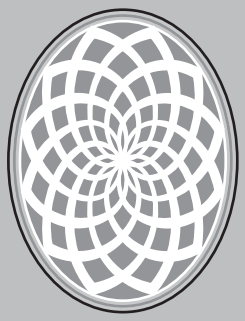

Industrial Mathematics Institute

2000:20

Active testing in a mixture of two Gaussian classes

R. Gribonval

IMI

Preprint Series

Department of Mathematics
University of South Carolina

Active Testing in a Mixture of two Gaussian Classes

R. Gribonval, Centre de Mathématiques Appliquées,
École Polytechnique, France

Abstract— **In a mixture of two high-dimensional Gaussian classes, the class can be identified with *active testing*, *i.e.* by a (sequential) adaptive selection of features from a redundant dictionary. Using the mutual entropy criterion, we provide an analytic characterization of this procedure in two situations. When the classes have the same covariance but different centers, the optimal features can be computed offline with an algorithm similar to the Orthonormal Matching Pursuit. When the classes are centered, the adaptive features are eigen-vectors of an operator determined by the covariances of the two classes, but we prove that the optimal schedule of observations of these eigen-vectors actually depends on the observed signal. The active testing process then takes the form of an (incomplete) binary decision tree. In the case of two stationary Gaussian noises, a low-frequency one and a high-frequency one, we provide numerical experiments showing that adaptivity does improve the accuracy of the estimation of the class with few tests. We also get a somewhat surprinsing result : the frequency components that are likely to be the least energetic bring the most information.**

Keywords—**sequential testing, adaptive feature selection, greedy algorithm, decision tree, mutual entropy, mutual information, infomax, Gaussian identification, classification, discriminant analysis.**

I. Introduction

The continuous increase in resolution and popularity of digital data acquisition devices (digital cameras, scanners, digital sound recorders ...) is giving birth to large databases of high-dimensional signals and images. Meanwhile, fast pattern recognition techniques are becoming crucial in critical systems monitoring (engines, planes, nuclear powerplants), biometrical identification (speaker, fingerprint or eye identification), office applications (optical character recognition, voice recognition), and many other fields including military applications, astronomy, satellite imaging, medical imaging, handwriting recognition, musical score recognition, *etc.* While higher-resolution data potentially means better rates of recognition, it also implies a larger computational cost which does not suit the increasing need for real-time processing.

Feature selection aims at reducing the computational burden of the identification of high-dimensional signals $x \in \mathbb{R}^N$ (typically, a speech sample or an image), by extracting a small number M of informative *features* $\{f_m(x)\}_{m=1}^M$. In a Bayesian setting, the observed data x is supposed to be a realization of a multi-dimensional random variable X, and the unknown class y (*e.g.* the name of the speaker in a speaker identification problem) is a realization of a joint random variable Y.

Linear Discriminant Analysis (LDA) [McL92] selects M linear features by maximizing some measure of the "information" that the M random variables $\{f_m(X)\}_{m=1}^M$ give about Y in the mean. In such a *passive* scheme, the features are selected once for all, during a learning procedure, *before* observing the data x that is to be classified. Several variants of LDA have been defined, using various information theoretic measures of information [CT91] (mutual entropy, Hellinger divergence, Kullback-Leibler divergence, ...).

Recently, the interaction of Feature Selection with Computational Harmonic Analysis proved quite succesfull, as computationally efficient techniques for passive feature selection [Wic91], [SC94], [Sai98], [LL99] were made possible by the restriction to linear features such as Fourier coefficients, wavelet coefficients [Dau88], [Mal89], wavepackets coefficients [CW92]. In the very last years, much effort was put in the development of feature extraction using "nonlinear" methods such as adaptive decompositions in redundant dictionaries that can be much larger than a basis [MZ93], [CD99] [JCMW98], [GDR+96], [MC97]. Among the reasons for this efforts was the clear superiority of adaptive methods over linear ones [DeV98] for approximation/regression. Moreover the push towards the use of redundant dictionaries was motivated by the belief that it would enable an even better fit to the signals of interest. To put it shortly, the combination of redundancy and adaptivity allows for sparser representations of the signals than linear techniques.

