Functional Learning with Wavelets

Laurent Rouvière

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
(e-mail: rouviere@ensam.inra.fr)

Abstract. Let X be a random variable taking values in $L_2\big([0,1]\big)$ and let Y be a random label with values in $\{0,1\}$. Given a class of classifiers and n independent copies (X_i, Y_i) of the pair (X, Y), we show how to select optimally a particular classifier in the class and derive its consistency properties. To build our classifier, we first reduce the dimension of the functional observations using a particular thresholding on the coefficients of the curves X_i expressed in a wavelet basis. Then a classification rule working in finite dimension is performed on the selected coefficients. The dimension is automatically selected by data-splitting and empirical risk minimization. An application of this technique to a signal discrimination problem involving speech recognition is presented.
Keywords: Functional Data Analysis, Classification, Wavelets.

1 Introduction

The problem of pattern recognition (or classification 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. In many real-life problems however, input data are in fact sampled functions rather than standard high dimensional vectors, and this casts the classification problem into the class of Functional Data Analysis.

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 *et al.*, 2003] for a detailed discussion, examples and counterexamples). In practice, before applying any learning technique to model real data, a preliminary dimension reduction or model selection step reveals crucial for appropriate smoothing and circumscription 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 observations by considering only the first d coefficients of the data on an appropriate basis. This

approach was followed by [Kirby and Sirovich, 1990], [Comon, 1994], [Belhumeur *et al.*, 1997], [Hall *et al.*, 2001], or [Amato *et al.*, 2005]. Given a collection of functions we wish to classify, [Biau *et al.*, 2005] propose to use first Fourier filtering on each function 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 aim of the present paper is to extend the data-based filtering approach of [Biau *et al.*, 2005] to wavelet bases and general discrimination rules. Our motivation is twofold.

- First, as pointed out for example in [Amato *et al.*, 2005], wavelet bases offer some significant advantages over other bases. Indeed, wavelets can be used successfully for compression of a stochastic process, in the sense that the sample paths can be accurately reconstructed from a fraction of the full set of wavelet coefficients. Further, the wavelet decomposition of the sample paths is a local one, so that if the information relevant to the classification problem is contained in a particular part of the sample functions, as typically it is, this information will be carried by a very small number of wavelet coefficients. Moreover, the ability of wavelets to model the signal at different levels of resolution means that we have the option of selecting from the paths at a range of bandwidths.
- Second, we seek for general performance bounds and consistency results when using (finite dimensional approximations of) the sample data in the selection of a discrimination rule and/or its parameters. This article offers both a practical methodology and general performance results for all those who are willing to use wavelet filtering as a dimension reduction step before effective classification.

Throughout the manuscript, we will adopt the point of view of automatic pattern recognition described, to a large extend, in [Devroye, 1988]. In this setup, one uses a test 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 regarding this classification paradigm are summarized in the next section. In Section 3, we outline the method and state consistency of our classification rule. Section 4 offers some experimental results on real-life data.

2 Automatic pattern recognition

This section gives a brief exposition and set up terminology of automatic pattern recognition. For a detailed introduction, the reader is referee to [Devroye, 1988].

To model the automatic learning problem, we introduce a probabilistic setting. Denote by $\mathcal{F} = L_2([0,1])$ the space of all square integrable functions on $[0,1]$. The data consist 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 the data are artificially split by us 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 *testing sequence*. A discrimination rule is a 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

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

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 that it cannot be smaller than the Bayes probability of error

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

(see [Devroye *et al.*, 1996], Theorem 2.1, page 10). In the learning process, we aim at constructing rules with small probability of error. To do this, we employ the learning sequence to design a class of data-dependent discrimination rules, and we use the testing 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 testing 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

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

and

$$\hat{g}(X_i) = \hat{g}\,(X_i, (X_1, Y_1), \ldots, (X_n, Y_n)),$$

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

[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. Even though there is much to say about multiclass classification, we will not approach this increasing field of research.

3 Dimension reduction for classification

3.1 Wavelet-based expansion of the observations

The theory of wavelets has recently undergone a period of rapid development with exciting implications for nonparametric function estimation. Wavelets are orthonormal basis functions that cut up signals into different frequency components, and then study each component with a resolution matched to its scale. The books of [Daubechies, 1992], [Meyer, 1992] and [Mallat, 1999] give detailed expositions of the mathematical aspects of wavelets.

