FUNCTIONAL CLASSIFICATION IN HILBERT SPACES

GÉRARD BIAU†, FLORENTINA BUNEA‡, AND MARTEN H. WEGKAMP‡

Abstract. Let X be a random variable taking values in a separable Hilbert space $\mathcal{X}$, with label $Y \in \{0, 1\}$. We establish universal weak consistency of a nearest neighbor-type classifier based on n independent copies (X_i, Y_i) of the pair (X, Y), extending the classical result of Stone (1977) to infinite dimensional Hilbert spaces. Under a mild condition on the distribution of X, we also prove strong consistency. We reduce the infinite dimension of $\mathcal{X}$ by considering only the first d coefficients of a Fourier series expansion of each X_i, and then we perform k-nearest neighbor classification in $\mathbb{R}^d$. Both the dimension and the number of neighbors are automatically selected from the data using a simple data-splitting device. An application of this technique to a signal discrimination problem involving speech recordings is presented.

1. Introduction

Classical pattern recognition deals with predicting the unknown nature Y, called a class or label, of an observation X in $\mathbb{R}^d$. Assume for simplicity that the class Y only takes two values, say 0 or 1. The statistician creates a classifier $\phi : \mathbb{R}^d \to \{0, 1\}$ which represents her guess on the label Y of X. Since it is not assumed that the covariate X fully determines the label – the same covariate x may give rise to different labels – it is certainly possible to misspecify its associated label. We err if $\phi(X)$ differs from Y, and the probability of error for a particular classifier ϕ is denoted by $L(\phi) =$

Date: April 13, 2005.

1991 *Mathematics Subject Classification.* Primary 62G10; secondary 62G05.

Key words and phrases. Classification, nearest neighbor rule, Fourier expansion, universal consistency.

†Corresponding author. Email: `biau@math.univ-montp2.fr`. Institut de Mathématiques et de Modélisation de Montpellier – UMR CNRS 5149, Equipe de Probabilités et Statistique, Université Montpellier II, CC 051 – Place Eugène Bataillon, 34095 Montpellier Cedex 5, France.

‡Research partially supported by NSF grant DMS 0406049. Department of Statistics, Florida State University, Tallahassee, FL 32306-4330, United States of America.

$\mathbb{P}\{\phi(X) \neq Y\}$. The Bayes classifier

$$\phi^*(x) \quad = \quad \begin{cases} 0 & \text{if } \mathbb{P}\{Y = 0|X = x\} \geq \mathbb{P}\{Y = 1|X = x\} \\ 1 & \text{otherwise} \end{cases}$$

has the smallest probability of error, that is

$$L^* = L(\phi^*) = \inf_{\phi:\mathbb{R}^d \to \{0,1\}} \mathbb{P}\{\phi(X) \neq Y\}, \tag{1}$$

cf. Devroye, Györfi, and Lugosi (1996, Theorem 2.1, page 10). Because this classifier depends on the unknown distribution of (X, Y), it cannot be computed on the basis of the data alone. Thus, the classification problem is to construct classifiers ϕ_n based on independent random observations $(X_1, Y_1), \ldots, (X_n, Y_n)$, with the same distribution as (X, Y). One can then assess the performance of a given classification scheme against the Bayes classifier. A classifier ϕ_n is called consistent if it achieves the Bayes error probability L^* in the limit, as $n \to \infty$, that is,

$$\lim_{n\to\infty} \mathbb{E}L(\phi_n) = L^*,$$

where $L(\phi_n) = \mathbb{P}\big\{\phi_n(X) \neq Y|(X_1, Y_1), \ldots, (X_n, Y_n)\big\}$ is the probability of error, conditionally on the data used to construct ϕ_n.

In this paper we generalize the problem of classifying elements of $\mathbb{R}^d$ to classifying random variables X_i that take values in an infinite dimensional separable Hilbert space $\mathcal{X}$; we consider here spaces of functions. We refer to this problem as functional classification. References to functional classification in the mainstream statistical literature are limited: Kulkarni and Posner (1995) study rates of convergence of k-nearest neighbor regression estimates in general spaces; Hall, Poskitt, and Presnell (2001) employ a functional data-analytic method for dimension reduction based on Principal Component Analysis (PCA) and perform Quadratic Discriminant Analysis (QDA) on the reduced space, and Ramsay and Silverman (1997, 2002) discuss similar techniques; Hastie, Buja, and Tibshirani (1995) set out the general idea of Functional Discriminant Analysis (FDA) making use of a roughness penalty approach to regularization; and Ferraty and Vieu (2002) estimate nonparametrically the posterior probability of an incoming curve in a given class. Cover and Hart (1967) study nearest neighbor classification of Banach-valued elements, but they do not establish consistency.

The k-nearest neighbor method of classification is central to our paper. This technique consists in constructing classifiers $\phi_n(X)$ by taking the majority vote over the labels of the k-nearest neighbors X_i of X [cf. Fix and Hodges (1951, 1952, 1991a, 1991b)]. 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 (1991) has provided a comprehensive collection of around 140 key papers. Stone (1977) proved the striking result that k-nearest neighbor classifiers are universally consistent if $X \in \mathcal{X} = \mathbb{R}^d$, provided $k \to \infty$ and $k/n \to 0$. Here universally consistent means that these rules are consistent, regardless of the underlying distribution of (X, Y). Universally consistent classifiers can also be obtained by other local averaging methods as long as $\mathcal{X} = \mathbb{R}^d$, see e.g. Devroye, Györfi, and Lugosi (1996). On the other hand, the story is radically different in general spaces $\mathcal{X}$ [cf. Kulkarni and Posner (1995), Diabo-Niang and Rhomari (2001), and Abraham, Biau, and Cadre (2003)]. Abraham, Biau, and Cadre (2003) presented counterexamples indicating that the moving window rule [Devroye, Györfi, and Lugosi (1996, Chapter 10)] is not universally consistent for general $\mathcal{X}$, and they argue that restrictions on the space $\mathcal{X}$ (in terms of metric covering numbers) and on the regression function $\eta(x) = \mathbb{E}[Y|X = x]$ cannot be dispensed with. By adapting the arguments in Abraham, Biau, and Cadre (2003), it can be shown that the k-nearest neighbor classifier is consistent, provided η is continuous on the separable Hilbert space $\mathcal{X}$, $k \to \infty$ and $k/n \to 0$. Hence at the price of requiring some regularity on η, we obtain consistency. This (direct) approach suffers from the curse of dimensionality as already pointed out by e.g. Ramsay and Silverman (2002, page 129), and we expect slow rates of convergence in general.

There are various ways to proceed in functional classification that try to deal with the curse of dimensionality. Ferraty, Peuch, and Vieu (2003) propose a single functional index method. Hastie, Buja, and Tibshirani (1995) suggest penalized discriminant analysis, which they illustrate with examples in speech recognition and handwritten character recognition.