However, identification requires the selection of *informative* features, which is quite different from the good *approximation power* of the features selected by adaptive decompositions. The statistical framework of Sequential Decision Theory [Wal49] is more suited to the design of adaptive feature selection than the theory of nonlinear approximation, with which it shares some principles.

In sequential tests [Wal45], [Fu68], one observes features $f_m(x)$ sequentially and takes a decision on the class y of x as soon as a stopping criterion is reached. As a result, the number $M(x)$ of observations $\{f_m(x)\}_{m=1}^{M(x)}$ *depends* on the observed data x. Similarly, *active testing* [BFOS84] can reduce $M(x)$ by selecting the $m+1$-th feature f_{m+1} once $f_m(x)$ has been observed, thus taking into account some specific characteristics of x. A good example of such an active identification strategy is the *Game of Twenty Questions*. In general, no closed form expression is available for the adaptive sequence of features, which is structured as a decision tree which may have infinitely many branches starting from each node.

Because of the lack of stochastic model for signals or images, the construction of the tree and the selection of features is usually done with data-driven techniques such as

The author is currently in the Department of Mathematics, University of South Carolina, Columbia, SC 29208, USA. e-mail : remi@math.sc.edu, phone : (803) 777-5299, fax : (803) 777-3783

CART [BFOS84]. Despite their success in character recognition [AG97], [AGW97] and speech recognition [AM99], these inductive techniques cannot build deep trees or allow many branches per node without losing their statistical significance. However, for some parametric models $\mathcal{P}_\Theta(x,y), \Theta \in \mathbb{R}^d$, an analytic characterization of the adaptive feature selection can be derived [GJ96], [Li99], making it possible to build on-line a deeper and/or more branched out tree.

In this paper we provide such an analytic characterization in the case of a mixture of two multi-dimensional Gaussian classes. In the case of two stationary Gaussian noises, a low-frequency one (low-pass filtered Gaussian white noise) and a high-frequency one (high-pass filtered), we get the somewhat surprinsing result (compared to what we would expect if we were performing *approximation* of the signal) that one should observe the frequencies where *the least amount of energy* is likely to be present in the signal.

The paper is organized as follows. In section II we define passive and adaptive (sequential) feature selection with the mutual entropy criterion, using extensively the analogy with linear and nonlinear approximation. We state in section III our main theorems, giving an analytic characterization of active testing for a mixture of two multi-dimensional Gaussian random variables. We discuss in section IV numerical experiments with a low-frequency and a high-frequency stationary Gaussian noises : we compare the passive and the active testing algorithms, and we point out the somewhat surprising fact that the optimal features for the identification grab "as little" energy of the analyzed signal as they can at each step. The proofs are collected in the appendix.

II. Sequential Feature Selection

Feature Selection is designed to extract "significant features" from a high-dimensional signal x, so as to identify its class y (a symbol in a finite alphabet). In an ideal world, each signal can be assigned a class without any uncertainty : the space $\mathbb{R}^N$ can be split into disjoint sets, each one exactly matched to one class. But in the real world (*e.g.* in a speech recognition process) there is some uncertainty on the class of a signal, corresponding to some overlap in the "partition". We model this uncertainty using a Bayesian setting : before observing any feature of x, the class has some prior distribution; the desired effect of the observation of features is a more "peaky" posterior distribution.

In this setting, the best estimator $\widehat{Y}(X)$ of the unknown class is the Bayesian estimator [Kay93], which minimizes the probability of misclassification $\mathcal{P}(\widehat{Y} \neq Y)$. In the case of uniform prior, the Bayesian estimator coincides with the Maximum Likelihood one. However, the high dimension of the signal x generally makes it difficult to actually build this classifier, and one uses a smaller set of M features $\{f_m(X)\}_{m=1}^M$ to build an estimator $\widehat{Y}(\{f_m(X)\}_{m=1}^M)$. Good features should lead to a "peaky" posterior distribution of the class. Information theory provides a nice framework to select such features.