To summarize in our context, we recall that $L_2([0,1])$ is 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 $\phi_{j,k}, k = 0, \ldots, 2^j - 1$, such that $\text{supp}(\phi_{j,k}) \subset [k2^{-j}, (k+1)2^{-j}]$. 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 $\psi_{j,k}, k = 0, \ldots, 2^j - 1$. Thus, the family

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

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

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

where

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

3.2 Consistent functional classification

In this paragraph, we present the construction of our classifier and discuss its consistency properties. Using the notation of Section 2, the data consist 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 $L_2([0,1])$ as explicited above, each observation X_i is expressed as a series expansion

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

For the sake of coherence, it will be convenient to reindex the sequence $\{\phi_{0,0}, \psi_{0,0}, \psi_{1,0}, \psi_{1,1}, \psi_{2,0}, \psi_{2,1}, \psi_{2,2}, \psi_{3,0}, ...\}$ into $\{\psi_1, \psi_2, \psi_3, ...\}$. With this scheme, expression (1) may be rewritten as

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

hence the random coefficients

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

Let $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots)$ be the coefficients associated with X_i. Recall that the Hilbert space $L_2([0,1])$ is isomorphic with $\ell_2 = \left\{\mathbf{x} = (x_1, x_2, \ldots) : \sum_{j=1}^{\infty} x_j^2 < \infty\right\}$. Consequently, knowing X_i is the same as knowing $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots)$. In our quest of dimension reduction, we first fix in (1) a maximum 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),

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

At this point, we could try to use these finite-dimensional approximations of the observations, and let the data select optimally one of the $2^{2^J} - 1$ subbases of $\{\psi_1, \ldots, \psi_{2^J}\}$. By doing so, we would face with an unreasonable overall algorithmic complexity, and therefore catastrophic subsequent performance bounds. Thus, in order to reduce the overall complexity of the 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 n d-dimensional learning data as input. For fixed training sequence $(x_1, y_1), \ldots, (x_n, y_n)$, denote by $\mathbf{C}_n^{(d)}$ the collection of all sets

$$\left\{\{x \in \mathbb{R}^d : \phi(x) = 1\} : \quad \phi \in \mathbf{D}_n^{(d)}\right\},$$

and define the shatter coefficient as

$$\mathbb{S}_{\mathbf{C}_n^{(d)}}(m) = \max_{z_1,\ldots,z_m \in \mathbb{R}^d} \mathrm{Card}\left\{\{z_1, \ldots, z_m\} \cap C : C \in \mathbf{C}_n^{(d)}\right\}.$$

With a slight abuse of notation, we will denote by $\mathbb{S}_{\mathbf{C}_n}^{(J)}(m)$ the shatter coefficient corresponding to the collection of all rules $\{g^{(d)} : d = 1, \ldots, 2^J\}$ embedded in $\mathbb{R}^{2^J}$. Observe that

$$\mathbb{S}_{\mathbf{C}_n}^{(J)}(m) \leq \sum_{d=1}^{2^J} \mathbb{S}_{\mathbf{C}_n^{(d)}}(m). \tag{3}$$

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

$$\sum_{i=1}^{n} X_{ij_1}^2 \geq \sum_{i=1}^{n} X_{ij_2}^2 \geq \ldots \geq \sum_{i=1}^{n} X_{ij_{2^J}}^2 . \tag{4}$$

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 $\mathbf{X}_i^{(d)} = (X_{ij_1}, \ldots, X_{ij_d})$ (without loose of generality, we assume implicitly that the sequence (j_k) is ordered – if not, just reorder it).

We select the dimension d and the rule simultaneously, using the data-splitting device described in Section 2. Precisely, we select both d and $g^{(d)}$ optimally by minimizing the empirical probability of error based on the independent validation set, that is

$$\left(\hat{d}, \hat{g}^{(\hat{d})}\right) = \underset{d=1,\ldots,2^J, g^{(d)} \in \mathbf{D}_n^{(d)}}{\operatorname{argmin}} \left[\frac{1}{m} \sum_{i=n+1}^{n+m} \mathbf{1}_{[g^{(d)}(X_i^{(d)}) \neq Y_i]} \right]. \tag{5}$$

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

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

The selected rule $\hat{g}$ satisfies the following optimal inequality.

