR	0.091
RF	0.096

TABLE 2
Estimated error rates for other methods.

Table 1 and Table 2 support the idea that the three methods using wavelets perform well on the present data and are robust with respect to the choice of bases. The methods MPLSR and RF are competitive procedures when compared to the others, and the NN-Direct algorithm (directly applied to the discretized functions, in $\mathbb{R}^{256}$) performs as well as the W-NN algorithm. The results of the Fourier-based procedure are still acceptable. Thus, for this data, the wavelet-based methods do not significantly outperform the other methods. However, the performance of the other methods can considerably deteriorate for time/frequency inhomogeneous signals, as illustrated in the next subsection. We note finally that Figure 3 exhibits further evidence that our wavelet approach allows a more significant dimension reduction than the Fourier filtering approach of Biau, Bunea, and Wegkamp [4].

3.2. *A simulation study.* We propose to investigate the performance of our method in the following simulated scenario. For each $i = 1, \ldots, n$, we generate pairs $(X_i(t), Y_i)$ via the scheme:

$$X_i(t) = \sin(F_i^1 \pi t) f_{\mu_i,\sigma_i}(t) + \sin(F_i^2 \pi t) f_{1-\mu_i,\sigma_i}(t) + \varepsilon_t,$$

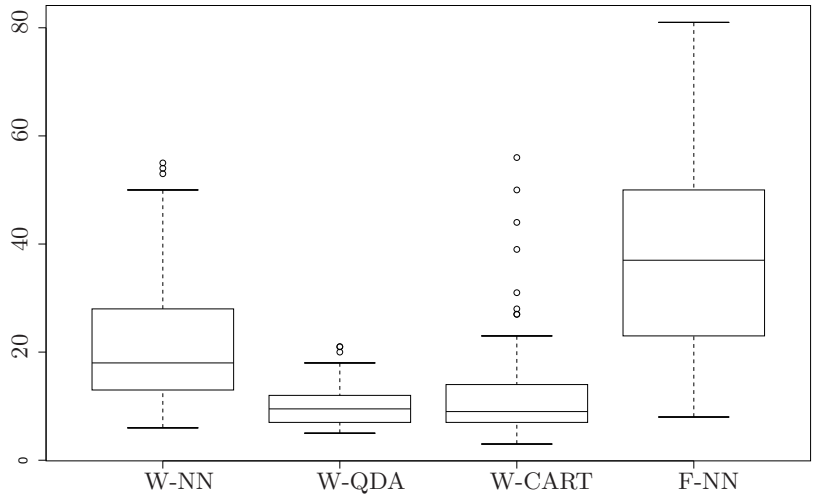

FIG 3. *Boxplots of the selected dimensions for wavelet (W-NN, W-QDA, W-CART) and Fourier methods (F-NN). The wavelet basis is daub4.*

where

- $f_{\mu,\sigma}$ stands for the normal density with mean μ and variance σ^2;
- F_i^1 and F_i^2 are uniform random variables on $[50, 150]$;
- μ_i is randomly uniform on $[0.1, 0.4]$;
- σ_i^2 is randomly uniform on $[0, 0.005]$;
- the ε_t's are mutually independent normal random variables with mean 0 and standard deviation 0.5.

The label Y_i associated to X_i is then defined, for $i = 1, \ldots, n$, by

$$Y_i = \begin{cases} 0 & \text{if } \mu_i \leq 0.25 \\ 1 & \text{otherwise.} \end{cases}$$

Figure 4 displays six typical realizations of the X_i's. We see that each curve $X_i(t)$, $t \in [0, 1]$, is composed of two different but symmetric signals, and the problem is thus to detect if the two signals are close (label 0) or enough distant (label 1). Curves are sampled at $1024 = 2^{10}$ equidistant points, and we choose therefore $J = 10$ for the maximum resolution level.

All the algorithms were tested over samples of size 50 for training and 50 for validation. The error rates (**ER**) were estimated on independent samples of size 500. They are reported on Table 3 and Table 4. All results are averaged over 100 repetitions.

