From projection pursuit and CART to adaptive discriminant analysis?

Rémi Gribonval

From Projection Pursuit and CART to Adaptive Discriminant Analysis?

Rémi Gribonval

***Abstract*—While many efforts have been put into the development of nonlinear approximation theory and its applications to signal and image compression, encoding and denoising, there seems to be very few theoretical developments of adaptive discriminant representations in the area of feature extraction, selection and signal classification. In this paper, we try to advocate the idea that such developments and efforts are worthwhile, based on the theorerical study of a data-driven discriminant analysis method on a simple—yet instructive—example. We consider the problem of classifying a signal drawn from a mixture of two classes, using its projections onto low-dimensional subspaces. Unlike the linear discriminant analysis (LDA) strategy, which selects subspaces that do not depend on the observed signal, we consider an adaptive sequential selection of projections, in the spirit of nonlinear approximation and classification and regression trees (CART): at each step, the subspace is enlarged in a direction that maximizes the mutual information with the unknown class. We derive explicit characterizations of this adaptive discriminant analysis (ADA) strategy in two situations. When the two classes are Gaussian with the same covariance matrix but different means, the adaptive subspaces are actually nonadaptive and can be computed with an algorithm similar to orthonormal matching pursuit. When the classes are centered Gaussians with different covariances, the adaptive subspaces are spanned by eigen-vectors of an operator given by the covariance matrices (just as could be predicted by regular LDA), however we prove that the order of observation of the components along these eigen-vectors actually depends on the observed signal. Numerical experiments on synthetic data illustrate how data-dependent features can be used to outperform LDA on a classification task, and we discuss how our results could be applied in practice.**

***Index Terms*—Classification and regression trees (CART), classification tree, discriminant analysis, mutual information, nonlinear approximation, projection pursuit, sequential testing.**

I. Introduction

In THE last decade, nonlinear approximation and sparse decompositions with wavelets or other dictionaries of functions [1] have emerged as very successful tools for compression, encoding and denoising of signals and images. Important efforts have been put into the development of a coherent theory of nonlinear approximation [2] and efficient practical algorithms have been introduced and studied [3]–[6].

The principle of nonlinear approximation consists in projecting a signal or an image x (seen as a vector of sample or pixel values in a high-dimensional vector space $\mathcal{H}$) onto a low-dimensional linear subspace $\mathcal{V}$ which is selected *adaptively*, that is to say $\mathcal{V}$ depends on x. Surprisingly, there seems to be very few investigations of what benefits this simple principle could bring to the area of feature extraction, selection and signal classification. Indeed, many dimension reduction techniques used in these domains are linear: they project the data x onto a *fixed* low-dimensional subspace $\mathcal{V}$ and try to keep in the projection $P_{\mathcal{V}}x$ as much as possible of the relevant information carried by x.

One of the objectives of this paper is to advocate and promote the development of a theoretical analysis of adaptive discriminant analysis (ADA), that is to say discriminant analysis based on data-adaptive projections. Our contribution to this development consists in the mathematical analysis of an ADA method—which is based on an information theoretic optimization criterion—in the context of discrimination between two Gaussian multivariate classes with either equal means or equal covariances.

The paper is organized as follows. In Section II, after recalling the principles of linear and nonlinear approximation, we switch to the framework of (statistical) classification and feature selection. A quick tour through classical techniques sheds light on their common structure, which resembles that of linear approximation: a linear projection onto a fixed subspace, independent of the test data to classify. In Section III, combining ideas from projection pursuit [7] and classification and regression trees (CART) [8], we give a theoretical description of what an ADA method should look like, and point out some of the difficulties and questions related with this approach.

Our main contribution starts in Section IV: we state our theorems about ADA on the example of a mixture of two multidimensional Gaussian random variables. Our main point is that, if the two Gaussians share the same mean but have different covariance matrices $\Sigma_0 \neq \Sigma_1$, then the feature sequence (chosen among the huge dictionary containing all possible unit vectors in $\mathbb{R}^N$), which consists in eigen-vectors of $\Sigma_0^{-1}\Sigma_1$ (Theorem 2), should be observed in an order that *depends* (Theorem 3) on the test data x that is to be classified. The theorems are proved in the Appendix. Based on our theoretical results, we describe in Section V an adaptive feature selection and classification algorithm and compare it numerically to linear discriminant analysis (LDA). The numerical experiments on synthetic data show that the data-adaptive classifier outperforms LDA.

II. From Approximation to Classification

In this section, we briefly introduce some general background about linear/nonlinear approximation, Bayesian decision theory and feature extraction.

Manuscript received March 10, 2003; revised August 30, 2004. This work was supported in part by the National Science Foundation (NSF) under Grant DMS-9872890.

The author is with the French National Center for Computer Science and Control (INRIA) at IRISA, 35042 Rennes, France (e-mail: remi.gribonval@inria.fr).

Digital Object Identifier 10.1109/TNN.2005.844900

A. Linear and Nonlinear Approximation

Given an orthonormal basis $\mathcal{B} = \{g_n\}_{n=1}^N$ of the finite dimensional Euclidian space $\mathbb{R}^N$, one can approximate (linearly) any input vector x by its orthonormal projection

$$P_{\mathcal{V}_M} x = \sum_{m=1}^{M} \langle x, g_m \rangle g_m$$

onto the subspace

$$\mathcal{V}_M := \operatorname{span}\{g_m, 1 \le m \le M\}.$$

If X is a random variable (in the rest of this paper we will use lower case letters for the realizations and upper case letters for random variables) and $\mathcal{B}$ is its Karhunen-Loève basis (or principal components), then it is well known that for each M, the subspace $\mathcal{V}_M$ minimizes the expected error $\mathbb{E}\{\|X - P_{\mathcal{W}_M} X\|^2\}$ of *linear approximation* over all M-dimensional subspaces $\mathcal{W}_M$. However, nonlinear approximation [2] provides for each input vector x a smaller approximation error than linear approximation. If $I_M(x)$ denotes the set of the M largest coefficients of x in the basis—which obviously depends on x—then the best nonlinear approximant of x is

$$\sum_{m \in I_M(x)} \langle x, g_m \rangle g_m.$$

B. Feature Extraction, Selection, and Classification

The classification of a high-dimensional signal $x \in \mathbb{R}^N$ (typically, a speech feature vector or an image), consists in finding its unknown class y, where y is a symbol in a finite alphabet, e.g., the name of the speaker in a speaker identification problem. Assuming the observed data x and its unknown class y are realizations of random variables X and Y with joint probability distribution $\mathcal{P}(X, Y)$, the classification of an observed data x can theoretically be performed using the maximum likelihood (ML) estimator $\hat{Y}_{\mathrm{ML}}(x) := \arg\max_y \mathcal{P}(x \,|\, y)$ or, in a Bayesian framework, the maximum *a posteriori* (MAP) estimator $\hat{Y}_{\mathrm{MAP}}(x) := \arg\max_y \mathcal{P}(y \,|\, x)$.

Remark 1: Generally, a training dataset $\{(x_t, y_t), 1 \le t \le T\}$ is used in a learning stage to estimate a model $\hat{\mathcal{P}}$ of $\mathcal{P}$. In this paper, we will not touch upon the intrinsic statistical problems of the estimation of $\mathcal{P}$; we will assume $\mathcal{P} = \hat{\mathcal{P}}$ is a perfectly known distribution, therefore, no training set is assumed. In this sense, this paper is of a theoretical nature.

Due to the high dimension of the random variable X and the possibly intricate structure of the conditional law $\mathcal{P}(\cdot \,|\, \cdot)$, the true ML/MAP estimator is not usable in practice and is commonly replaced with some new estimator $\hat{Y}(\{f_m(x)\}_{m=1}^M)$ where $\{f_m(x)\}_{m=1}^M, M \ll N$ is a low-dimensional vector of *features*. A new problem becomes the selection of appropriate features that do not degrade the performance of the classification. The rest of this paper is focussed on the selection of *linear features* of the form $f_m(x) = \langle x, g_m \rangle$, for some vector $g_m \in \mathbb{R}^N$.

Several techniques of feature selection have been introduced and thoroughly studied in the literature of pattern recognition [9]–[13]. Let us make a quick tour.

1) Principal Component Analysis: First comes to mind principal component analysis (PCA), which we described briefly above: it selects features according to their *approximation power*, but we have already seen that a nonlinear approximation strategy is more efficient, i.e., it better describes the data. Moreover, classification requires the selection of *informative* features, which is quite different from the good approximation power of the features selected by linear/nonlinear approximations.

2) Independent Component Analysis and Sparse Coding: Alternatives to PCA are independent component analysis (ICA) [14] and sparse coding [15], which have been the subject of intense research in the last decade. While the Karhunen–Loève (orthonormal) basis $\mathcal{B} = \{g_n\}_{n=1}^N$ selected by PCA merely decorrelates the components $\{\langle X, g_n \rangle, 1 \le k \le N\}$, ICA and sparse coding attempt to find independent (resp. *sparse*) components. This is generally done with nonlinear optimization techniques relying on higher order moments. Independence (resp. sparsity) of the features $f_m(x) = \langle x, g_m \rangle$ is certainly a desirable property for many applications such as source separation or compression, however it does not guarantee that each feature brings any valuable information about the unknown class y. In other words, independent (resp. sparse) components are not necessarily discriminant.