Filtering is another popular dimension reduction method in signal and speech processing, and this is the central approach we take in this paper. Filtering reduces the infinite dimension of the space of functions by considering only the first d coefficients of a Fourier series expansion of each function; see, e.g. Kirby and Sirovich (1990), Comon (1994), Belhumeur, Hepana, and Kriegman (1997), and Hall, Poskitt, and Presnell (2001). Given a collection of functions we wish to classify, we suggest the following procedure: first use filtering on each function and then perform k-nearest neighbor classification in $\mathbb{R}^d$. We propose a data-driven procedure that selects simultaneously both the dimension d and the optimal number of neighbors k. We found that this method is easily implementable, works fast, and exhibits excellent finite sample behavior. Our main theoretical result is that the procedure yields a universally consistent classifier, thereby extending Stone's classical result from $\mathbb{R}^d$ to general separable Hilbert spaces.

The article is organized as follows. In Section 2, we outline the method and state consistency of our classification rule. In Section 3 we present an application of our method to speech recognition. Section 4 explores, via a simulation study, the small sample properties of the classifier. All proofs are collected in Section 5.

2. Consistent functional classification

In this section we present the construction of our classifier and show that it is universally consistent. The proofs are collected in Section 5.

We begin by introducing the theoretical framework. Let X be a random variable taking values in an infinite dimensional, separable Hilbert space $\mathcal{X}$ with inner product $< \cdot, \cdot >$ and norm $\|.\|$. Let Y be the random variable with values 0 and 1 which represents the label associated with X. The distribution of the pair (X, Y) is completely specified by the probability measure of X, and by $\eta(x) = \mathbb{P}\{Y = 1|X = x\}$, the regression function of Y on X. The data consist of $(X_1, Y_1), \ldots, (X_n, Y_n)$, assumed to be independent copies of (X, Y). The goal is to construct a consistent classifier based on this data.

In deriving both our method and our consistency result, the assumption that $\mathcal{X}$ is separable is important. Note that in this case each element X_i may be expressed as a series expansion

$$X_i = \sum_{j=1}^{\infty} X_{ij}\psi_j, \tag{2}$$

where $\{\psi_j\}_{j=1}^{\infty}$ form a complete, orthonormal system of $\mathcal{X}$ and hence the random coefficients $X_{ij} =< X_i, \psi_j >$. Let $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots)$ be the coefficients associated with X_i. Recall that any infinite dimensional, separable Hilbert space $\mathcal{X}$ is isomorphic with $\ell_2 = \big\{\mathbf{x} = (x_1, x_2, \ldots) : \sum_{j=1}^{\infty} x_j^2 < +\infty\big\}$. Consequently, knowing X_i is the same as knowing $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots)$. Thus, recalling identity (1), the consistency of the classifier we construct below will be established relative to

$$L^* = \inf_{\phi:\mathcal{X}\to\{0,1\}} \mathbb{P}\{\phi(X) \neq Y\} = \inf_{\phi:\ell_2\to\{0,1\}} \mathbb{P}\{\phi(\mathbf{X}) \neq Y\}.$$

In Sections 3 and 4 we consider $\mathcal{X} = L_2\big([0,1]\big)$ and the trigonometric basis

$$\psi_1(t) = 1, \quad \psi_{2j}(t) = \sqrt{2}\cos(2\pi jt)\,, \quad \psi_{2j+1}(t) = \sqrt{2}\sin(2\pi jt)\,, \quad j = 1, 2, \ldots,$$

which is a complete orthonormal system in $L_2\big([0,1]\big)$ [cf. Szegö (1959), Zygmund (1959), and Sansone (1969)].

We suggest the following procedure:

- Select the *effective* dimension d to approximate each X_i by the sum $\sum_{j=1}^{d} X_{ij}\psi_j$.

- Perform nearest neighbor classification based on the coefficients $(X_{i1}, \ldots, X_{id}) \in \mathbb{R}^d$, $1 \leq i \leq n$, for the selected dimension d.

We select the dimension d and the number of neighbors k simultaneously, using a simple data-splitting device. The data are split into a training set $\{(X_i, Y_i), i \in \mathcal{I}_\ell\}$ of length ℓ and a validation set $\{(X_j, Y_j), j \in \mathcal{J}_m\}$ of length m such that $n = \ell + m$ (ℓ and m possibly functions of n, and $1 \leq \ell \leq n-1$). For each $d \geq 1$, $1 \leq k \leq \ell$ and $x \in \mathbb{R}^d$, let $\phi_{\ell,k,d}(x)$ be the k-nearest neighbor rule based on the training set. We define it as follows. Let $\mathbf{X}_i^{(d)} = (X_{i1}, \ldots, X_{id})$ be the first d coefficients in expansion (2) of X_i, $i \in \mathcal{I}_\ell$ (similarly, we will denote by $\mathbf{X}^{(d)}$ the first d coefficient vector in expansion (2) of a generic X). Then, we reorder the training data

$$\big(\mathbf{X}_{(1)}^{(d)}(x), Y_{(1)}(x)\big), \ldots, \big(\mathbf{X}_{(\ell)}^{(d)}(x), Y_{(\ell)}(x)\big)$$

according to increasing Euclidean distances $\|\mathbf{X}_i^{(d)} - x\|$ of the $\mathbf{X}_i^{(d)}$ to $x \in \mathbb{R}^d$. In other words, $\mathbf{X}_{(i)}^{(d)}(x)$ is the i-th nearest neighbor of x amongst $\mathbf{X}_j^{(d)}$, $j \in \mathcal{I}_\ell$. If distance ties occur, a tie-breaking strategy must be defined. For example, in case of $\|\mathbf{X}_i^{(d)} - x\| = \|\mathbf{X}_j^{(d)} - x\|$, $\mathbf{X}_i^{(d)}$ may be declared closer to x if $i < j$, i.e., the tie-breaking is done by indices. The k-nearest neighbor classification rule is then defined as

$$(3) \qquad \phi_{\ell,k,d}(x) = \begin{cases} 0 & \text{if } \sum_{i=1}^k \mathbf{1}_{\{Y_{(i)}(x)=0\}} \geq \sum_{i=1}^k \mathbf{1}_{\{Y_{(i)}(x)=1\}} \\ 1 & \text{otherwise.} \end{cases}$$

The validation set is used to select both d and k optimally by minimizing a penalized empirical probability of error based on the independent validation set,

$$(4) \quad (\hat{d}, \hat{k}) = \text{argmin}_{d \geq 1, 1 \leq k \leq \ell} \left[\frac{1}{m} \sum_{j \in \mathcal{J}_m} \mathbf{1}_{\left\{\phi_{\ell,k,d}(\mathbf{X}_j^{(d)}) \neq Y_j\right\}} + \frac{\lambda_d}{\sqrt{m}} \right].$$

Here $\lambda_d/\sqrt{m}$ is a given penalty term which tends to infinity with d to prevent overfitting.