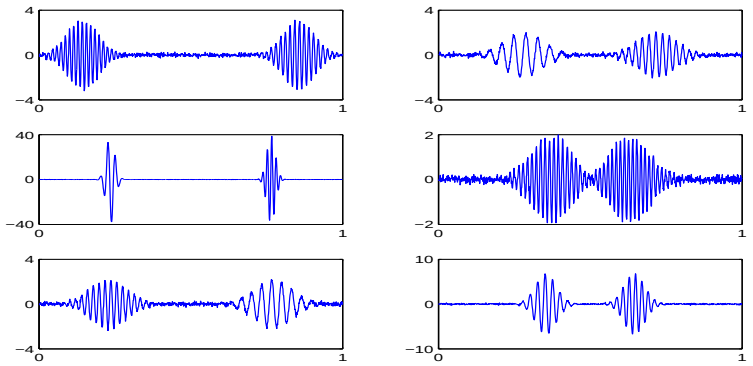

FIG 4. *Six typical realizations of simulated curves with label* 1 *(left) and label* 0 *(right).*

Method \ Basis	daub1	daub2	daub4	daub6	daub8	daub10
W-NN	0.146	0.143	0.146	0.156	0.155	0.159
W-QDA	0.078	0.082	0.085	0.082	0.085	0.084
W-CART	0.185	0.174	0.170	0.177	0.179	0.161

TABLE 3
Estimated error rates for wavelet filtering methods.

Table 3 and Table 4 further emphasize the good results achieved by the wavelet classification algorithms, and their robustness with respect to the choice of bases. We note in particular the excellent performance of W-QDA, which achieves, on average, the best error rates, together with the RF algorithm. On the other hand, we note that the choice of the method performed on the wavelet coefficients is crucial, as W-QDA clearly outperforms W-NN and W-CART. The rather poor results obtained by the F-NN method are not surprising. In effect, due to the penalty term ($\lambda_d = 0$ for $d \leq n$ and $\lambda_d = \infty$ for $d > n$), this procedure retains only the first n coefficients of the Fourier expansion. This maximal number of coefficients is definitely too low here since frequencies of the two signals can typically approach 150 Hz. The problem of the calibration of the penalty term is discussed in detail in Biau, Bunea, and Wegkamp [4] and Fromont and Tuleau [21].

To illustrate the importance of the wavelet shrinkage approach, we ran all the wavelet methods without reordering the 2^J basis functions. Table 5 summarizes the results, Figure 5 displays boxplots of the estimated error

Method	ER
F-NN	0.212
NN-Direct	0.182
MPLSR	0.483
RF	0.060

TABLE 4
Estimated error rates for other methods.

rates, and Figure 6 shows boxplots of the selected dimensions.

Method \ Basis	daub1	daub2	daub4	daub6	daub8	daub10
W-NN	0.170	0.189	0.185	0.193	0.192	0.190
W-QDA	0.066	0.104	0.288	0.406	0.455	0.467
W-CART	0.300	0.348	0.446	0.465	0.475	0.485

TABLE 5
Estimated error rates for wavelet filtering methods without reordering the basis functions.

Table 5, Figure 5 and Figure 6 illustrate the clear advantages of reordering the data, as shown by the error rates as well as by the dimension reduction. We note finally that the performance of the approach without basis reordering is not robust with respect to the choice of basis. In effect, the estimated error rate of W-QDA increases from 0.06 when the method is performed with the daub1 basis to 0.467 when it is performed with the daub10 basis. This drawback can clearly be avoided by the reordering strategy, as illustrated in Table 3. In practice, when one has no or little a priori information to support a particular choice of wavelet basis, the automatic selection approach discussed in the present paper is thus preferable.