The mutual entropy [CT91] between two (possibly multi-dimensional) random variables Z and Y is defined as the symmetric functional $\mathcal{I}(Z;Y) = \mathcal{H}(Y) - \mathcal{H}(Y|Z) = \mathcal{H}(Z) - \mathcal{H}(Z|Y)$ where $\mathcal{H}(.)$ is the entropy and $\mathcal{H}(.|.)$ the conditional entropy. Fano's inequality shows that if $\mathcal{P}(\widehat{Y}(Z) \neq Y)$ is small for some estimator $\widehat{Y}(Z)$, then $\mathcal{H}(Y|Z)$ must be small, hence $\mathcal{I}(Z;Y)$ must be large, *i.e.* close to $\mathcal{H}(Y)$. Hence M "good" features should lead to a large mutual entropy $\mathcal{I}\left(\{f_m(X)\}_{m=1}^M; Y\right)$. In practice one has to choose them from a limited set of easily computable features.

In signal and image processing [Mal98], it is common to consider linear features, that is to say linear functionals $f_g : x \mapsto f_g(x) = \langle x, g\rangle = \sum_{n=0}^{N-1} x[n]g[n]$ where g is a vector in $\mathbb{R}^N$. Some examples of widely used linear features are

Example 1: coordinates in the Dirac basis : $x[n]$.

$$g[n] = \delta[n-n_0] = \begin{cases} 1, n = n_0 \\ 0, n \neq n_0 \end{cases} . \tag{1}$$

Example 2: Fourier coefficients : $\widehat{x}[k]$,

$$g[n] = \frac{1}{\sqrt{N}} \exp\left(\frac{2i\pi kn}{N}\right). \tag{2}$$

Example 3: wavelet coefficients : $\langle x, \psi_{j,k}\rangle$.

$$g[n] = \psi_{j,k}[n]. \tag{3}$$

See [Dau88], [Dau92], [Mal89], [Mal98] for references on wavelets.

A *dictionary* $\mathcal{D}$ of linear features is any subset $\mathcal{D}$ of the unit sphere in $\mathbb{R}^N$ that spans $\mathbb{R}^N$. It can be as small as an orthonormal basis and as redundant as the whole unit sphere. Examples include the Dirac-Fourier dictionary [MZ93], the wavepackets dictionary [CW92], the Gabor dictionary [Tor91], [QC94], [MZ93], as well as data-driven dictionaries [MC97].

Maximizing the mutual entropy

$$\{g_1^\star, \ldots, g_M^\star\} = \arg \max_{g_1,\ldots,g_M \in \mathcal{D}} \mathcal{I}\Big(\{f_{g_m}(X)\}_{m=1}^M; Y\Big)$$

in order to select M "good" features is a difficult optimization problem. It is easier to proceed in a sequential, greedy way, in the spirit of the Projection Pursuit/Matching Pursuit algorithm [FT74], [FS81], [Hub85], [MZ93]. A first feature $g_1 = \arg\max_{g\in\mathcal{D}} \mathcal{I}(f_g(X);Y)$ is selected. Then, using the chain rules for mutual entropy [CT91], the following ones are iteratively defined as

$$\begin{aligned} g_m &= \arg\max_{g\in\mathcal{D}} \quad \mathcal{I}\Big(f_{g_1}(X), \ldots, f_{g_{m-1}}(X), f_g(X); Y\Big) \\ &= \arg\max_{g\in\mathcal{D}} \quad \Big[\mathcal{I}\Big(f_g(X); Y \Big| \{f_{g_l}(X)\}_{l=1}^{m-1}\Big) \\ &\qquad\qquad +\mathcal{I}\Big(\{f_{g_l}(X)\}_{l=1}^{m-1}; Y\Big)\Big] \\ &= \arg\max_{g\in\mathcal{D}} \quad \mathcal{I}\Big(f_g(X); Y \Big| \{f_l(X)\}_{l=1}^{m-1}\Big). \end{aligned} \tag{4}$$