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

$$L(\hat{\phi}_n) = \mathbb{P}\big\{\hat{\phi}_n(\mathbf{X}) \neq Y \mid (\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\big\}.$$

The subscript n in $\hat{\phi}_n$ may be a little confusing, since $\hat{\phi}_n$ is within a class of classifiers using the first ℓ pairs $\mathcal{I}_\ell$ only. However, $\hat{\phi}_n$ depends on the entire data set, as the rest of the data is used for selecting the classifiers. As pointed out by a referee, the classifier $\phi_{n,\hat{k},\hat{d}}$ could be a better choice, at least from a practical point of view. However, for sake of coherence, we did not follow this approach.

The selected classifier $\hat{\phi}_n$ satisfies the following oracle inequality:

Lemma 1. *Assume that*

$$\Delta \equiv \sum_{d=1}^{\infty} e^{-2\lambda_d^2} < +\infty\,. \tag{5}$$

Then there exists a constant $c > 0$, only depending on Δ, such that, for every integer $\ell > 1/\Delta$ and m with $\ell + m = n$,

$$\begin{aligned} &\mathbb{E}L(\hat{\phi}_n) - L^* \\ &\leq \inf_{d\geq 1}\left[(L_d^* - L^*) + \inf_{1\leq k\leq \ell}(\mathbb{E}L(\phi_{\ell,k,d}) - L_d^*) + \frac{\lambda_d}{\sqrt{m}}\right] + c\sqrt{\frac{\log \ell}{m}}\,. \end{aligned} \tag{6}$$

Here

$$L_d^* = \inf_{\phi:\mathbb{R}^d\to\{0,1\}} \mathbb{P}\{\phi(\mathbf{X}^{(d)}) \neq Y\}$$

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

We may view the first term, $L_d^* - L^*$, on the right of the oracle inequality as an approximation term – the price to be paid for using a finite dimensional approximation – and it converges to zero by Lemma 5 below. The second term, $\inf_{1\leq k\leq\ell}(\mathbb{E}L(\phi_{\ell,k,d}) - L_d^*)$, can be handled by the results obtained for nearest neighbor classification in finite dimensions: for fixed d, it converges to zero provided $k \to \infty$ and $k/\ell \to 0$ as $\ell \to \infty$. Since the infimum in (6) is taken over all $d \geq 1$, convergence is ensured.

Theorem 2. *Under the assumption (5) and*

$$\lim_{n\to\infty} \ell = \infty,\ \lim_{n\to\infty} m = \infty,\ \lim_{n\to\infty} \frac{\log \ell}{m} = 0\,, \tag{7}$$

the classifier $\hat{\phi}_n$ is universally consistent, that is,

$$\lim_{n\to\infty} \mathbb{E}L(\hat{\phi}_n) = L^*\,,$$

for any distribution of $(X,Y) \in \mathcal{X} \times \{0,1\}$.

This consistency result is new and is especially valuable since few theoretical results have been established for functional classification. Abraham, Biau, and Cadre (2003) investigate asymptotic properties of the classical moving window classification rule [Devroye, Györfi, and Lugosi (1996, Chapter 10)] in the context of random functions. In particular, these authors discuss the necessity of placing restrictions both on the functional space and the distribution of (X,Y) in order to obtain suitable consistency properties. In contrast, Theorem 2 shows that no assumptions are needed for the nearest neighbor-type classifier (3), extending the result obtained by Stone (1977).

As noted by a referee, similar results can be obtained if we replace the k-nearest neighbor procedure by other universally consistent classifiers in $\mathbb{R}^d$. Kernel classification rules studied in detail in Devroye, Györfi, and Lugosi (1996, Chapter 10), provide such a class of universally consistent rules. However, for sake of concreteness, we only present nearest neighbor rules.

Rather than using the penalized criterium (4), we can select $(\hat{d},\hat{k})$ as

(8) $$(\hat{d},\hat{k}) = \mathrm{argmin}_{1\le d\le d_n, 1\le k\le \ell} \sum_{j\in\mathcal{J}_m} \mathbf{1}_{\left\{\phi_{\ell,k,d}(\mathbf{X}_j^{(d)})\neq Y_j\right\}}.$$

This is equivalent to (4) with $\lambda_d = 0$ for $d \le d_n$ and $\lambda_d = +\infty$ for $d \ge d_n+1$. Inspection of the proof of Theorem 2 reveals that its conclusion still holds, provided that

$$\lim_{n\to\infty} \frac{\log d_n \log \ell}{m} = 0.$$

A safe choice seems $d_n = n$, see Sections 4 and 5.

The type of consistency employed in Theorem 2 is called *weak* consistency in the literature. Under a mild assumption on the distribution of X_i, namely

(9) The distribution of $\mathbf{X}_i^{(d)} = (X_{i1},\dots,X_{id})$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$ for all $d \ge 1$,

we can strengthen the result of Theorem 2 to *strong* consistency of the classifier $\hat{\phi}_n$ using the following oracle inequality:

Lemma 3. *Let Δ and L_d^* be as in Lemma 1. Then for every integer $\ell > 1/\Delta$ and m with $\ell + m = n$, and all $\delta > 0$,*

$$\mathbb{P}\left\{L(\hat{\phi}_n) - L^* \geq \inf_{d \geq 1}\left[(L_d^* - L^*) + \inf_{1 \leq k \leq \ell}(L(\phi_{\ell,k,d}) - L_d^*) + \frac{2\lambda_d}{\sqrt{m}}\right] + \delta\right\}$$
$$\leq 2\Delta\ell \exp\left(-\frac{m\delta^2}{2}\right).$$

Theorem 4. *Assume that the assumptions (5) and (9) hold. Assume moreover that*

$$\lim_{n\to\infty} \ell = \infty, \quad \lim_{n\to\infty} m = \infty, \qquad \text{and} \qquad \sum_{n=1}^{\infty} \ell \exp\left(-\frac{m\delta^2}{2}\right) < +\infty \text{ for all } \delta > 0. \tag{10}$$

Then the classifier $\hat{\phi}_n$ is strongly consistent, that is,

$$\mathbb{P}\left\{\lim_{n\to\infty} L(\hat{\phi}_n) = L^*\right\} = 1,$$

for any distribution of $(X, Y) \in \mathcal{X} \times \{0, 1\}$.

Possible choices for ℓ and m satisfying (10) are, for instance, $\ell = n/2$ for n even, and $\ell = (n+1)/2$ for n odd.