3) LDA: In Fisher's LDA [9], [11], [12], a basis $\mathcal{B} = \{g_n\}_{n=1}^N$ is defined such that for each M, the M first basis vectors $\{g_1, \ldots, g_M\}$ maximize a discriminant measure. Computationally efficient LDA algorithms were recently defined, where a (suboptimal) basis is selected from a library of bases [3], [16]–[18]. A common aspect of LDA techniques is that, for a prescribed number M of features, the very same set of basis vectors $\{g_1, \ldots, g_M\}$ (i.e., the same linear projector $P_{\mathcal{V}_M}$ on the same subspace $\mathcal{V}_M$) is used for every input data x that needs to be classified.

III. From LDA to Adaptive Feature Selection

LDA is analogue in structure to linear approximation, with the difference that the chosen subspace $\mathcal{V}_M$ maximizes a discriminant measure instead of minimizing an approximation error. In this section, we consider what could be the analogue of nonlinear approximation for discriminant analysis, that is to say how one could select a data-dependent subspace $\mathcal{V}_M(x)$ with a discriminant measure. First, we see how *nonadaptive, embedded* subspaces $\mathcal{V}_1 \subset \ldots \subset \mathcal{V}_M \subset \mathcal{V}_{M+1} \subset \ldots$ can be selected using a projection pursuit approach [7]. Then, using ideas from CART [8], we will propose a theoretical mean of changing viewpoint and making the sequential selection adaptive. In Section IV, we will provide a more explicit construction of data-dependent embedded subspaces for a classification problem with two Gaussian classes. We will describe in Section V an algorithm that is based on this construction and compare it numerically to LDA on an example.

A. Sequential LDA and Projection Pursuit

In the spirit of the projection pursuit/matching pursuit algorithm [7], [19], one can select iteratively (linear) features from a *dictionary* $\mathcal{D} = \{g : g \in \mathcal{D}\}$ of unit vectors with an information criterion. A first vector

$$g_1 := \arg\max_{g\in\mathcal{D}} \mathcal{I}(Y; \langle X, g\rangle) \tag{1}$$

is selected, where $\mathcal{I}(Y; f(X))$ is a measure of the "average information" that the random variable $f(X)$ gives about Y. In this paper, we will focus on the case where $\mathcal{I}(Y; f(X)) := \mathcal{H}(Y) - \mathcal{H}(Y \,|\, f(X)) = \mathcal{H}(f(X)) - \mathcal{H}(f(X)\,|\,Y)$ is the mutual information [20], however, one could consider several other measures of information, such as the Hellinger distance, the Kullback–Leibler divergence, etc.

Using the chain rules for mutual information [20], the following vectors are iteratively defined (for $m \geq 1$) as:

$$\begin{aligned} g_{m+1} &:= \arg\max_{g\in\mathcal{D}} \mathcal{I}(Y; \langle X, g_1\rangle, \ldots, \langle X, g_m\rangle, \langle X, g\rangle) \\ &= \arg\max_{g\in\mathcal{D}} \mathcal{I}\left(Y; \langle X, g\rangle \,|\, \{\langle X, g_l\rangle\}_{l=1}^{m}\right). \end{aligned} \tag{2}$$

Remark 2: There is, in general, no simple expression of the conditional mutual information described previously, hence, real-life algorithms must estimate it in practice with Monte Carlo methods, and its estimate for large m has poor statistical significance. In order to overcome this issue, some authors [18], [21], [22] replace (2) with

$$g_{m+1} := \arg\max_{g\in\mathcal{D}} \mathcal{I}(Y; \langle X - P_{\mathcal{V}_m} X, g\rangle). \tag{3}$$

How different is the feature sequence selected with (3) from the one chosen through (2) ? We will give an answer to this question for an example in Section IV.

B. Adaptive Feature Selection and CART

Using (2) or the heuristics (3) noticeably leads to (nonadaptive) variants of LDA. Instead, a data-dependent choice $g_{m+1}(x)$ would depend explicitly on the observations $\{\langle x, g_l\rangle\}_{l=1}^{m}$ that were already collected about x at the previous iterations, in the spirit of CART [8]. In a binary decision tree, each node of the tree is associated to a binary test Q. Starting from the root node, a signal x is classified by descending recursively through the branches of the tree as follows: if the binary test at the current node answers 0, the signal is sent to the left child of the node, else it is sent to the right one. The process ends when the signal reaches a leaf node and is assigned the class label of this leaf.

Here, instead of a binary tree, we have a tree where the number of branches starting from each node is essentially the number of linear features g in the dictionary $\mathcal{D}$. At the root node, a first feature g_1 is chosen using only our prior knowledge, i.e., the distribution $\mathcal{P}(X, Y)$ just as in (1). Then the feature $f_1(x) = \langle x, g_1\rangle$ is observed, and we now know that x belongs to the set of realizations of the random variable X that satisfy $\langle X, g_1\rangle = f_1(x)$. Using this information, a second vector $g_2(x) := \arg\max_{g\in\mathcal{D}} \mathcal{I}(Y; \langle X, g\rangle \,|\, \langle X, g_1\rangle = \langle x, g_1\rangle)$ is chosen, and a second feature $f_2(x) = \langle x, g_2(x)\rangle$ is observed. The process goes on iteratively for $m \geq 1$ with the selection of $g_{m+1}(x)$ as

$$\arg\max_{g\in\mathcal{D}} \mathcal{I}\left(Y; \langle X, g\rangle \,|\, \{\langle X, g_l(x)\rangle = \langle x,\, g_l(x)\rangle\}_{l=1}^{m}\right). \tag{4}$$

While, for obvious reasons, g_1 cannot depend on x, it is natural to wonder whether and how $g_m(x)$ depends on x for $m \geq 2$. Such questions will be studied and answered in the next section on two examples. Such an adaptive feature sequence (AFS) $\{g_1, g_2(x), \ldots, g_m(x), \ldots\}$ has a tree structure which is quite different from the nonadaptive sequences that can be selected through (2) or (3). Indeed, the difference between the adaptive selection rule and the nonadaptive ones is of the same nature as the difference between linear and nonlinear approximation: they correspond to a different approach to the feature selection problem. The adaptive rule can bring some improvement in classification performance when very few features ($M \ll N$) are selected compared to the nonadaptive ones, but it is also more complex to understand, analyze and implement than the nonadaptive one. We dedicate the rest of this paper to its theoretical and practical analysis on a tractable example.

Remark 3: After the observation of each feature $\langle x, g_m(x)\rangle$, it is possible to either select and observe the next one, or to stop and make a decision on the class y of x. Wald's sequential decision theory [10], [23] would help design a stopping criterion to select adaptively the number $M(x)$ of observations, this would correspond to *pruning* the AFS tree. Note that in CART and its variants [24], decision trees are also pruned but for a different reason: the aim is to avoid fitting the training data they were learned from.

IV. Adaptive Features for Two Gaussian Classes

As we already pointed out, the conditional mutual information has in general no simple analytic expression. The explicit computation of the AFS $\{g_1, g_2(x), \ldots, g_m(x), \ldots\}$ from (4) is in general almost impossible, and one has to rely either on Monte Carlo estimation or on a nicely structured distribution $\mathcal{P}(X, Y)$.

In a very specific, well-structured problem, Geman and Jedynak [25] were able to exhibit a simple algorithmic rule ("active testing") that computes a maximizer of a suitably modified version of (4). Later on, Geman and Li [26] dealt with the case of $\mathcal{P}$ a mixture of (multidimensional) Gaussian classes and $\mathcal{D} = \mathcal{B}$ a *given* basis, however they replaced the mutual information with the Hellinger distance, for which they were able to derive an analytic expression. In this section, we state similar results for $\mathcal{P}$ a mixture of two (multidimensional) Gaussian classes, using the mutual information—which is more complex to manipulate than the Hellinger distance—and choosing linear features in a redundant dictionary $\mathcal{D}$.

A. Notations

We consider a mixture of two Gaussian classes $\mathcal{N}(\vec{\mu}_0, \Sigma_0)$ and $\mathcal{N}(\vec{\mu}_1, \Sigma_1)$, with mixture parameter $p_0 \in [0, 1]$. That is to say: the conditional distribution of X under the hypothesis $Y = y$ ($y = 0, 1$) is the multivariate normal distribution $\mathcal{N}(\vec{\mu}_y, \Sigma_y)$ with mean $\vec{\mu}_y$ and covariance Σ_y; the *prior* distribution of the two classes is given by $p_0 = \mathcal{P}(Y = 1)$. We assume that Σ_y has

full rank, $y = 0, 1$. Given a dictionary $\mathcal{D}$, what is the AFS $x \mapsto \{g_1, g_2(x), \ldots, g_m(x), \ldots\}$? The answer depends on whether the two Gaussians share the same mean or covariance matrix or not, and how large the dictionary is.

B. Case where $\Sigma_0 = \Sigma_1 =: \Sigma$

If the Gaussians share the same covariance matrix Σ, but have different means, then the best strategy is to select one vector colinear to the matched filter [27] $\Sigma^{-1}(\vec{\mu}_1 - \vec{\mu}_0)$, if such a vector is contained in the dictionary; otherwise the best sequence can be computed with an algorithm similar to an orthonormal matching pursuit [28]. This is our first theorem, which is similar to a result of Li [26, p. 44] with the Hellinger distance, $\Sigma = \mathrm{Id}$ and $\mathcal{D}$ an orthonormal basis:

Theorem 1: Assume $\Sigma_1 = \Sigma_0 = \Sigma$ and $\vec{\mu}_1 \neq \vec{\mu}_0$, and let $\mathcal{D}$ be *any* dictionary. A sequence $\{x \mapsto g_m(x)\}_{m=1}^N$ is an AFS if, and only if, for $0 \leq m \leq N-1$

$$g_{m+1} := \arg\max_{g \in \mathcal{D}} |\langle \Sigma^{-1}(\vec{\mu}_1 - \vec{\mu}_0), R_m g / \|R_m g\|_\Sigma \rangle_\Sigma| \quad (5)$$

where $\langle \cdot, \cdot \rangle_\Sigma := \langle \cdot, \Sigma \cdot \rangle$ defines a weighted inner product on $\mathbb{R}^N$, $\| \cdot \|_\Sigma$ is the associated weighted Euclidian norm and R_m is the orthonormal projector (with respect to this inner product) onto the orthonormal complement of $\mathcal{V}_m := \mathrm{span}\{g_l, 1 \leq l \leq m\}$.