An analogue of this sequential optimization problem is the sequential minimization of the expected error of ap-

proximation $\mathbb{E}\left\{ \|X - P_{\mathcal{V}_m} X\|^2 \right\}$ of the signal x by its orthonormal projection onto the subspace $\mathcal{V}_m = \text{span}\{g_l, 1 \leq l \leq m\}$. The "optimal features" for approximation are the elements of the Karhunen-Loeve basis. However the theory of *nonlinear approximation* [DeV98] shows that the sequential approximation scheme associated with the Karhunen-Loeve basis is actually far from being "optimal", especially when the distribution is not Gaussian, because it does not adapt to the vector x which is being approximated. On the contrary, in an orthonormal basis $\mathcal{B} = \{g_n\}_{n=1}^N$, nonlinear approximation can do better by approximating x with the m vectors that correspond to its m largest coefficients $|\langle x, g_{n_1}\rangle| \geq \ldots \geq |\langle x, g_{n_m}\rangle|$.

Similarly, the sequential choice (4) of features is passive : it does not depend on the observations $\{f_{g_l}(x)\}_{l=1}^{m-1}$ that were already collected about x. Another possibility is to choose $g_m(x)$ adaptively, using what is known so far about x :

$$g_m(x) := \arg\max_{g \in \mathcal{D}} \mathcal{I}\Big(f_g(X); Y \Big| \{f_{g_l(x)}(X) = f_{g_l(x)}(x)\}_{l=1}^{m-1}\Big). \quad (5)$$

Such an adaptive choice of g_m brings at least as much information at each step as the passive one does, and leads to a natural tree structure [BFOS84], where there may be infinitely many branches starting from each node. Combining it with sequential testing techniques [Wal49], [Fu68], one can expect to reach a reliable decision on the class with a smaller number of observations $M(x)$.

Unfortunately, the conditional mutual entropy in (4) or (5) cannot generally be estimated with statistical significance from a reasonable amount of learning data. As a heuristics, some passive sequential strategies [Phi98], [JGL99], [LL99] use linear projections instead of conditioning, *i.e.* $\mathcal{I}\big(\langle x - P_{\mathcal{V}_{m-1}} x, g\rangle ; Y\big)$ becomes the criterion to be maximized. However, there are cases where active testing can actually be implemented : with a parametric model $\mathcal{P}_\theta(x, y)$ for which an analytic computation of the conditional entropy is possible. Using such a model with a dictionary of nonlinear features, Geman and Jedynak [GJ96] were able to actively track roads in satellite images by building on-line, during the recognition process, the only branch of interest of a deep, heavily branched out tree. The structure of their model was such that they did not even need to store the tree.

In [Li99], with a model of mixture of Gaussian random variables and linear features consisting in the coordinates of x in a given basis, an analytic computation of the *Hellinger divergence* [GL00] was given. It resulted in an explicit algorithm for the adaptive selection of coordinates with this criterion. In the next section we state similar results using the *mutual entropy* —which is more complex to manipulate than the Hellinger divergence— and choosing linear features in a redundant dictionary.

III. Adaptive Features for Gaussian classes

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 the signal 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 *a priori* distribution of the two classes is given by $p_0 = \mathcal{P}(Y = 0)$. We assume that Σ_y has full rank.

Let $\mathcal{D}$ be some dictionary of vectors in $\mathbb{R}^N$, and $\{x \mapsto \langle x, g\rangle, g \in \mathcal{D}\}$ the associated dictionary of linear features. Our purpose in this section is to characterize the sequences of functions $\{x \mapsto g_m(x)\}_{m=1}^N$ satisfying (5) for all x and $1 \leq m \leq N$. Such sequences are called *adaptive feature sequences* (AFS).

We first get the following characterization.