3. Application: speech recognition

We created a small test problem by selecting two short words for classification based on digitized speech frames. The data comprise three sets, each containing $n = 100$ recordings of two different words. The first set corresponds to the words *yes* (label 1, 48 items) and *no* (label 0, 52 items); the second to the words *boat* (label 1, 55 items) and *goat* (label 0, 45 items); the third to the phonemes *sh* (as in *she*, label 1, 42 items) and *ao* (as in *water*, label 0, 58 items). The data in each of our three examples arise from the discretization of analog signals and consist of 100 time series of length 8192, with known class (word) membership. Figure 1 shows two typical speech frames in each data set. All data are available in compressed Matlab format at http://www.math.univ-montp2.fr/∼biau/bbwdata.tgz. We opted for the trigonometric basis since the Fast Fourier Transform (FFT) algorithm enables us to find all the coefficients extremely efficiently.

In the examples and simulations, we observed that a good choice of the penalty term λ_d is needed for the success of our classification algorithm. Practically, an abusive penalization of high dimensions can mask information that is helpful for discrimination. The theory developed in Section 2 shows that any penalty term satisfying (5) could be used, and there is an uncountable number of such (subjective) choices. For these reasons, we decided to restrict ourselves to dimensions up to $d = 100$, and compute $(\hat{d}, \hat{k})$ according to (8). Formally, we just set $\lambda_d = 0$ for $d \leq 100$ and $\lambda_d = +\infty$ for $d \geq 101$. This bound on the selected dimension appears as a safe choice for the data at hand, since the effective selected dimensions do not go beyond 45 (see below). This procedure is the one used in what follows and it is referred to as FOURIER.

In addition, our nonparametric functional classification methodology was compared with two existing alternative approaches:

- NN-DIRECT denotes the k-nearest neighbor rule directly applied to the data $X_1, \ldots, X_n$ without reducing the dimension. This approach is somehow similar to the approach presented in Abraham, Biau, and Cadre (2003) for the moving window rule. As for the FOURIER method, the optimal number of neighbors $\hat{k}$ is selected using the data-splitting procedure [for more on this, see Devroye, Györfi, and Lugosi (1996, Chapters 22 and 26)].
- QDA stands for the standard Quadratic Discriminant Analysis [Mardia, Kent, and Bibby (1979)] performed in the low-dimensional space selected, as in the FOURIER method, by data-splitting. This approach is studied in detail in Hall, Poskitt, and Presnell (2001) and used as a benchmark in our context.

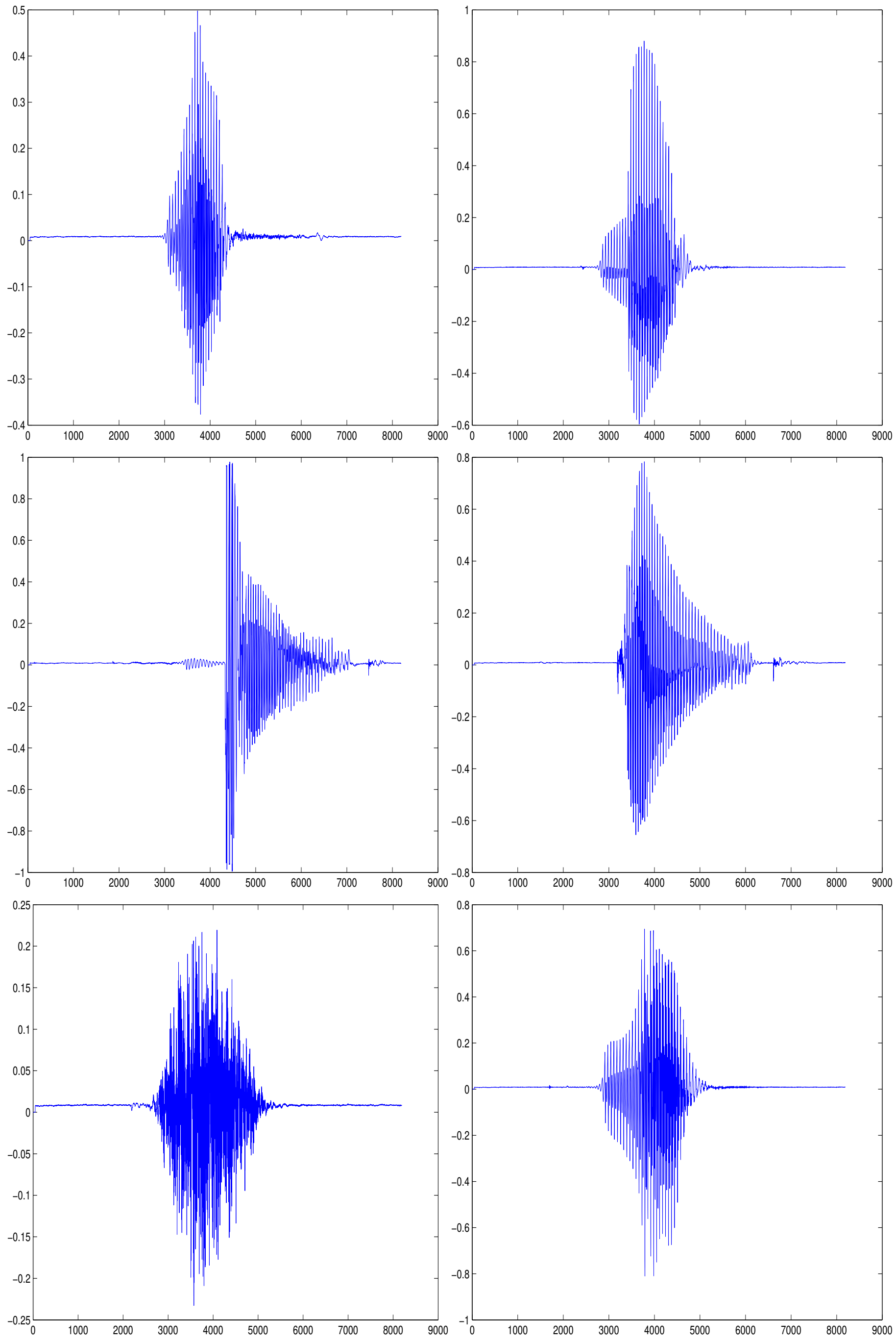

Figure 1. Typical digitized speech frames for the words *yes* (top, left) and *no* (top, right), *boat* (middle, left) and *goat* (middle, right), *sh* (bottom, left) and *ao* (bottom, right).