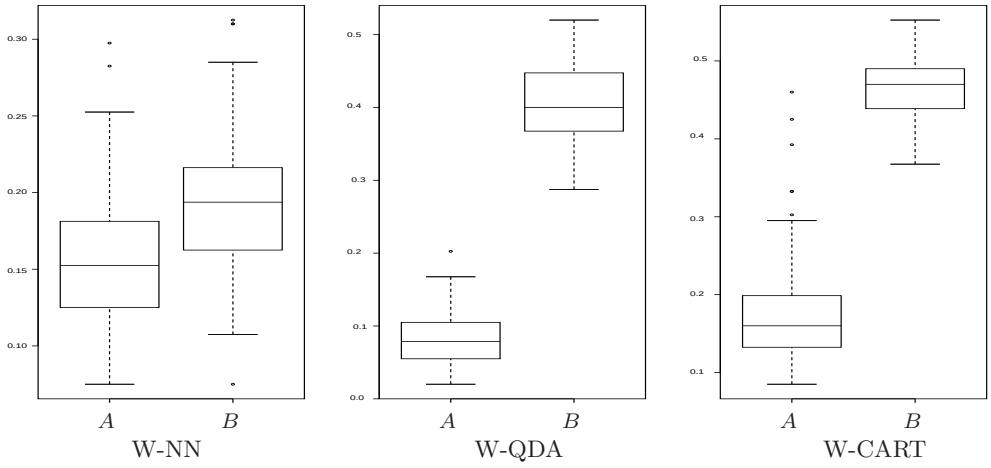

FIG 5. *Boxplots of the estimated error rates. A: wavelet filtering with reordering of the basis functions; B: wavelet filtering methods without reordering of the basis functions. The wavelet basis is daub6.*

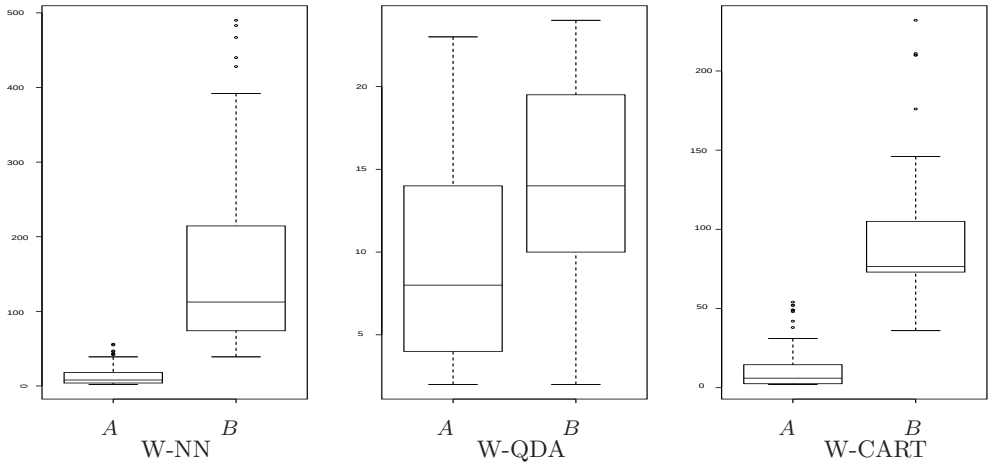

FIG 6. *Boxplots of the selected dimensions. A: wavelet filtering methods with reordering of the basis functions; B: wavelet filtering methods without reordering of the basis functions. The wavelet basis is daub6.*

Acknowledgements. The authors greatly thank an anonymous referee for his comments and suggestions and for having pointed out additional references.

References.

[1] C. Abraham, G. Biau, and B. Cadre. On the kernel rule for function classification. *Annals of the Institute of Statistical Mathematics*, 58:619–633, 2006.

[2] U. Amato, A. Antoniadis, and I. De Feis. Dimension reduction in functional regression with applications. *Computational Statistics and Data Analysis*, 50:2422–2446, 2006.

[3] P. N. Belhumeur, J. P. Hepana, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:711–720, 1997.

[4] G. Biau, F. Bunea, and M. H. Wegkamp. Functional classification in Hilbert spaces. *IEEE Transactions on Information Theory*, 51:2163–2172, 2005.

[5] S. Boucheron, O. Bousquet, and G. Lugosi. *Advanced Lectures in Machine Learning*, chapter Introduction to Statistical Learning Theory, pages 169–207. Springer, Heidelberg, 2004.