The proof is in the Appendix. In this case, the discrimination problem is indeed very close to an approximation problem, and the AFS coincides with the features predicted by Fisher's LDA that are *independent* of the test data x.

Looking at the proof would also show that here, selecting the features with the simplified criterion (3) would yield the same nonadaptive feature sequence, and this may explain the good behavior of nonadaptive feature selection strategies [18], [21], [22] that use linear projections instead of conditional mutual information estimation.

C. Case where $\vec{\mu}_0 = \vec{\mu}_1 =: \vec{\mu}$

The situation becomes completely different if the two Gaussians share the same mean but have different covariance matrices. In this case, Theorem 2 will show that the AFS among a huge dictionary—containing all possible unit vectors in $\mathbb{R}^N$—is a sequence of eigenvectors of $\Sigma_0^{-1}\Sigma_1$ (which was predicted by LDA using the Bhattacharyya discriminant measure [11, pp. 456–457]). However, and this is the main point of this paper, Theorem 2 will show that the order of the sequence is now *data-dependent*, i.e., it depends on x.

Theorem 2: Assume that $\Sigma_1 \neq \Sigma_0$ and $\vec{\mu}_1 = \vec{\mu}_0 = \vec{\mu}$. Let $\mathcal{D}$ be the whole unit sphere of $\mathbb{R}^N$ and $\{v_k\}_{k=1}^N$ a basis of unit eigenvectors of $\Sigma_0^{-1}\Sigma_1$. There exists an (adaptive) permutation $\{x \mapsto k_m(x)\}_{m=1}^N$ of $\{1, 2, \ldots, N\}$ such that $\{x \mapsto g_m(x)\}_{m=1}^N$ is an AFS if, and only if, for all x and $1 \leq m \leq N$

$$\mathrm{span}\{g_l(x), 1 \leq l \leq m\} = \mathrm{span}\{v_{k_l(x)}, 1 \leq l \leq m\}. \quad (6)$$

The proof is in the Appendix. In this case, the AFS *potentially* depends on the test data x, in that the order $\{k_m(x)\}_{m=1}^N$ of the observations $\langle x, v_k \rangle$ may depend on x. The following definition will be useful to characterize the dependence.

Definition 1: Let $\{\lambda_k\}_{k=1}^N$ be the eigenvalues of $\Sigma_0^{-1}\Sigma_1$ associated to the unit eigenvectors $\{v_k\}_{k=1}^N$, and for $0 \leq m \leq N-1$, let $A_m(x), B_m(x) \in \{1, \ldots, N\}$ be the indexes of the extremal remaining eigenvalues after m steps

$$\lambda_{A_m(x)} = \min\{\lambda_k, k \notin \{k_l(x)\}_{l=1}^m\} \quad (7)$$

$$\lambda_{B_m(x)} = \max\{\lambda_k, k \notin \{k_l(x)\}_{l=1}^m\}. \quad (8)$$

For $m = 0$, A_0, and B_0 are actually independent of x. We already noticed that the first feature vector g_1 cannot depend on x, hence, k_1, A_1, and B_1 are independent of x. For $m \geq 2$, $k_m(x)$ (and by consequence $A_m(x)$ and $B_m(x)$) can generally depend on x. The reason for the previous definition is the following theorem, which gives more information on the nature of this dependence. Let us denote

$$p_m(x) := \mathcal{P}(Y = 1 \,|\, \{\langle X, g_l(x) \rangle = \langle x, g_l(x) \rangle\}_{l=1}^m) \quad (9)$$

the *a posteriori* probability that $Y = 1$ after m observations.

Theorem 3: Let $a(\lambda) := \lambda - \log\lambda$ with 'log' the natural logarithm and define

$$C_0(\lambda_A, \lambda_B) := a(\lambda_B) - a(\lambda_A) \quad (10)$$

$$C_1(\lambda_A, \lambda_B) := a\left(\lambda_B^{-1}\right) - a\left(\lambda_A^{-1}\right). \quad (11)$$

The adaptive order $\{k_m(x)\}$ has the following properties:

a) for all x and m, $k_{m+1}(x) \in \{A_m(x), B_m(x)\}$;
b) if $p_m(x) \approx 1/2$ then $\lambda_{k_{m+1}(x)} = \max(1/\lambda_{A_m(x)}, \lambda_{B_m(x)})$;
c) if $1 \leq \lambda_{A_m(x)}$ then $k_{m+1}(x) = B_m(x)$, and this goes recursively because $A_{m+1}(x) = A_m(x)$;
d) if $\lambda_{B_m(x)} \leq 1$ then $k_{m+1}(x) = A_m(x)$, and this goes recursively because $B_{m+1}(x) = B_m(x)$;
e) if $C_0(\lambda_A, \lambda_B) > 0$ and $C_1(\lambda_A, \lambda_B) < 0$, then the set $\{x \in \mathbb{R}^N, A_m(x) = A, B_m(x) = B\}$ can be partitioned into two nontrivial subsets where $k_{m+1}(x) = A_m(x)$ and $k_{m+1}(x) = B_m(x)$, respectively.

The proof is in the Appendix. It is rather technical and does not provide an explicit decision rule to select $k_{m+1}(x)$ when $C_0(\lambda_A, \lambda_B) < 0$ and $C_1(\lambda_A, \lambda_B) > 0$. One should also note that Theorem 3 only gives sufficient conditions for $\{k_m(x)\}$ to depend/not to depend on x, but there are cases where we do not know whether there is a dependence or not. The analysis is indeed quite technical and it is not clear whether one can find necessary and sufficient conditions and explicit decision rules.

A couple of simple examples can illustrate the implications of Theorem 3.

Example 1: Assume $\lambda_{A_0} \geq 1$, i.e., all the eigenvalues of $\Sigma_0^{-1}\Sigma_1$ are larger than one. Then, by Theorem 3-(c), $\lambda_{A_m(x)} \geq 1$ for all m, and the AFS is independent of x. The order of observations is actually that of the decreasing rearrangement of the eigenvalues of $\Sigma_0^{-1}\Sigma_1$.

Example 2: A simple function study shows that for any $\lambda_A \leq 1 \leq \lambda_B$, either $C_0(\lambda_A, \lambda_B) \geq 0$ and $C_1(\lambda_A, \lambda_B) \geq 0$, or $C_1(\lambda_A, \lambda_B) \leq 0$ and $C_1(\lambda_A, \lambda_B) \leq 0$, or $C_0(\lambda_A, \lambda_B) < 0$ and $C_1(\lambda_A, \lambda_B) > 0$. On Fig. 1, the white region corresponds to the couples $(\log\lambda_A, \log\lambda_B)$ where $C_0(\lambda_A, \lambda_B) > 0$ and $C_1(\lambda_A, \lambda_B) < 0$; the quadrants that are not displayed correspond to where either $\lambda_A \geq 1$ or $\lambda_B \leq 1$; the grey regions corre-

spond to the configurations ($C_0(\lambda_A, \lambda_B) \leq 0$ or $C_1(\lambda_A, \lambda_B) \geq 0$) not dealt with in Theorem 3. Take $N = 7$ and $\lambda_1 < \cdots < \lambda_4 < 1 < \lambda_5 < \cdots < \lambda_7$, with $\lambda_1 < 1/\lambda_7$. The circles correspond to the possible pairs $(\log \lambda_{A_m(x)}, \log \lambda_{B_m(x)})$, and close to each circle is the corresponding value of m.

- For $m = 0, A_0 = 1$, and $B_0 = 7$. Because $\lambda_1 < 1/\lambda_7$, we know that $k_1 = 1$, and it follows that $A_1 = 2$ and $B_1 = 7$. This is illustrated by an arrow that joins $(\log \lambda_{A_0}, \log \lambda_{B_0})$ to $(\log \lambda_{A_1}, \log \lambda_{B_1})$, with the indication $k_1 = 1$.
- For $m = 1$, as the pair $(\log \lambda_{A_1}, \log \lambda_{B_1})$ is in the white region, we deduce from Theorem 3-(e) that (depending on x) we may either have $k_2(x) = 2$ or $k_2(x) = 7$, which is depicted on Fig. 1 by the two corresponding plain arrows.
- At any later stage of the selection process, the behavior is the same, provided that the pair $(\log \lambda_{A_m(x)}, \log \lambda_{B_m(x)})$ is in the white region.
- When the process leads us to leave the white region (for example by choosing $k_4(x) = 5$ from the situation $A_3(x) = 2, B_3(x) = 5$, as depicted on Fig. 1 by the dashed arrow), this results either in $\lambda_{A_m(x)} \geq 1$ or $\lambda_{B_m(x)} \leq 1$ (on the example, $\lambda_{B_5(x)} \leq 1$). Afterwards, the remaining eigenvectors are observed in the order of decreasing magnitude of $(\lambda_k + 1/\lambda_k)$.