The error rate for classifying new observations is unknown, but it can be estimated using cross-validation: select one curve with its label from the 99 remaining observations, and treat this curve as a new curve. Next, determine the label using the proposed procedure on the 99 remaining data, and compare the estimated label with the true one. This process, repeated for each of the 100 curves, provides us with an estimate of the error rate as well as 100 estimated parameters $(\hat{d}, \hat{k})$ (for FOURIER), $\hat{k}$ (for NN-DIRECT) and $\hat{d}$ (for QDA). Interestingly, we found that the selected parameters were stable. In the first data set (*yes* and *no*), the pairs $(\hat{d}, \hat{k}) = (8, 2)$ and (14,4) were selected by the method FOURIER in 98 and 2 cases, respectively. Method NN-DIRECT always selected $\hat{k} = 2$, and method QDA selected $\hat{d} =$ 12, 13, 14, 15 in 2, 11, 86 and 1 cases, respectively. In the second data set (*boat* and *goat*), method FOURIER selected the pairs $(\hat{d}, \hat{k}) =$ (15,42), (37,4), (40,2), (41,2), (45,3), (46,2), (48,3) in 1, 1, 10, 3, 1, 83, 1 cases, respectively. Method NN-DIRECT selected $\hat{k} =$ 39, 40, 41 in 2, 67, 31 cases, respectively, and method QDA selected $\hat{d} =$ 1, 3, 4 in 94, 2, 4 cases, respectively. Finally, in the third data set (*sh* and *ao*), method FOURIER selected $(\hat{d}, \hat{k}) =$ (9,1) and (19,1) in 1 and 99 cases, respectively; method NN-DIRECT selected $\hat{k} =$ 38, 40 in 20, 80 cases, respectively; and method QDA selected $\hat{d} =$ 7, 9, 12 in 1, 98 and 1 cases, respectively.

Table 1 summarizes the results obtained for the three data sets with $\ell = m = n/2 = 50$. We see that method FOURIER achieves the best estimated error rates in the last two data sets, and is slightly inferior to method QDA in the first data set. The rather poor results of the method NN-DIRECT are not surprising. Due to the discretizing process, this approach amounts to applying the k-nearest neighbor rule within a space of dimension 8192! The classical QDA produces stable results, similar but slightly inferior to those achieved by our method. Significant improvements of discriminant analysis for situations such as those obtained by discretizing a function can be found in Hastie, Tibshirani, and Buja (1994), Hastie, Buja, and Tibshirani (1995), and Ferraty and Vieu (2003).

Method	Estimated error rate
FOURIER	0.10
NN-DIRECT	0.36
QDA	0.07

Method	Estimated error rate
FOURIER	0.21
NN-DIRECT	0.42
QDA	0.35

Method	Estimated error rate
FOURIER	0.16
NN-DIRECT	0.42
QDA	0.19

TABLE 1. Results obtained for (from top to bottom) the first data set (*yes* and *no*), second data set (*boat* and *goat*), and the third data set (*sh* and *ao*).

4. A small simulation study

The data-splitting device can be unstable due to the variability caused by taking a random partition, as observed by Hengartner, Matzner-Løber, and Wegkamp (2002) in the context of bandwidth selection in local linear regression smoothers. Figure 2 below shows that this remains true in the classification context: for three different toy models (one per row, see below for details), different random splits lead to different selected dimensions and number of neighbors. This suggests that, as a general strategy, one needs to consider B different random partitions of the data and then combine the B corresponding classifiers, where B is a user-specified number. In this section we explore, via a small simulation study, two ways of taking into account the variability induced by random partitions.

For each random partition b, $1 \leq b \leq B$, we denote by $\hat{d}_b, \hat{k}_b$ the selected dimension and number of neighbors, respectively. Let $D = \text{median}_{1\leq b\leq B}\hat{d}_b$ and $K = \text{median}_{1\leq b\leq B}\hat{k}_b$. Our first method uses the classifier $\hat{\phi}_{1,n}$ which is the K-nearest neighbor rule in dimension D.

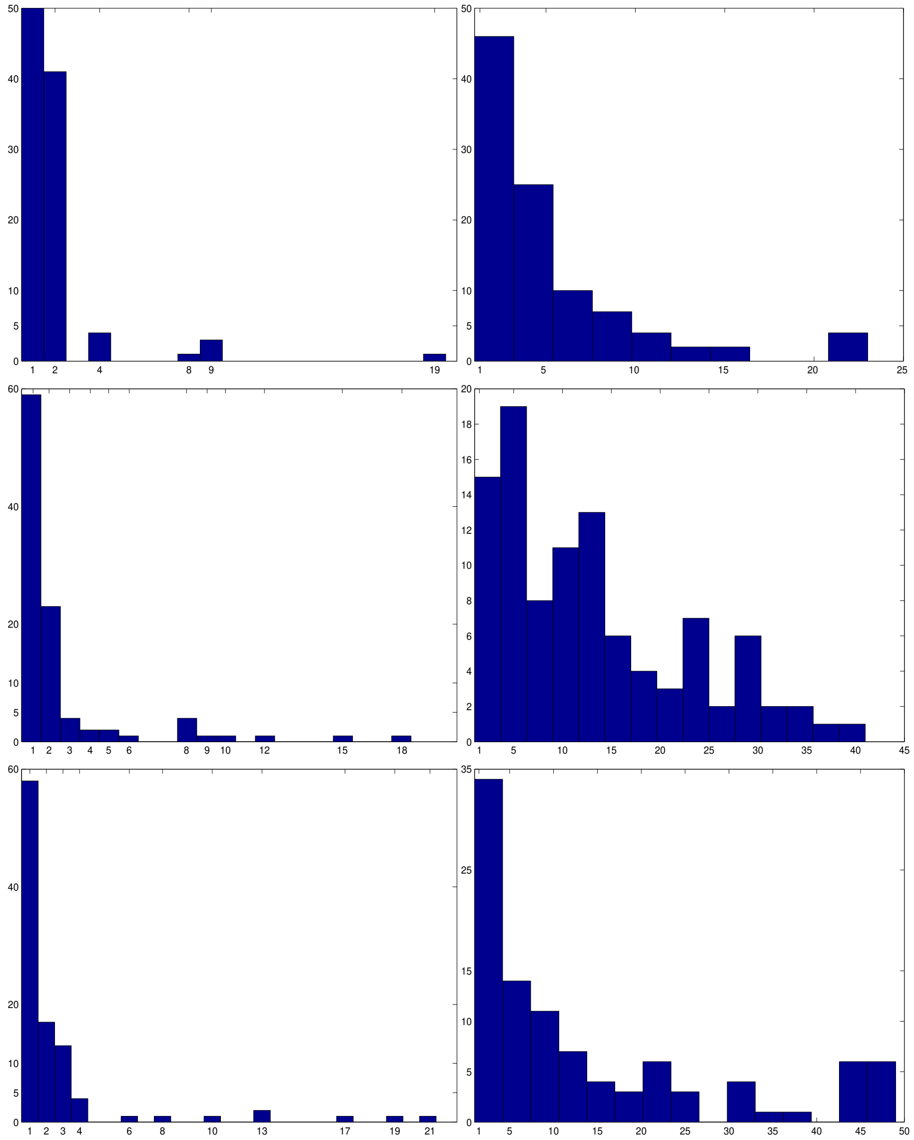

FIGURE 2. Histograms of the selected dimensions (first column) and the selected number of neighbors (second column) based on 100 random partitions for $p = 0.05$ (top), $p = 0.45$ (middle) and $p = 0.5$ (bottom).