[6] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

[7] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[8] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Belmont, Wadsworth, 1984.

[9] H. Cardot and P. Sarda. Estimation in generalized linear models for functional data via penalized likelihood. *Journal of Multivariate Analysis*, pages 24–41, 2005.

[10] F. Cérou and A. Guyader. Nearest neighbor classification in infinite dimension. *ESAIM: P&S*, 10:340–355, 2006.

[11] P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36: 287–314, 1994.

[12] A. Cuevas, M. Febrero, and R. Fraiman. Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics*, 22: 481–496, 2007.

[13] B. V. Dasarathy. *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*. IEEE Computer Society Press, Los Alamitos, 1991.

[14] I. Daubechies. *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, Philadelphia, 1992.

[15] L. Devroye. Automatic pattern recognition: A study of the probability of error. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10:530–543, 1988.

[16] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New-York, 1996.

[17] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: Asymptopia? With discussion and a reply by the authors. *Journal of the Royal Statistical Society Series B*, 57:545–564, 2002.

[18] F. Ferraty and P. Vieu. The functional nonparametric model and application to spectrometric data. *Computational Statistics*, 17:545–564, 2002.

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

[20] F. Ferraty and P. Vieu. *Nonparametric Functional Data Analysis. Theory and Practice*. Springer, New York, 2006.

[21] M. Fromont and C. Tuleau. Functional classification with margin conditions. In *Lecture Notes in Computer Science, Learning Theory*, volume 4005, pages 94–108. COLT, 2006.

[22] A. Haar. Zur Theorie der orthogonalen Funktionensysteme. *Mathematische Annalen*, 69:331–371, 1910.

[23] P. Hall, D. S. Poskitt, and B. Presnell. A functional data-analytic approach to signal discrimination. *Technometrics*, 43:1–9, 2001.
[24] T. Hastie, A. Buja, and R. Tibshirani. Penalized discriminant analysis. *The Annals of Statistics*, 23:73–102, 1995.
[25] M. Kirby and L. Sirovich. Application of the Karhunen-Loeve procedure for the characterization of human faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12:103–108, 1990.
[26] S. R. Kulkarni and S. E. Posner. Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Transactions on Information Theory*, 41:1028–1039, 1995.
[27] Y. Meyer. *Wavelet and Operators*. Cambridge University Press, Cambridge, 1992.
[28] C. Preda and G. Saporta. PLS regression on a stochastic process. *Computational Statistics and Data Analysis*, 48:149–158, 2005.
[29] C. Preda, G. Saporta, and C. Lévéder. PLS classification of functional data. *Computational Statistics*, 22:223–235, 2007.
[30] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer-Verlag, New York, 1997.
[31] J. O. Ramsay and B. W. Silverman. *Applied Functional Data Analysis. Methods and Case Studies*. Springer-Verlag, New York, 2002.
[32] F. Rossi and N. Villa. Support vector machine for functional data classification. *Neurocomputing*, 69:730–742, 2006.
[33] C. J. Stone. Consistent nonparametric regression. *The Annals of Statistics*, 5:595–645, 1977.
[34] V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280, 1971.

Alain Berlinet
Institut de Mathématiques
et de Modélisation
de Montpellier, UMR CNRS 5149
Université Montpellier II
Place Eugène Bataillon
34095 Montpellier Cedex 5, France
E-mail: berlinet@math.univ-montp2.fr

Gérard Biau
Laboratoire de Statistique
Théorique et Appliquée
Université Pierre et Marie Curie
Paris VI
Boîte 158, 175 rue du Chevaleret
75013 Paris, France
E-mail: biau@ccr.jussieu.fr
URL: http://www.lsta.upmc.fr/biau.html

Laurent Rouvière
Institut de Recherche Mathématique de Rennes
UMR CNRS 6625
Université Rennes II-Haute Bretagne
Campus Villejean
Place du Recteur Henri Le Moal, CS 24307
35043 Rennes Cedex, France
E-mail: laurent.rouviere@univ-rennes2.fr
URL: http://www.uhb.fr/sc_sociales/labstats/ROUVIERE