By comparison, LDA based on the Bhattacharyya discriminant measure [11, pp. 456–457] on this example would select the very same features, but in a fixed order corresponding to the decreasing magnitude of $(\lambda_k + 1/\lambda_k) : \{k_1 = 1; k_2 = 7; k_3 = 2; k_4 = 6; k_5 = 3; k_6 = 5; k_7 = 4\}$.

V. Numerical Experiments

In this section, we provide numerical evidence that ADA can outperform LDA. First, we describe a feature selection and classification algorithm that implements the ADA strategy corresponding to the theory developed so far. Then, we discuss the experimental setup and the results of the numerical experiments.

A. ADA Algorithm

We consider a restricted model with two Gaussian classes where Theorem 2 and 3 take the simplest form: we assume centered Gaussian classes ($\vec{\mu}_0 = \vec{\mu}_1 = 0$) and diagonal covariance matrices $\Sigma_i = \mathrm{diag}(\sigma^2_{i,k})_{k=1}^N$. It follows that $\Sigma_0^{-1}\Sigma_1 = \mathrm{diag}(\lambda_k)_{k=1}^N$ with $\lambda_k = \sigma^2_{1,k}/\sigma^2_{0,k}$, and the AFS simply corresponds to projections onto the canonical coordinates, i.e., $\langle x, g_{k_m(x)} \rangle = x[k_m(x)]$. We denote $I(\eta, \lambda, p)$ the mutual information $\mathcal{I}(Y; Z)$ between a class variable Y with $p = \mathcal{P}(Y = 1)$ and a one-dimensional (1-D) observation Z drawn according to $\mathcal{N}(\eta, 1)$ if $Y = 0$ and $\mathcal{N}(0, \lambda)$ if $Y = 1$.

The adaptive order of observations $\{k_m(x)\}_{m=1}^N$ for an input $x = (x[1], \ldots, x[N])$ is (theoretically) computed as follows.

Step 1) Initialization: Set $m = 0, A_0 = \min_k \lambda_k, B_0 = \max_k \lambda_k, p_0 = \mathcal{P}(Y = 1)$, and $S_0 := \log(\mathcal{P}(Y = 1)/\mathcal{P}(Y = 0)) = \log p_0/(1 - p_0)$.

Step 2) Adaptive selection:

$$k_{m+1}(x) := \arg \max_{k \in \{A_m(x), B_m(x)\}} I(0, \lambda_k, p_m(x))$$

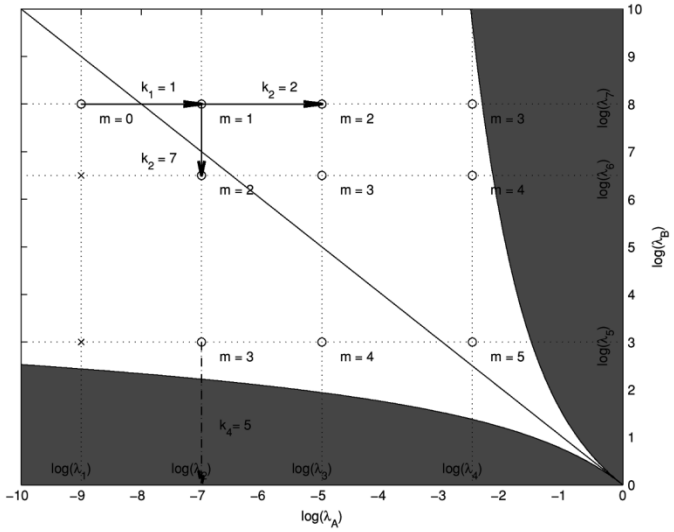

Fig. 1. White region corresponds to the set of values $(\log \lambda_A, \log \lambda_B)$ where $C_0(\lambda_A, \lambda_B) > 0$ and $C_1(\lambda_A, \lambda_B) < 0$. It contains the "line" $\lambda_B = 1/\lambda_A$. The lower-left grey region corresponds to $C_0(\lambda_A, \lambda_B) \leq 0$ while the upper-right one corresponds to $C_1(\lambda_A, \lambda_B) \geq 0$, which are the cases not dealt with in Theorem 3.

Step 3) Observation: Observe $\langle x\, g_{k_{m+1}(x)} \rangle = x[k_{m+1}(x)]$;

Step 4) Update: Update $A_{m+1}(x), B_{m+1}(x)$ according to (7) and (8), and compute the new *score*[1]

$$\begin{aligned} S_{m+1}(x) &:= \log \frac{p_{m+1}(x)}{1 - p_{m+1}(x)} \\ &= S_m(x) - \frac{1}{2} \log \lambda_{k_{m+1}(x)} \\ &\quad + \frac{|x[k_{m+1}(x)]|^2}{2\sigma^2_{1,k_{m+1}(x)}} \left(\lambda_{k_{m+1}(x)} - 1\right). \end{aligned} \tag{12}$$

Step 5) Either:
- increment m and go back to Step 2; or
- stop and make a decision

$$\hat{Y}^{\mathrm{ADA}}_{m+1}(x) := \begin{cases} 1, & \text{if } S_{m+1}(x) \geq 0 \\ 0, & \text{if } S_{m+1}(x) < 0 \end{cases}. \tag{13}$$

The adaptive selection step (Step 2) is not explicit but could be implemented using a table of sampled numerical values of $I(0, \lambda, p)$. Instead, we propose to rely on Theorem 3 (see also Lemma 2 in the Appendix) and replace Step 2 with a modified adaptive selection rule.

Step 2bis) Semiheuristic adaptive selection:

a) if $\lambda_{A_m(x)} \geq 1$, then $k_{m+1}(x) := B_m(x)$;
b) else if $\lambda_{B_m(x)} \leq 1$, then $k_{m+1}(x) := A_m(x)$;
c) else if $C_0(\lambda_{A_m(x)}, \lambda_{B_m(x)}) \leq 0$, then $k_{m+1}(x) := A_m(x)$;
d) else if $C_1(\lambda_{A_m(x)}, \lambda_{B_m(x)}) \geq 0$, then $k_{m+1}(x) := B_m(x)$;
else compute the heuristic threshold

$$\theta_m(x) := \log \frac{C_0\left(\lambda_{A_m(x)}, \lambda_{B_m(x)}\right)}{-C_1\left(\lambda_{A_m(x)}, \lambda_{B_m(x)}\right)} \tag{14}$$

[1]The expression of the update rule (12) is justified in the Appendix.

on the log-likelihood ratio to get

e.1) if $S_m(x) \geq \theta_m(x)$, then $k_{m+1}(x) := A_m(x)$;

e.2) if $S_m(x) < \theta_m(x)$, then $k_{m+1}(x) := B_m(x)$.

The rationale behind Step 2bis is as follows.

- In case (a) [resp. (b)] it gives the optimal choice as shown by Theorem 3-(c) [resp Theorem 3-(d)].
- In case (c) [resp. (d)], when $p_m(x)$ is sufficiently close to 0 or to 1, Step 2bis gives the best choice according to Lemma 2. Heuristically we keep the same choice for any $0 \leq p_m(x) \leq 1$.
- In case (d), Lemma 2 shows that Step 2bis also gives the best choice whenever $S_m(x)$ is big enough ($p_m(x)$ close enough to 1) or small enough ($p_m(x)$ close enough to 0). Heuristically, we use a threshold to specify what is "big enough."

Though it is partially heuristic, the previous approach is compatible with Theorem 3 and Lemma 2 and provides *some* adaptive order of observations which can outperform LDA, as shown by our experiments in the following.

B. Experimental Setup

We conducted experiments in a very simple setting where

$$\sigma^2_{0,k} = \begin{cases} c^{-1}, & 1 \leq k \leq N/2 \\ c, & N/2+1 \leq k \leq N \end{cases} \tag{15}$$

$$\sigma^2_{1,k} = \begin{cases} c, & 1 \leq k \leq N/2 \\ c^{-1}, & N/2+1 \leq k \leq N \end{cases} \tag{16}$$

with $c > 1$ some constant, and consequently

$$\lambda_k = \begin{cases} c^2, & 1 \leq k \leq N/2 \\ c^{-2}, & N/2+1 \leq k \leq N \end{cases}. \tag{17}$$

As already noticed, on such a discrimimation problem, LDA with the Bhattacharyya discriminant measure [11, pp. 456–457] selects the eigenvectors of $\Sigma_0^{-1}\Sigma_1$ with any fixed order $\{k_m^{\mathrm{LDA}}\}_{m=1}^N$ of observations such that $\{\lambda_{k_m^{\mathrm{LDA}}} + 1/\lambda_{k_m^{\mathrm{LDA}}}\}_{m=1}^N$ is nonincreasing. By observing the components in such an order, one can iteratively update a log-likelihood ratio $S_m^{\mathrm{LDA}}(x)$ and, at any step, make a decision $\hat{Y}_{m+1}^{\mathrm{LDA}}(x)$ using the analogue of (12) and (13). In our case, since $\lambda_k + 1/\lambda_k = c^2 + 1/c^2$ is constant, there is no preferred order and we used a random order.

In the ADA framework, at any step where $\lambda_{A_m(x)} = c^{-2} < 1$ and $\lambda_{B_m(x)} = c^2 > 1$, we have

$$\begin{aligned} C_0(\lambda_A, \lambda_B) = -C_1(\lambda_A, \lambda_B) &= a(c^2) - a(1/c^2) \\ &= c^2 - 1/c^2 - 4\log c > 0 \end{aligned}$$

hence, $\theta_m(x) = 0$ and the rule to select the next component to observe is $k_{m+1}(x) = A_m(x)$ iff $S_m(x) \geq 0$, until the remaining available eigen-values are either all smaller or all larger than one (note that this can only happen when $m \geq N/2$).