In our second method we combine the final votes directly by computing the convex combination classifier

$$\hat{\phi}_{2,n}(x) \quad = \quad \begin{cases} 0 & \text{if } \dfrac{1}{B}\displaystyle\sum_{b=1}^{B}\hat{\phi}_{b,n}(x) \leq 1/2 \\ 1 & \text{otherwise,} \end{cases}$$

where $\hat{\phi}_{b,n}$ is the nearest neighbor classifier based on the selected $\hat{d}_b, \hat{k}_b$ in partition b. Literature on combining classifiers is quite extensive, as they comprise popular techniques such as bagging and boosting. For a description of various methods, we refer to Hastie, Tibshirani, and Buja (1994), and for a recent survey on theoretical advances, we refer to Boucheron, Bousquet, and Lugosi (2004).

We investigate the performance of the two combined estimates suggested above in the following scenario. For each $1 \le i \le n$ we generate pairs $(X_i(t), Y_i)$ as follows: $X_i(t) = U_i \exp(-U_i t)$, where U_i is a uniform random variable on $[1, 11]$ and t ranges in $[0, 1]$. For a given $0 < p < 1$, if $U_i < 6$, we generate Y_i from a Bernoulli $(1-p)$ distribution and if $U_i \ge 6$, we generate Y_i from a Bernoulli (p) distribution. One sample of $n = 100$ observations was generated for each value of p. Figure 3 plots five typical realizations of $X_i(t)$ for $p = 0.25$.

Figure 2 shows histograms of the selected $\hat{d}_b$ and $\hat{k}_b$ based on $B = 100$ data splits for different values of $p = 0.05, 0.45$ and 0.5, respectively. For each split, both parameters were simultaneously selected by the algorithm described in Section 2, with $\ell = m = n/2 = 50$. Note also that each curve was sampled on a uniform mesh of 256 points. The results are very similar for the three values of p: the empirical distribution of the selected dimension concentrates around low values, but there are few outliers in each case. The distribution of the selected number of neighbors is markedly more spread out.

Thus, to reduce the variability induced by taking a random partition of the data, we first computed the pairs $(\hat{d}_b, \hat{k}_b)$ for $B = 20$ splits (still keeping $\ell = m = n/2 = 50$). Then we estimated on a new independent sample of size 100 the misclassification rate of the K-nearest neighbor classifier $\hat{\phi}_{1,n}$, with input parameters $D = \text{median}_{1 \le b \le B} \hat{d}_b$ and $K = \text{median}_{1 \le b \le B} \hat{k}_b$, as well as the misclassification rate for the convex combination classifier $\hat{\phi}_{2,n}$.

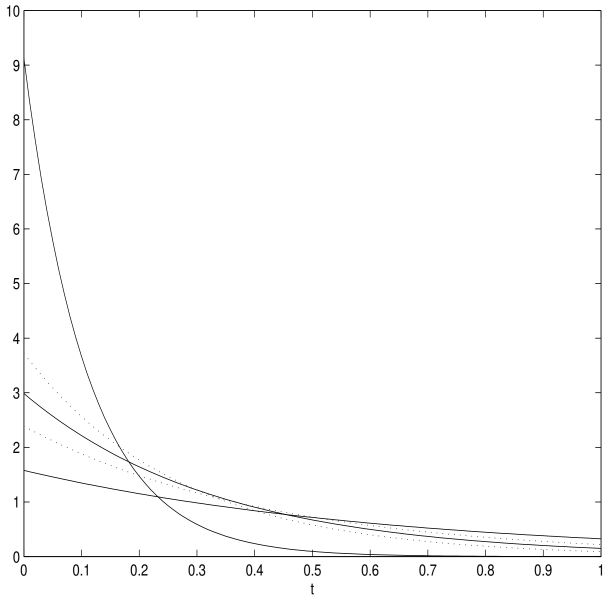

FIGURE 3. Five typical realizations of simulated curves with labels 1 (continuous line) and 0 (dashed line) ($p = 0.25$).

Table 2 summarizes the average simulation results, over 100 replications, obtained for both classifiers $\hat{\phi}_{1,n}$ and $\hat{\phi}_{2,n}$.

We compare the error corresponding to each classifier with the Bayes error using the formula [cf. Devroye, Györfi, and Lugosi (1996, Chapter 2)]

$$L^*(p) = \frac{1}{2} - \frac{1}{2}\mathbb{E}\left[|2\eta(X) - 1|\right].$$

Since for our simulated data we have $\eta(X) = (1-p)\mathbf{1}_{\{U<6\}} + p\mathbf{1}_{\{U\geq 6\}}$, $X(t) = U\exp(-Ut)$ and $U \sim$ Unif $[1, 11]$, we obtain $L^*(p) = p$ for $p \leq 1/2$.

Average estimates	$\mathbf{p = L^*(p) = 0.25}$	$\mathbf{p = L^*(p) = 0.45}$
Average D	1.29	1.35
Average K	10.86	12.94
Average estimated error rate for $\hat{\phi}_{1,n}$	0.30	0.49
Average estimated error rate for $\hat{\phi}_{2,n}$	0.27	0.48

TABLE 2. Simulation results based on 100 replications for $p = 0.25$ and $p = 0.45$.

The results in Table 2 strongly suggest that both methods perform very well, even for small ($n = 100$) sample sizes. The estimated misclassification errors based on either of the classifiers $\hat{\phi}_{1,n}$ and $\hat{\phi}_{2,n}$ are quite similar, and close to the the optimal error rate $L^*(p)$.

5. Proofs

Proof of Theorem 2. By Lemma 5 below, there exists d_0 such that $L_d^* - L^* \leq \varepsilon$ for $d \geq d_0$. Stone (1977) proved that for any $d \geq 1$, there exists a suitable sequence $k_\ell \leq \ell$ with $k_\ell \to \infty$ and $k_\ell/\ell \to 0$ such that $\mathbb{E}L(\phi_{\ell,k_\ell,d}) \to L_d^*$ as $\ell \to \infty$. Invoke oracle inequality (6) to conclude that

$$\begin{aligned}
&\mathbb{E}L(\hat{\phi}_n) - L^* \\
&\quad \leq \inf_{d\geq 1}\left[(L_d^* - L^*) + \inf_{1\leq k\leq \ell}(\mathbb{E}L(\phi_{\ell,k,d}) - L_d^*) + \frac{\lambda_d}{\sqrt{m}}\right] + c\sqrt{\frac{\log \ell}{m}} \\
&\quad \leq (L_{d_0}^* - L^*) + \left(\mathbb{E}L(\phi_{\ell,k_\ell,d_0}) - L_{d_0}^*\right) + \frac{\lambda_{d_0}}{\sqrt{m}} + c\sqrt{\frac{\log \ell}{m}} \\
&\quad \leq \varepsilon + \mathrm{o}(1), \ \text{ as } n \to \infty,
\end{aligned}$$

under (7). Since ε is arbitrary, the theorem follows. □

Recall that $\mathbf{X}^{(d)}$ is the projection vector $(X_1, \ldots, X_d)$ of $\mathbf{X}$ onto $\mathbb{R}^d$. L_d^* is the Bayes error in $\mathbb{R}^d$ achieved by the Bayes rule

$$\phi_d^*(x) = \begin{cases} 0 & \text{if } \mathbb{E}[Y|\mathbf{X}^{(d)} = x] \leq 1/2 \\ 1 & \text{otherwise.} \end{cases}$$