Theorem 1

$$\mathbf{E}\big\{L_{n+m}(\hat{g})\big\} - L^* \leq L_{2^J}^* - L^* + \mathbf{E}\big\{ \inf_{\substack{d=1,\ldots,2^J \\ g^{(d)} \in \mathbf{D}_n^{(d)}}} L_n(g^{(d)}) \big\} - L_{2^J}^*$$

$$+ 2\,\mathbf{E}\left\{ \sqrt{\frac{8 \log\left(4\mathbb{S}_{\mathbf{C}_n}^{(J)}(2m)\right)}{m}} + \frac{1}{\sqrt{(m/2)\log\left(4\mathbb{S}_{\mathbf{C}_n}^{(J)}(2m)\right)}} \right\}.$$

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 be paid for using a finite dimensional approximation – and it converges to zero. The second term,

$$\mathbf{E}\Big\{ \inf_{\substack{d=1,\ldots,2^J \\ g^{(d)}\in\mathbf{D}_n^{(d)}}} L_n(g^{(d)})\Big\} - L^*_{2^J}$$

can be handled by standard results on classifications. Let us first recall the definition of a *consistent* rule: a rule g is consistent if $\mathbf{E}\big\{L_n(g)\big\} \to L^*$ as $n \to \infty$.

Corollary 1 *Let $J \geq 0$ be a fixed integer. Assume that from each $\mathbf{D}_n^{(2^J)}$, $n \geq 1$, we can pick one $g_n^{(2^J)}$ such that the sequence $(g_n^{(2^J)})_{n\geq 1}$ is consistent for a certain class of distributions. Then the automatic rule $\hat{g}$ defined in (5) is consistent for the same class of distributions,* i.e.*,*

$$\mathbf{E}\big\{L_{n+m}(\hat{g})\big\} \to L^* \quad \textit{as } n \to \infty$$

if

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

This consistency result is new and is especially valuable since few theoretical results have been established for functional classification. Corollary 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 just worried about consistency, Corollary 1 reassures us that nothing is lost as long as we take m much larger than $\log \mathbf{E}\left\{\mathbb{S}_{\mathbf{C}_n^{(J)}}(2m)\right\}$. Often, this reduces to a very weak condition on the size m of the testing set and the maximum resolution J. 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 $\mathbb{S}^{(J)}_{\mathbf{C}_n}(2m)$. For example, if $\mathbf{D}_n^{(d)}$ contains all nearest-neighbor rules, a trivial bound 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, 1977] proved the striking result that k-nearest neighbor classifiers are consistent if $X \in \mathbb{R}^d$, provided $k \to \infty$ and $k/n \to \infty$. Thus we see that our strategy leads to a consistent rule whenever $J/m \to 0$ and $\log n/m \to 0$ as $n \to \infty$. For other examples, we refer to [Devroye, 1988].

4 Application to a speech recognition problem

In this section, we illustrate performance of our method. To this aim, we study a part of TIMIT database which was investigated in [Hastie *et al.*, 1995]. 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 : "aa" as the vowel in "dark" (695 items), "a0" as the first vowel in "water" (1022 items), "dcl" as in "dark" (757 items), "iy" as the vowel in "she" (1163 items) and "sh" as in "she" (872 items). The database is a multispeaker database. Each speaker is recorded at a 16k-Hz sampling rate and we retain only the first 256 frequencies (see Figure 1). Thus, the data consist of 4509 series of length 256 with known class membership.