After m observations, the two strategies lead to two estimators $\hat{Y}_m^{\mathrm{ADA}}(x)$ and $\hat{Y}_m^{\mathrm{LDA}}(x)$. Our experiments consisted in estimating the probability of error $\mathcal{P}(\hat{Y}_m(x) \neq Y)$ of these estimators for $0 \leq m \leq N$. To do that we draw L test samples $\{(x_l, y_l)\}_{l=1}^L$ according to the Gaussian mixture $\mathcal{P}(X, Y) = (1/2)\mathcal{N}(0, \Sigma_0) + (1/2)\mathcal{N}(0, \Sigma_1)$, applied the feature selection strategy and the corresponding classifier (ADA, or LDA with a different random order $\{k_m\}_{m=1}^N$ for each index l) to each x_l, and measured the average number of errors.

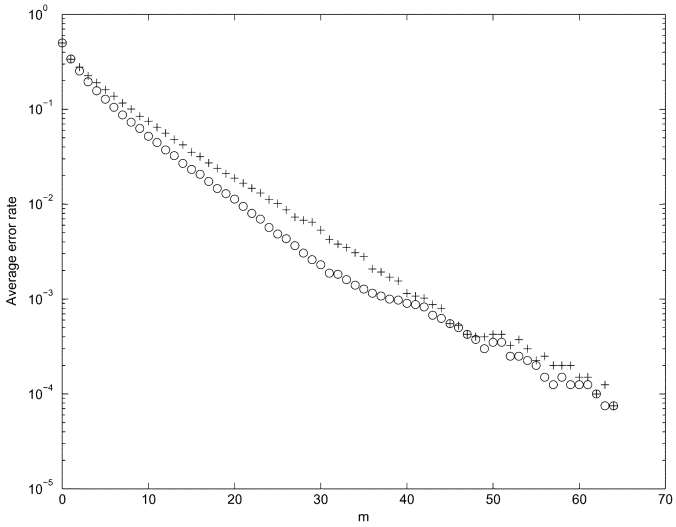

Fig. 2. Average error rates (in logarithmic scale) of the fixed feature sequence classifier $\hat{Y}_m^{\mathrm{LDA}}(x)$ ('+' signs) and the AFS one $\hat{Y}_m^{\mathrm{ADA}}(x)$ (circles) for an increasing number m of observations.

C. Numerical Results

We ran the experiment in dimension $N = 64$, with $c = 2$, using $L = 40\,000$ test samples (x_l, y_l) of the mixture model and an equal *a priori* probability $p_0 = \mathcal{P}(Y = 1) = 1/2$. Fig. 2 compares the average error rates (in a logarithmic scale) of the LDA and ADA strategies for increasing values of m.

For $m = 0$ (the signal is not observed at all) the error rate is $1/2$ since the decision is made completely at random. In both strategies, the first observation k_1^{LDA} and $k_1^{\mathrm{ADA}}(x)$ does not depend on x, which explains why the error rates also coincide for $m = 1$. For $m \geq 2$, ADA provides in few steps quite a smaller error rate that LDA. For small m (typically for $m \leq N/2$, i.e., when the adaptive strategy is possible because $\lambda_{A_m(x)} = c^{-2} < 1$ and $\lambda_{B_m(x)} = c^2 > 1$) the error rates of both estimators seem to decrease linearly in a logarithmic scale, however the slope of this linear decrease is clearly stronger for the ADA strategy. When m becomes closer to the dimension of the full feature space N, the LDA and ADA estimators both converge to the "full" Bayesian estimator obtained with the complete knowledge of x, so the gap between their performances decreases and they have equal performance for $m = N$. One should notice that when the error rate becomes small, the total number of errors over the L test samples become a small integer and the estimate of the error rates becomes less reliable, which explains the irregular behavior of the curves for large m.

This example shows that one can obtain the same classification performance with less observations using ADA instead of LDA. For example, one can read on the figure that an error rate below one percent (10^{-2}) is achieved after $m = 21$ observations with ADA and $m = 26$ observations with LDA. Since the only overhead in the algorithmic implementation of ADA versus LDA consists only in a few tests, this means that the same classification performance can be achieved with about 20% fewer

computations. The adaptive strategy is likely to display an even higher gain in performance compared to the passive one when the dimension of the data becomes larger.

D. Perspective of Application

Before concluding, let us briefly mention an example where ADA with the model of mixture of centered Gaussians might be useful in a practical setting. The Wiener filter is a well known tool to perform signal denoising, however it is based on a (generally centered) Gaussian model of the signal of interest and of the noise. Real signals are never purely Gaussian, and in some cases even the noise is not Gaussian: this is, e.g., the case when the signal of interest is speech and the noise is some background music. However, it is often reasonable to model the signal as locally Gaussian, with power density spectrum that varies with time. Thus, denoising becomes feasible provided that the proper (centered) Gaussian model is used at each time: the problem becomes the identification of the most likely couple of power spectrums (of the signal and of the noise). Such an adaptive Wiener filter approach has been proposed for single channel signal separation [29] and our ADA technique could be used to estimate the most likely Gaussian model on each time frame with a fast algorithm, so as to denoise the signal with a low numerical complexity.

VI. Conclusion

In this paper, we have explored the theory of adaptive projections for feature extraction and classification. Combining the spirit of nonlinear approximations, projection pursuit and CART, we have proposed a formalism for ADA.

In the case of two Gaussian classes entirely characterized by their mean (i.e., with $\Sigma_0 = \Sigma_1$), we showed that ADA is a nonadaptive variant of LDA that corresponds to an approximation strategy similar to the orthogonal matching pursuit. Our main point is that, in the case of classes entirely characterized by their covariance structure, ADA is actually data-dependent. We provide numerical evidence that, with a small number of tests, a more reliable classification can be obtained with the data-dependent strategy than with LDA.

The question of the actual performance of ADA compared to LDA is of course fundamental for applications: it should be carefully studied in a proper experimental setup that is not the purpose of this contribution. We can nevertheless mention some issues that will certainly come up for such a comparison. First, we have assumed Gaussian models with known covariance matrices (Σ_0, Σ_1). In practice, the matrices must be estimated from training data that are not necessarily Gaussian: how robust is ADA feature selection and classification to modeling error and/or inaccurate estimates of the eigenstructure of $\Sigma_0^{-1}\Sigma_1$? Is it robust enough that we can use fast computational harmonic analysis techniques to select approximate eigenvectors in a finite dictionary (e.g., wave-packets [3])? Can we observe on some real data as substantial an improvement in classification performance (for fixed number of observations) as in our numerical experiments with synthetic data? Another related question is how to design a suitable modification of the sequential probability ratio test [23], [10] to use a data-dependent number $m(x)$ of steps of ADA before stopping and making a decision.

Appendix

VII. Proof of the Theorems

It is a good exercise of probability (we leave it to the reader) to check, for any unit vectors $\{g_l\}_{l=1}^m, m \geq 1$, and g, the following algebraic expressions of conditional expectations:

$$\begin{aligned}
&\mathbb{E}[g_1, \ldots, g_m](g) := \mathbb{E}\left\{\langle X, g\rangle \,|\, Y, \{\langle X, g_l\rangle\}_{l=1}^m\right\} \\
&\quad \overset{a.s.}{=} \langle X, g\rangle - \langle X - \vec{\mu}_Y, R_{\mathcal{V}_m, \Sigma_Y} g\rangle \\
&\mathbb{E}\left\{(\langle X, g\rangle - \mathbb{E}[g_1, \ldots, g_m](g))^2 \,|\, Y, \{\langle X, g_l\rangle\}_{l=1}^m\right\} \\
&\quad \overset{a.s.}{=} \|R_{\mathcal{V}_m, \Sigma_Y} g\|_{\Sigma_Y}^2
\end{aligned}$$

where $R_{\mathcal{V}_m, \Sigma_y}$ is the projector perpendicular to $\mathcal{V}_m := \operatorname{span}\{g_l\}_{l=1}^m$ with respect to the inner product $\langle \cdot, \cdot \rangle_{\Sigma_y} := \langle \cdot, \Sigma_y \cdot \rangle$, i.e., the projector onto $\mathcal{V}_m^{\perp_{\Sigma_y}} = (\Sigma_y \mathcal{V}_m)^\perp = \Sigma_y^{-1}\mathcal{V}_m^\perp$ with kernel $\mathcal{V}_m$. The expressions are also true for $m = 0$ with the convention $\operatorname{span}\{\emptyset\} = \{0\}$, i.e., $R_0 = \mathrm{Id}$.