Lemma 5. *We have*

$$L_d^* - L^* \to 0 \quad \textit{as } d \to \infty\,.$$

Proof. From Devroye, Györfi, and Lugosi (1996, Theorem 2.2, page 16), we have, for every $d \geq 1$,

$$L_d^* - L^* = \mathbb{P}\big\{\phi_d^*(\mathbf{X}^{(d)}) \neq Y\big\} - \mathbb{P}\big\{\phi^*(\mathbf{X}) \neq Y\big\} \leq 2\mathbb{E}\big|\mathbb{E}[Y|\mathbf{X}^{(d)}] - \mathbb{E}[Y|\mathbf{X}]\big|\,.$$

Note that $M_d = \mathbb{E}[Y|\mathbf{X}^{(d)}] = \mathbb{E}[Y|X_1, \ldots, X_d]$ is a uniformly bounded martingale with respect to the natural filtration $\sigma(X_1, \ldots, X_d)$. By the martingale convergence theorem [cf. Pollard (2002, Corollary 27, page 151)], it follows that M_d converges in L_1 to a limit M_∞, which equals $\mathbb{E}[Y|\mathbf{X}]$ since

$|M_d| \le 1$ with probability one [cf. Pollard (2002, Theorem 36, page 154)]. The claim of the lemma follows. $\square$

Proof of Lemma 1. To simplify notation, we define

$$L(k,d) = \mathbb{P}\big\{\phi_{\ell,k,d}(\mathbf{X}^{(d)}) \neq Y \,|\, (\mathbf{X}_i, Y_i), 1 \le i \le n\big\}$$

and

$$\hat{L}(k,d) = \frac{1}{m} \sum_{j \in \mathcal{J}_m} \mathbf{1}_{\{\phi_{\ell,k,d}(\mathbf{X}_j^{(d)}) \neq Y_j\}},$$

so that the criterion to be minimized in d and k reads

$$\hat{L}(k,d) + \frac{\lambda_d}{\sqrt{m}}.$$

Fix $\varepsilon > 0$. For every $d \ge 1$ and every k satisfying $1 \le k \le \ell$, we have

$$\mathbb{P}\left\{L(\hat{k},\hat{d}) - \hat{L}(k,d) > \frac{\lambda_d}{\sqrt{m}} + \varepsilon\right\} \le \mathbb{P}\left\{L(\hat{k},\hat{d}) - \hat{L}(\hat{k},\hat{d}) \ge \frac{\lambda_{\hat{d}}}{\sqrt{m}} + \varepsilon\right\},$$

since by definition of $\hat{k}$ and $\hat{d}$,

$$\hat{L}(\hat{k},\hat{d}) + \frac{\lambda_{\hat{d}}}{\sqrt{m}} \le \hat{L}(k,d) + \frac{\lambda_d}{\sqrt{m}}.$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\left\{L(\hat{k},\hat{d}) - \hat{L}(k,d) > \frac{\lambda_d}{\sqrt{m}} + \varepsilon\right\} \\
&\quad \le \sum_{d=1}^{\infty} \sum_{k=1}^{\ell} \mathbb{P}\left\{L(k,d) - \hat{L}(k,d) \ge \frac{\lambda_d}{\sqrt{m}} + \varepsilon\right\} \\
&\qquad \text{(by the union bound)} \\
&\quad = \sum_{d=1}^{\infty} \sum_{k=1}^{\ell} \mathbb{E}\mathbb{P}\left\{L(k,d) - \hat{L}(k,d) \ge \frac{\lambda_d}{\sqrt{m}} + \varepsilon \,|\, (\mathbf{X}_i, Y_i),\, i \in \mathcal{I}_\ell\right\} \\
&\quad \le \sum_{d=1}^{\infty} \ell \exp\left(-2m\left[\frac{\lambda_d}{\sqrt{m}} + \varepsilon\right]^2\right) \\
&\qquad \text{[by Hoeffding's inequality, see e.g. Devroye et al. (1996, p. 122)]} \\
&\quad \le \Delta \ell e^{-2m\varepsilon^2},
\end{aligned}$$

where $\Delta = \sum_{d=1}^{\infty} e^{-2\lambda_d^2} < +\infty$ by assumption (5). Since, for every $d \ge 1$ and k with $1 \le k \le \ell$,

$$\mathbb{E}L(\hat{k},\hat{d}) \le \mathbb{E}\hat{L}(k,d) + \frac{\lambda_d}{\sqrt{m}} + \int_0^{\infty} \mathbb{P}\left\{L(\hat{k},\hat{d}) - \hat{L}(k,d) > \frac{\lambda_d}{\sqrt{m}} + \varepsilon\right\} \mathrm{d}\varepsilon$$

we obtain, for every $u > 0$,

$$\mathbb{E}L(\hat{k},\hat{d}) \le \mathbb{E}\hat{L}(k,d) + \frac{\lambda_d}{\sqrt{m}} + u + \Delta\ell \int_u^\infty e^{-2m\varepsilon^2}\mathrm{d}\varepsilon\,.$$

Note that

$$\begin{aligned}\int_u^\infty e^{-2m\varepsilon^2}\mathrm{d}\varepsilon &\le \frac{1}{2}\int_u^\infty \Big(2+\frac{1}{2m\varepsilon^2}\Big)e^{-2m\varepsilon^2}\mathrm{d}\varepsilon \\ &= -\frac{1}{2}\Big[\frac{1}{2m\varepsilon}e^{-2m\varepsilon^2}\Big]_u^\infty \\ &= \frac{1}{4mu}e^{-2mu^2}\,,\end{aligned}$$

whence, choosing $u = \sqrt{\log(\Delta\ell)/(2m)}$, we obtain

$$\mathbb{E}L(\hat{k},\hat{d}) \le \mathbb{E}\hat{L}(k,d) + \frac{\lambda_d}{\sqrt{m}} + \sqrt{\frac{\log(\Delta\ell)}{2m}} + \frac{1}{2\sqrt{2}\Delta\ell\sqrt{m\log(\Delta\ell)}}\,.$$