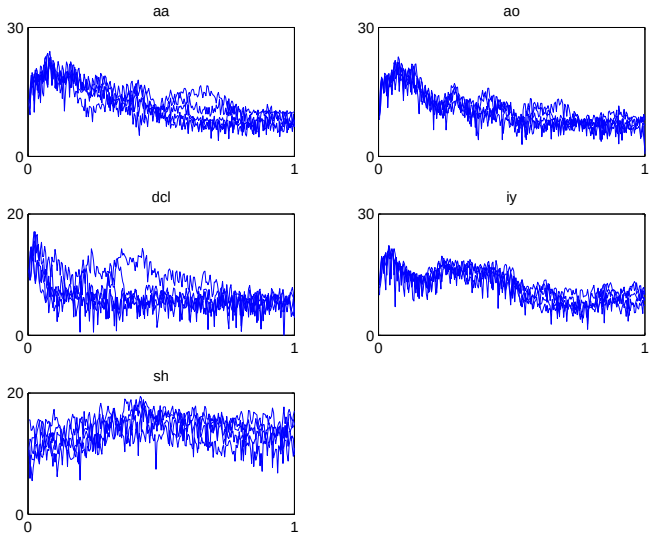

Fig. 1. A sample of 5 log-periodograms per class.

We first compute the wavelet filtering approach described in Section 3 using three collections of rules $\mathbf{D}_n^{(d)}$ working in $\mathbb{R}^d$. Precisely:

- W-LDA denotes the wavelet filtering followed by the class $\mathbf{D}_n^{(d)}$ of all linear discrimination rules.
- W-NN denotes the wavelet filtering followed by the class $\mathbf{D}_n^{(d)}$ of all nearest-neighbor rules.
- W-T denotes the wavelet filtering followed by the class $\mathbf{D}_n^{(d)}$ of all binary trees in which each internal node corresponds to a split perpendicular to one of the axes [Devroye *et al.*, 1996].

In addition, we propose to compare our algorithm with two existing alternative approaches:

- F-NN refers to the Fourier filtering approach combined with the k nearest-neighbor rule described in [Biau *et al.*, 2005].
- MPLSR refers to the multivariate partial least square regression. This approach is studied in detail in [Preda and Saporta, 2002] and is used as a benchmark in our context. The number of PLS components is selected by minimizing the empirical probability of error based on the testing sequence.

We use the split sample approach presented in Section 2 to select the free parameters. The *training sequence* and the *testing sequence* both contain 250 observations. The error rate (*e.r.*) for classifying new observations is unknown, but it can be estimated using the rest of the data:

$$e.r. = \frac{1}{3509} \sum_{i=501}^{4509} \mathbf{1}_{[\hat{g}(X_i) \neq Y_i]} \,,$$

where $\hat{g}$ denotes the selected rule. Table 1 displays the estimated error rates for the different methods together with the dimensions selected (number of PLS components for MPLSR). Results are averaged over 50 random partitions of the data.

Method	*e.r.*	$\hat{d}$
W-LDA	0.0854	18.70
W-NN	0.1096	19.52
W-T	0.1253	9.10
F-NN	0.1277	48.76
MPLSR	0.0904	5.96

Table 1. Estimated error rates.

We see that method W-LDA achieves the best estimated error rates, and that its results are slightly inferior to method MPLSR. The results of the

Fourier-based algorithm are still acceptable, because of a good localisation of the signal.

References

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

[Amato *et al.*, 2005]U. Amato, A. Antoniadis, and I. De Feis. Dimension reduction in functional regression with applications. *Computational Statistics and Data Analysis*, 2005. In press.

[Belhumeur *et al.*, 1997]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.

[Biau *et al.*, 2005]G. Biau, F. Bunea, and M. Wegkamp. Functional classification in hilbert spaces. *IEEE Transactions on Information Theory*, 2005. In press.

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

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

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

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

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

[Hastie *et al.*, 1995]T. Hastie, A. Buja, and R. Tibshirani. Penalized discriminant analysis. *Annals of Statistics*, 23:73–102, 1995.

[Kirby and Sirovich, 1990]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.

[Mallat, 1999]S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 1999. 2nd edition.

[Meyer, 1992]Y. Meyer. *Wavelet and Operators*. Cambridge University Press, Cambridge, 1992.

[Preda and Saporta, 2002]C. Preda and G. Saporta. Régression PLS sur un processus stochastique. *Revue de Statistique Appliquée*, 50(2), 2002.

[Stone, 1977]C.J. Stone. Consistent nonparametric regression. *Annals of Statistics*, 5:595–645, 1977.