For any vector $x \in \mathbb{R}^N$, let $\{g_1, g_2(x), \ldots, g_N(x)\}$ an AFS, and for $m \geq 0$

$$\mathcal{V}_m(x) := \operatorname{span}\{g_l(x)\}_{l=1}^m. \tag{18}$$

It follows from the previous expressions that for any $x \in \mathbb{R}^N$ and $y \in \{0, 1\}$, the conditional distribution of $\langle X, g\rangle$ under the assumption that $\{\langle X, g_l(x)\rangle = \langle x, g_l(x)\rangle\}_{l=1}^m$ and $Y = y$ is a one-dimensional Gaussian $\mathcal{N}(\mu_{m,x,y}(g), \sigma_{m,x,y}^2(g))$ with

$$\mu_{m,x,y}(g) := \langle x, g\rangle - \left\langle x - \vec{\mu}_y, R_{\mathcal{V}_m(x), \Sigma_y} g\right\rangle \tag{19}$$

$$\sigma_{m,x,y}^2(g) := \left\|R_{\mathcal{V}_m(x), \Sigma_y} g\right\|_{\Sigma_y}^2. \tag{20}$$

Using the invariance of mutual information [20] with respect to translations and dilations of X, it is easy to show that the property (4) of the AFS is equivalent to selecting $g_{m+1}(x)$ as

$$\arg\max_{g \in \mathcal{D}} I\left(\left|\frac{\mu_{m,x,1}(g) - \mu_{m,x,0}(g)}{\sigma_{m,x,0}(g)}\right|, \frac{\sigma_{m,x,1}^2(g)}{\sigma_{m,x,0}^2(g)}, p_m(x)\right) \tag{21}$$

where $I(\eta, \lambda, p)$ denotes the mutual information $\mathcal{I}(Y; Z)$ between a class variable Y with $p = \mathcal{P}(Y = 1)$ and a 1-D observation Z drawn according to $\mathcal{N}(\eta, 1)$ if $Y = 0$ and $\mathcal{N}(0, \lambda)$ if $Y = 1$, with $0 \leq p \leq 1, \eta \in \mathbb{R}, 0 < \lambda < \infty$.

The proof of our theorems relies on the variations of $\lambda \mapsto I(0, \lambda, p)$ and $\eta \mapsto I(\eta, 1, p)$. They are summarized in the following lemma, which we prove in the Appendix.

Lemma 1: For any value of $0 < p < 1$:

a) the even function $\eta \mapsto I(\eta, 1, p)$ is increasing with $|\eta|$.
b) the function $\lambda \mapsto I(0, \lambda, p)$ is strictly decreasing on $(0, 1]$ and strictly increasing on $[1, \infty)$.

Let us now proceed to the proof of the theorems. In order to simplify the notations, we will generally not write the dependence of $\mathcal{V}_m, R_{\mathcal{V}_m, \Sigma_y}, \mu_{m,y}(g)$, and $\sigma_{m,y}^2(g)$ on x.

Proof of Theorem 1: As $\Sigma_1 = \Sigma_0 = \Sigma$, the projector $R_m := R_{\mathcal{V}_m,\Sigma_1} = R_{\mathcal{V}_m,\Sigma_0}$ is independent of y and (20) can be written as $\sigma^2_{m,y}(g) = \langle R_m g, \Sigma R_m g\rangle$. As a result $\sigma^2_{m,1}/\sigma^2_{m,0} = 1$ for all g, and (21) combined with Lemma 1-(a) shows that

$$g_{m+1}(x) = \arg\max_{g\in\mathcal{D}} \left| \frac{\mu_{m,1}(g) - \mu_{m,0}(g)}{\sigma_{m,0}(g)} \right|.$$

Using (19) and (20) this becomes

$$g_{m+1}(x) = \arg\max_{g\in\mathcal{D}} \left| \left\langle \Sigma^{-1}(\vec{\mu}_1 - \vec{\mu}_0), \frac{R_m g}{\|R_m g\|_\Sigma} \right\rangle_\Sigma \right| \quad (22)$$

where the weighted inner product $\langle\cdot,\cdot\rangle_\Sigma = \langle\cdot,\Sigma\cdot\rangle$ and its associated weighted norm $\|\cdot\|_\Sigma$ define an Euclidian structure on $\mathbb{R}^N$. It follows by induction that the AFS $\{g_m\}_{m=1}^N$ is nonadaptive: let us show that (22) corresponds to a variant of the orthogonal matching pursuit [19], [28] on the signal $\Sigma^{-1}(\vec{\mu}_1 - \vec{\mu}_0)$.

With respect to the new Euclidian structure, the Gram-Schmidt orthonormalization of $\{g_m\}_{m=1}^N$ is $\{u_m\}_{m=1}^N$ where $u_m := R_{m-1}g_m/\|R_{m-1}g_m\|_\Sigma$, and $P_m = \mathrm{Id} - R_m$ is the orthonormal projector onto $\mathcal{V}_m$, hence

$$\|P_{m+1}\Sigma^{-1}(\vec{\mu}_1 - \vec{\mu}_0)\|^2_\Sigma = \sum_{l=1}^{m+1} \left|\langle \Sigma^{-1}(\vec{\mu}_1 - \vec{\mu}_0), u_l\rangle^2_\Sigma\right|.$$

and (22) corresponds to choosing g_{m+1} so as to maximize the increase

$$\|P_{m+1}\Sigma^{-1}(\vec{\mu}_1 - \vec{\mu}_0)\|^2_\Sigma - \|P_m\Sigma^{-1}(\vec{\mu}_1 - \vec{\mu}_0)\|^2_\Sigma,$$

in the (weighted) energy of the projection onto $\mathcal{V}_{m+1}$. □

Proof of Theorem 2: Let $\{v_k\}_{k=1}^N$ be a basis of unit eigenvectors for $\Sigma_0^{-1}\Sigma_1 : \Sigma_0^{-1}\Sigma_1 v_k = \lambda_k v_k$. We shall prove by induction that there exists $\{k_m(x)\}_{m=1}^N$ such that for $0 \le m \le N$

$$\mathcal{V}_m = \mathrm{span}\{v_{k_l(x)}, 1 \le l \le m\}. \quad (23)$$

The relation is true for $m = 0$ with the convention $\mathrm{span}\{\emptyset\} = \{0\}$. Let us show that if (23) holds for m, then it is also true for $m+1$. The induction hypothesis implies that $\Sigma_0^{-1}\Sigma_1\mathcal{V}_m = \mathcal{V}_m$, i.e.

$$\Sigma_1\mathcal{V}_m = \Sigma_0\mathcal{V}_m. \quad (24)$$

Hence, the two projectors $R_m := R_{\mathcal{V}_m,\Sigma_1} = R_{\mathcal{V}_m,\Sigma_0}$ onto $(\Sigma_1\mathcal{V}_m)^\perp = (\Sigma_0\mathcal{V}_m)^\perp$ with kernel $\mathcal{V}_m$ are equal. Using (19) we get $\mu_{m,1}(g) - \mu_{m,0}(g) = 0$ for all g and (20) combined with Lemma 1-(b) and (20) shows that $g_{m+1}(x)$ corresponds to an extremum of

$$\sigma^2_{m,1}(g)/\sigma^2_{m,0}(g) = \langle R_m g, \Sigma_1 R_m g\rangle / \langle R_m g, \Sigma_0 R_m g\rangle.$$

Using Lagrange multipliers, such an extremum is obtained when, for some λ

$$R^*_m\Sigma_1 R_m g = \lambda R^*_m \Sigma_0 R_m g. \quad (25)$$

The notation $R^\star_m$ stands for the adjoint of R_m in the standard Euclidian structure. Linear algebra shows that condition (25) is equivalent to $(\Sigma_1 - \lambda\Sigma_0)R_m g \in (\mathrm{Im}R_m)^\perp = \Sigma_y\mathcal{V}_m, y = 0, 1$, that is to say $(\Sigma_0^{-1}\Sigma_1 - \lambda\mathrm{Id})R_m g \in \mathcal{V}_m = \mathrm{Ker}R_m$.

From (24), we know that the projector R_m commutes with $\Sigma_0^{-1}\Sigma_1$, because its range $\Sigma_1^{-1}\mathcal{V}_m^\perp$ and kernel $\mathcal{V}_m$ are stable under $\Sigma_0^{-1}\Sigma_1$. Thus, (25) $\Leftrightarrow R_m(\Sigma_0^{-1}\Sigma_1 - \lambda\mathrm{Id})g \in \mathrm{Ker}R_m \Leftrightarrow R_m(\Sigma_0^{-1}\Sigma_1 - \lambda\mathrm{Id})g = 0 \Leftrightarrow (\Sigma_0^{-1}\Sigma_1 - \lambda\mathrm{Id})R_m g = 0 \Leftrightarrow R_m g$ is either zero or an eigenvector of $\Sigma_0^{-1}\Sigma_1$. But $R_m g$ cannot be zero, for it would mean $g \in \mathcal{V}_m$, and such a g cannot bring *any* additional information on the class Y. It follows that $R_m g_{m+1}(x)$ is an eigenvector of $\Sigma_0^{-1}\Sigma_1$. As a result

$$g_{m+1}(x) = \underbrace{(\mathrm{Id} - R_m)g_{m+1}(x)}_{\in\mathcal{V}_m} + \underbrace{R_m g_{m+1}(x)}_{\in\mathrm{span}\{v_{k_{m+1}(x)}\}}$$

and (23) is true at step $m+1$. □

Proof of Theorem 3: Let $m \ge 0$. Clearly $k_{m+1}(x) \notin \{k_l(x)\}_{l=1}^m$ because an already observed feature cannot bring any additional information. Moreover for $k \notin \{k_l(x)\}_{l=1}^m, \sigma^2_{m,1}[v_k]/\sigma^2_{m,0}[v_k] = \lambda_k$, hence

$$k_{m+1}(x) = \arg\max_{k\notin\{k_l(x)\}_{l=1}^m} I(0, \lambda_k, p_m(x))$$

Using Lemma 1-(b) and the notations from Definition 1 we immediately get that

$$k_{m+1}(x) = \arg\max_{k\in\{A_m(x), B_m(x)\}} I(0, \lambda_k, p_m(x)) \quad (26)$$

which gives Theorem 3-(a). Lemma 1-(b) shows that $\lambda_{A_m(x)} \ge 1$ (resp. $\lambda_{B_m(x)} \le 1$) is a sufficient condition for the maximization (26) to be independent of x: for example if $\lambda_{B_m(x)} \le 1$ then $I(0, \lambda_{A_m(x)}, p) > I(0, \lambda_{B_m(x)}, p)$ for every value of p, and $k_{m+1}(x) = A_m(x)$ is the only possible choice. The statements Theorem 3-(c)-(d) immediately follow. It is easy to show by a change of variables that $I(0, \lambda, p) = I(0, 1/\lambda, 1-p)$, hence, $I(0, \lambda, 1/2) = I(0, 1/\lambda, 1/2)$ and we get Theorem 3-(b) by Lemma 1-(b). The following lemma, proved in the Appendix, shows that the maximization (26) can depend on the value of x.