Since k and d are arbitrary,

$$\mathbb{E}L(\hat{k},\hat{d}) \le \inf_{d\ge 1, 1\le k\le \ell}\Big(\mathbb{E}\hat{L}(k,d) + \frac{\lambda_d}{\sqrt{m}}\Big) + \sqrt{\frac{\log(\Delta\ell)}{2m}} + \frac{1}{2\sqrt{2}\Delta\ell\sqrt{m\log(\Delta\ell)}}\,.$$

Invoke that $\mathbb{E}\hat{L}(k,d) = \mathbb{E}L(k,d)$ for each fixed k, d to conclude (6). □

Proof of Lemma 3. The proof is similar to the one of Lemma 1. Again, by definition of $(\hat{d},\hat{k})$, we find that for all $d \ge 1$ and $1 \le k \le \ell$,

$$\begin{aligned}L(\hat{k},\hat{d}) &= \Big\{\hat{L}(\hat{k},\hat{d}) + \frac{\lambda_{\hat{d}}}{\sqrt{m}}\Big\} + \Big\{(L-\hat{L})(\hat{k},\hat{d}) - \frac{\lambda_{\hat{d}}}{\sqrt{m}}\Big\} \\ &\le \Big\{\hat{L}(k,d) + \frac{\lambda_d}{\sqrt{m}}\Big\} + \Big\{(L-\hat{L})(\hat{k},\hat{d}) - \frac{\lambda_{\hat{d}}}{\sqrt{m}}\Big\} \\ &\le \Big\{L(k,d) + \frac{2\lambda_d}{\sqrt{m}}\Big\} + R_n,\end{aligned}$$

where

$$\begin{aligned}R_n \equiv& \sup_{d\ge 1, 1\le k\le \ell}\Big\{(L-\hat{L})(k,d) - \frac{\lambda_d}{\sqrt{m}}\Big\} \\ &+ \sup_{d\ge 1, 1\le k\le \ell}\Big\{(\hat{L}-L)(k,d) - \frac{\lambda_d}{\sqrt{m}}\Big\}.\end{aligned}$$

Again by the union bound and Hoeffding's inequality, we find for all $\delta > 0$,

$$\mathbb{P}\left\{R_n \ge \delta\right\} \le 2\Delta\ell\exp\left(-\frac{m\delta^2}{2}\right),$$

and the conclusion of the lemma follows easily. □

Proof of Theorem 4. The Borel-Cantelli lemma and Lemma 3 yield that

$$L(\hat{\phi}_n) - L^* \leq \inf_{d\geq 1}\left[(L_d^* - L^*) + \inf_{1\leq k\leq \ell}(L(\phi_{\ell,k,d}) - L_d^*) + \frac{2\lambda_d}{\sqrt{m}}\right]$$

with probability one, as $n \to \infty$. Under assumption (9), by Theorem 11.1, page 170, in Devroye, Györfi, and Lugosi (1996),

$$L(\phi_{\ell,k_\ell,d}) - L_d^* \to 0,$$

with probability one, for all fixed $d \geq 1$, provided $k_\ell \to \infty$ and $k_\ell/\ell \to 0$. Invoke Lemma 5 and the same reasoning applied in the proof of Theorem 2 to conclude the proof. □

Acknowledgments. The authors thank the Associate Editor and three referees for their constructive remarks and Laurent Rouvière for his help with the simulation study.

References

Abraham, C., Biau, G., and Cadre, B. (2003). On the kernel rule for function classification. *Technical report*, University Montpellier II. http://www.math.univ-montp2.fr/~biau/publications.html .

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

Boucheron, S., Bousquet, O., and Lugosi, G. (2004). Theory of classification: A survey of some recent advances. *ESAIM: Probability and Statistics* (in press).

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

Cover, T.M. and Hart, P.E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13:21 – 27.

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

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

Diabo-Niang, S. and Rhomari, N. (2001). Nonparametric regression estimation when the regressor takes its values in a metric space. *Technical report*, University Paris VI. http://www.ccr.jussieu.fr/lsta .

Ferraty, F., Peuch, A., and Vieu, P. (2003). Modèle à indice fonctionnel simple. *Comptes Rendus de l'Académie des Sciences de Paris*, 336:1025 – 1028.

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

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

Fix, E. and Hodges, J.L. (1951). Discriminatory analysis. Nonparametric discrimination: Consistency properties. *Technical Report 4*, Project Number 21-49-004, USAF School of Aviation Medicine, Randolph Air Field, TX.

Fix, E. and Hodges, J.L. (1952). Discriminatory analysis: Small sample performance. *Technical Report 21-49-004*, USAF School of Aviation Medicine, Randolph Air Field, TX.

Fix, E. and Hodges, J.L. (1991a). Discriminatory analysis, nonparametric discrimination, consistency properties. In *Nearest Neighbor Pattern Classification Techniques*, ed. Dasarathy, B.V., pp. 32 – 39, IEEE Computer Society Press, Los Alamitos, CA.

Fix, E. and Hodges, J.L. (1991b). Discriminatory analysis: Small sample performance. In *Nearest Neighbor Pattern Classification Techniques*, ed. Dasarathy, B.V., pp. 40 – 56, IEEE Computer Society Press, Los Alamitos, CA.

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

Hastie, T., Tibshirani, R., and Buja, A. (1994). Flexible discriminant analysis by optimal scoring. *Journal of the Royal Statistical Society, Series B*, 89:1255 – 1270.

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

Hengartner, N.W., Matzner-Løber, E., and Wegkamp, M.H. (2002). Bandwidth selection for local linear regression. *Journal of the Royal Statistical Society, Series B*, 64:1 – 14.

Kirby, M. and Sirovich, L. (1990). Application of the Karhunen-Loeve procedure for the characterization of human faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12:103 – 108.

Kulkarni, S.R. and Posner, S.E. (1995). Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Transactions on Information Theory*, 41:1028 – 1039.

Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979). *Multivariate Analysis.* Academic Press, London.

Pollard, D.B. (2002). *A User's Guide to Measure Theoretic Probability.* Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge.

Ramsay, J.O. and Silverman, B.W. (1997). *Functional Data Analysis.* Springer-Verlag, New York.

Ramsay, J.O. and Silverman, B.W. (2002). *Applied Functional Data Analysis. Methods and Case Studies*. Springer-Verlag, New York.

Sansone, G. (1969). *Orthogonal Functions*. Interscience, New York.

Stone, C.J. (1977). Consistent nonparametric regression. With discussion and a reply by the author. *The Annals of Statistics*, 5:595 – 645.

Szegö, G. (1959). *Orthogonal Polynomials, Volume 32*. American Mathematical Society, Providence, RI.

Zygmund, A. (1959). *Trigonometric Series I*. University Press, Cambridge.