Lemme 2: Let $0 < \lambda_A < 1 < \lambda_B$, and $a(\lambda) := \lambda - \log\lambda$ with 'log' the natural logarithm. For p close to 0

$$I(0, \lambda_B, p) - I(0, \lambda_A, p) \sim \frac{p}{2} \cdot C_0(\lambda_A, \lambda_B) \quad (27)$$

while for p close to 1

$$I(0, \lambda_B, p) - I(0, \lambda_A, p) \sim \frac{1-p}{2} \cdot C_1(\lambda_A, \lambda_B). \quad (28)$$

Assume that after m observations $\{\langle X, v_{k_l(x)}\rangle = \langle x, v_{k_l(x)}\rangle\}_{l=1}^m, (m \ge 1)$, the extremal remaining eigenvalues $\lambda_{A_m(x)}$ and $\lambda_{B_m(x)}$ satisfy $C_0(\lambda_A, \lambda_B) > 0$ and $C_1(\lambda_A, \lambda_B) < 0$. Then, if $p_m(x)$ is close enough to 0 (which is true for some values of x)

$$I(0, \lambda_{B_m(x)}, p_m(x)) > I(0, \lambda_{A_m(x)}, p_m(x))$$

while if $p_m(x)$ is close enough to 1 (which holds for some other values of x),

$$I(0,\lambda_{B_m(x)},p_m(x)) < I(0,\lambda_{A_m(x)},p_m(x)).$$

The last statement of Theorem 3 follows. □

VIII. Variations of the Mutual Information

Notations 1: Let $\phi(t) = (1/\sqrt{2\pi})e^{-t^2/2}$ be the Gaussian probability density function (pdf) of unit variance. The entropy of $Z \sim \mathcal{N}(\eta,\lambda)$ is $(\log 2\pi e\lambda)/2$ ([20], Example 9.1.2). So as to simplify future computations, let $\nu = \lambda^{-1/2}$. The pdf of the mixture is $h(y) = (1-p)\phi(y-\eta) + p\nu\phi(\nu y)$. Let us denote $\psi(x) = x\log x$.

The mutual information can be written

$$I(\eta,\nu^{-2},p) = -\int \psi[h(y)]\,dy - \frac{1}{2}\log 2\pi e + p\log\nu. \quad (29)$$

Proof of Lemma 1-(a): We compute $(\partial/\partial\eta)I(\eta,1,p)$

$$\begin{aligned}
&= -\int \frac{\partial}{\partial\eta}h(y)\psi'[h(y)]\,dy \\
&= +\int (1-p)\phi'(y-\eta)[1+\log h(y)]\,dy \\
&\overset{(a)}{=} -(1-p)\int \phi(y-\eta)\frac{h'(y)}{h(y)}\,dy \\
&= +(1-p)\int \phi(y-\eta) \\
&\quad\times \frac{(1-p)\cdot(y-\eta)\phi(y-\eta) + py\phi(y)}{h(y)}\,dy \\
&\overset{(b)}{=} (1-p)\int \phi(y-\eta)\frac{(y-\eta)h(y) + p\eta\phi(y)}{h(y)}\,dy \\
&\overset{(c)}{=} \eta\underbrace{p(1-p)\int \frac{\phi(y)\phi(y-\eta)}{h(y)}\,dy}_{>0}.
\end{aligned}$$

In (a) we integrated by parts, in (b) we introduced $h(y)$ at the numerator, and in (c) we used the cancellation of the integral of the odd function $y\phi(y)$. Hence, the result. □

Proof of Lemma 1-(b): We compute $(\partial/\partial\nu)I(0,\nu^{-2},p)$

$$\begin{aligned}
&= -\int \frac{\partial}{\partial\nu}h(y)\psi'[h(y)]\,dy + \frac{p}{\nu} \\
&= \frac{p}{\nu} - \int \frac{p}{\nu}\nu\phi(\nu y)[1-(\nu y)^2][1+\log h(y)]\,dy \\
&\overset{(a)}{=} \frac{p}{\nu}\left\{1 - \int \overbrace{\phi(u)(1-u^2)}^{-\phi'(u)}\left(1+\log h\left(\frac{u}{\nu}\right)\right)\,du\right\} \\
&\overset{(b)}{=} \frac{p}{\nu}\left\{1 - \int \phi'(u)\frac{\frac{1}{\nu}h'\left(\frac{u}{\nu}\right)}{h\left(\frac{u}{\nu}\right)}\,du\right\} \\
&= \frac{p}{\nu}\left\{1 - \int u\phi(u)\frac{1}{\nu}\frac{(1-p)\frac{u}{\nu}\phi\left(\frac{u}{\nu}\right) + p\nu^3\frac{u}{\nu}\phi(u)}{h\left(\frac{u}{\nu}\right)}\,du\right\} \\
&= \frac{p}{\nu}\left\{1 - \int u^2\phi(u)\frac{(1-p)\frac{1}{\nu^2}\phi\left(\frac{u}{\nu}\right) + p\nu\phi(u)}{h\left(\frac{u}{\nu}\right)}\,du\right\}.
\end{aligned}$$

In (a) we used the change of variable $u = \nu y$ and in (b) we integrated by parts. As $\phi(u)$ is a pdf of unit variance $\int u^2\phi(u)du = 1$, the computation of $(\partial/\partial\nu)I(0,\nu^{-2},p)$ goes on

$$\begin{aligned}
&= \frac{p}{\nu}\int u^2\phi(u)\frac{h\left(\frac{u}{\nu}\right) - (1-p)\frac{1}{\nu^2}\phi\left(\frac{u}{\nu}\right) - p\nu\phi(u)}{h\left(\frac{u}{\nu}\right)}\,du \\
&= \frac{p}{\nu}\int u^2\phi(u)\frac{(1-p)\left(1-\frac{1}{\nu^2}\right)\phi\left(\frac{u}{\nu}\right)}{h\left(\frac{u}{\nu}\right)}\,du \\
&= \left(\nu - \frac{1}{\nu}\right)\underbrace{p(1-p)\int \frac{y^2\phi(y)\phi(\nu y)}{h(y)}\nu\,dy}_{>0}.
\end{aligned}$$

This shows that the sign of $(\partial/\partial\nu)I(0,\nu^{-2},p)$ is that of $\nu - 1/\nu$, hence, the result. □

IX. Proof of Lemma 2

In order to prove Lemma 2, we need a technical lemma first.

Lemme 3: The mutual information, in nats (i.e., defined using the natural logarithm 'log' rather than the logarithm in base 2 "$\log_2$" [20]) can be developed as

$$I(0,\nu^{-2},p) = \frac{p}{2}\left\{-1 + \frac{1}{\nu^2} + \log\nu^2\right\} + o(p). \quad (30)$$

Proof: We use the notations of Section I-B, and 'log' is the natural logarithm. Let us denote $r(y) = (p)/((1-p)(\nu\phi(\nu y)/\phi(y)))$, which enables us to write $\int\psi[h] = \int h\log h$ as $\int h\log p\phi + \int h\log[1+r]$

$$\begin{aligned}
&= \int h(y)\left[\log(1-p) - \frac{\log 2\pi}{2} - \frac{y^2}{2}\right]dy \\
&\quad + \int h\log[1+r] \\
&= \log(1-p) - \frac{\log 2\pi}{2} - \int h(y)\frac{y^2}{2}\,dy \\
&\quad + \int h\log[1+r] \quad (31)
\end{aligned}$$

because $\int h = 1$. Using the variances (1 and $1/\nu^2$) of the pdf $\phi(y)$ and $\nu\phi(\nu y)$, we can compute

$$\begin{aligned}
\int h(y)y^2 &= (1-p)\int \phi(y)y^2 \\
&\quad + p\int \nu\phi(\nu y)y^2 = 1 - p + \frac{p}{\nu^2}. \quad (32)
\end{aligned}$$

Collecting (29), (31), and (32), we get the estimate

$$\begin{aligned}
I(0,\nu^{-2},p) &= -\log(1-p) + \frac{\log 2\pi}{2} + \frac{1-p}{2} + \frac{p}{2\nu^2} \\
&\quad - \frac{\log 2\pi e}{2} + p\log\nu - \int h\log[1+r] \\
&= \frac{p}{2}\left\{-1 + \frac{1}{\nu^2} + \log\nu^2\right\} \\
&\quad - \log(1-p) - \int h\log[1+r]. \quad (33)
\end{aligned}$$

Let us now estimate the remaining integral term. The dominated convergence theorem shows that

$$\lim_{p\to 0}\int \nu\phi(\nu y)\log[1+r(y)]\,dy = 0$$

hence

$$\int h\log[1+r] = (1-p)\int \phi\log[1+r] + o(p).$$

As $\forall r \geq 0, 0 \leq \log(1+r) \leq r$, we get

$$0 \leq \int \phi\log[1+r] \leq \int \phi r = \frac{p}{1-p}$$

which leads to

$$\int h\log[1+r] = p + o(p). \tag{34}$$

Combined with the development $\log(1-p) = -p + o(p)$, (33) and (34) finally lead to (30). □

Lemma 2 is actually a corollary of Lemma 3. It is easy to show by a change of variables that $I(0,\lambda,1-p) = I(0,1/\lambda,p)$, hence, using Lemma 3 we get for $\lambda_A < 1 < \lambda_B$

$$\begin{aligned}\lim_{p\to 0}\frac{2}{p}\left(I(0,\lambda_B,p) - I(0,\lambda_A,p)\right) &= a(\lambda_B) - a(\lambda_A)\\ &= C_0(\lambda_A,\lambda_B);\\ \lim_{p\to 1}\frac{2}{1-p}\left(I(0,\lambda_B,p) - I(0,\lambda_A,p)\right) &= a\left(\lambda_B^{-1}\right) - a\left(\lambda_A^{-1}\right)\\ &= C_1\left(\lambda_A,\lambda_B\right).\end{aligned}$$

which gives (27) and (28). □

X. Update of the Log-Likelihood Ratio

In this section, we prove that in the simplified model corresponding to our numerical experiments, the log-likelihood ratio can be updated at each step as expressed in (12). To simplify the notations, we do not write the dependence of $\{k_m(x)\}_{m=1}^N$ on x. Using Bayes rule and the conditional independence of the coordinates conditionally to the class standard computations we get

$$\begin{aligned}S_{m+1}(x) &:= \log\frac{\mathcal{P}\left(Y=1 \,|\, \{\langle X,g_{k_l}\rangle = \langle x,g_{k_l}\rangle\}_{l=1}^{m+1}\right)}{\mathcal{P}\left(Y=0 \,|\, \{\langle X,g_{k_l}\rangle = \langle x,g_{k_l}\rangle\}_{l=1}^{m+1}\right)}\\ &= \log\frac{\mathcal{P}\left(Y=1, \{\langle X,g_{k_l}\rangle = \langle x,g_{k_l}\rangle\}_{l=1}^{m+1}\right)}{\mathcal{P}\left(Y=0, \{\langle X,g_{k_l}\rangle = \langle x,g_{k_l}\rangle\}_{l=1}^{m+1}\right)}\\ &= \log\frac{\mathcal{P}\left(Y=1, \{\langle X,g_{k_l}\rangle = \langle x,g_{k_l}\rangle\}_{l=1}^{m}\right)}{\mathcal{P}\left(Y=0, \{\langle X,g_{k_l}\rangle = \langle x,g_{k_l}\rangle\}_{l=1}^{m}\right)}\\ &\quad + \log\frac{\mathcal{P}\left(\langle X,g_{k_{m+1}}\rangle = \langle x,g_{k_{m+1}}\rangle \,|\, Y=1\right)}{\mathcal{P}\left(\langle X,g_{k_{m+1}}\rangle = \langle x,g_{k_{m+1}}\rangle \,|\, Y=0\right)}\\ &= S_m(x) + \log\frac{\mathcal{P}\left(\langle X,g_{k_{m+1}}\rangle = \langle x,g_{k_{m+1}}\rangle \,|\, Y=1\right)}{\mathcal{P}\left(\langle X,g_{k_{m+1}}\rangle = \langle x,g_{k_{m+1}}\rangle \,|\, Y=0\right)}.\end{aligned}$$

Since

$$\mathcal{P}(\langle X,g_{k_{m+1}}\rangle = u \,|\, Y=i) = \frac{1}{\sqrt{2\pi\sigma_{i,k_{m+1}}^2}}\exp\left(-\frac{u^2}{2\sigma_{i,k_{m+1}}^2}\right)$$

we obtain

$$\begin{aligned}S_{m+1}(x) &= S_m(x) + \log\frac{\sigma_{0,k_{m+1}}}{\sigma_{1,k_{m+1}}}\\ &\quad + \frac{1}{2}\left(\frac{1}{\sigma_{0,k_{m+1}}^2} - \frac{1}{\sigma_{1,k_{m+1}}^2}\right)\left\langle x,g_{k_{m+1}}\right\rangle^2\\ &= S_m(x) - \frac{1}{2}\log\lambda_{k_{m+1}}\\ &\quad + \frac{1}{2\,\sigma_{1,k_{m+1}}^2}\left(\lambda_{k_{m+1}}-1\right)\left\langle x,g_{k_{m+1}}\right\rangle^2.\end{aligned}$$

Acknowledgment

The author would like to thank S. Mallat and E. Bacry for their motivating interest in this research, as well as D. Geman, G. Gravier, and F. Bimbot for their valuable support and remarks.

References

[1] S. Mallat, *A Wavelet Tour of Signal Processing*, CA: Academic, 1998.
[2] R. A. DeVore, "Nonlinear approximation," in *Acta Numerica*. Cambridge, U.K.: Cambridge Univ. Press, 1998, vol. 7, pp. 51–150.
[3] R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Trans. Inf. Theory*, vol. 38, no. 2, pp. 713–718, Mar. 1992.
[4] J. M. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 41, no. 12, pp. 3445–3462, Dec. 1993.
[5] R. A. DeVore, B. Jawerth, and V. Popov, "Compression of wavelet decompositions," *Amer. J. Math.*, vol. 114, no. 4, pp. 737–785, 1992.
[6] D. L. Donoho and I. M. Johnstone, "Ideal denoising in an orthonormal basis chosen from a library of bases," *C. R. Acad. Sci. Paris Ser. I*, vol. 319, pp. 1317–1322, 1994.
[7] J. H. Friedman and J. W. Tukey, "A projection pursuit algorithm for exploratory data analysis," *IEEE Trans. Comput.*, vol. C-23, no. 9, pp. 881–889, Sep. 1974.
[8] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. London, U.K.: Chapman & Hall, 1984.
[9] R. A. Fisher, *The Design of Experiments*, 2nd ed. London, U.K.: Oliver & Boyd, 1937.
[10] K. S. Fu, *Sequential Methods in Pattern Recognition and Machine Learning, Vol. 52 of Mathematics in Science and Engineering*. New York: Academic, 1968.
[11] K. Fukunaga, *Introduction to Statistical Pattern Recognition, Electrical Science*. New York: Academic, 1972.
[12] G. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*. New York: Wiley, 1992.
[13] R. Duda and P. Hart, *Pattern Recognition and Scene Analysis*. New York: Wiley, 1973.
[14] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Netw.*, vol. 13, no. 4–5, pp. 411–430, 2000.
[15] D. J. Field and B. A. Olshausen, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607–609, 1996.
[16] N. Saito and R. R. Coifman, "Local discriminant bases," *J. Math. ImagVis.*, vol. 5, no. 4, pp. 337–358, Dec. 1995.
[17] N. Saito, "Least statistically-dependent basis and its application to image modeling," in *Wavelet Applications in Signal and Image Processing VI*, A. F. Laine, M. A. Unser, and A. Aldroubi, Eds. San Diego, CA, Jul. 1998, pp. 24–37. vol. 3458 of *Proc. SPIE*.

[18] B. Liu and S.-F. Ling, "On the selection of informative wavelets for machine diagnosis," *J. Mech. Syst. Signal Process.*, vol. 13, no. 1, pp. 145–162, 1999.
[19] S. Mallat and Z. Zhang, "Matching pursuit with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.
[20] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[21] P. J. Phillips, "Matching pursuit filters applied to face identification," *IEEE Trans. Image Process.*, vol. 7, no. 8, pp. 1150–1164, Aug. 1998.
[22] Q. Jiang, S. S. Goh, and Z. Lin, "Local discriminant time-frequency atoms for signal classification," *Signal Process.*, vol. 72, pp. 47–52, 1999.
[23] A. Wald, "Sequential tests of statistical hypothesis," *Ann. Math. Statist.*, vol. 16, no. 2, pp. 117–186, Jun. 1945.
[24] F. Esposito, D. Malerba, and G. Semeraro, "A comparative analysis of methods for pruning decision trees," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 5, pp. 476–491, May 1997.
[25] D. Geman and B. Jedynak, "An active testing model for tracking roads in satellite images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 1, pp. 1–14, Jan. 1996.
[26] C. Li, "Classification by active testing with applications to imaging and change detection," Ph.D. dissertation, Univ. Massachussets, Amherst, Feb. 1999.
[27] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory, Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[28] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthonormal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. 27th Annu. Asilomar Conf. Signals, Systems Computers*, Nov. 1993.
[29] L. Benaroya and F. Bimbot, "Wiener-based source separation with HMM/GMM using a single sensor," in *Proc. 4th Int. Symp. Independent Component Analysis Blind Signal Separation (ICA'03)*, Nara, Japan, Apr. 2003.

Remi Gribonval graduated from École Normale Supérieure, Paris, France in 1997 and received the Ph.D. degree in applied mathematics from the University of Paris-IX Dauphine, Paris, France, in 1999.

From 1999 to 2001, he was a Visiting Scholar at the Industrial Mathematics Institute (IMI), Department of Mathematics, University of South Carolina, Columbia. He is currently a Research Associate with the French National Center for Computer Science and Control (INRIA) at IRISA, Rennes, France. His research interests are in adaptive techniques for the representation and classification of audio signals with redundant systems, with a particular emphasis in blind audio source separation.