Theorem 1: Assume $\Sigma_1 = \Sigma_0$ 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 $1 \leq m \leq N$

$$g_m := \arg\max_{g \in \mathcal{D}} \left| \left\langle \Sigma^{-1}(\vec{\mu}_1 - \vec{\mu}_0), R_{m-1} g / \|R_{m-1} g\|_\Sigma \right\rangle_\Sigma \right|, \quad (6)$$

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

Remark 1: When some vector $g_1 \in \mathcal{D}$ is colinear to the matched filter $\Sigma_0^{-1}(\vec{\mu}_1 - \vec{\mu}_0)$, g_1 is the first and only feature to be used : after observing $\langle x, g_1\rangle$, no other feature will give any information on the class.

Remark 2: When $\mathcal{D}$ is an orthonormal basis with respect to $\langle ., . \rangle_\Sigma$, this AFS with the mutual entropy criterion could equivalently be obtained with the Hellinger divergence criterion [Li99] (page 44-45).
A consequence of Theorem 1 is that, when $\Sigma_0 = \Sigma_1$, AFS is essentially *unique* and actually *independent* of the observed data x. It can indeed be computed off-line. In such a case adaptivity is simply useless. The similarity of the computation of the AFS with an Orthonormal Matching Pursuit [PRK93] (see the appendix) may explain the good behavior of passive feature selection strategies [Phi98], [JGL99], [LL99] that use linear projections instead of conditional mutual entropy estimation. This identification problem is indeed very close to an approximation problem.

The following theorem shows that AFS have a totally different structure when $\Sigma_1 \neq \Sigma_0$ and $\vec{\mu}_1 = \vec{\mu}_0$.

Theorem 2: Assume that $\Sigma_1 \neq \Sigma_0$ and $\vec{\mu}_1 = \vec{\mu}_0$. Let $\mathcal{D}$ be the whole unit sphere of $\mathbb{R}^N$ and $\{u_k\}_{k=1}^N$ a basis of unit eigen-vectors 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 $1 \leq m \leq N$

$$\text{span}\{g_l(x), 1 \leq l \leq m\} = \text{span}\{u_{k_l(x)}, 1 \leq l \leq m\}. \quad (7)$$

The proof is in the appendix.

Remark 3: Notice that $\Sigma_0^{-1}\Sigma_1$ is not necessarily a symmetric matrix. However it can be diagonalized, because it

is similar to the symmetric matrix $\Sigma_0^{-1/2}\Sigma_1\Sigma_0^{-1/2}$. A basis of unit eigen-vectors $\{u_k\}_{k=1}^N$ can thus be computed off-line (without observing the signal x which has to be classified) but is generally not orthonormal. Hence such a basis is generally distinct from the Least Statistically Dependant (Orthonormal) Basis of Saito [Sai98].

Remark 4: The components $\{\langle X, u_k\rangle, 1 \leq k \leq N\}$ are not necessarily independent random variables, whether under the mixture distribution or under one of the conditional (Gaussian) distribution. This shows the difference between the AFS approach and Independent Component Analysis [Com94].

The feature family is indeed the same as the one Linear Discriminant Analysis (LDA) produces for this problem [Fuk72]. However the potential difference with LDA is that the *schedule* $\{k_m(x)\}_{m=1}^N$ of the observations $\langle x, u_k\rangle$ in an AFS may actually depend on the observed data x, and may have to be decided on-line. We will need the following definition.

Definition 1: For any $p \in [0,1]$, η and $0 < \lambda < \infty$, let $\mathcal{I}(\eta, \lambda, p) = \mathcal{I}(Z;Y)$ where Z is a mixture of two one-dimensional Gaussian classes $\mathcal{N}(\eta, 1)$ and $\mathcal{N}(0, \lambda)$ with mixture parameter p and Y is the class variable.

Let us give some precision on the optimal schedule of the observations when $\Sigma_0 \neq \Sigma_1$ and $\vec{\mu}_0 = \vec{\mu}_1$.

Lemma 1: Let $\{\lambda_k\}_{k=1}^N$ be the eigen-values of $\Sigma_0^{-1}\Sigma_1$ associated to the unit eigen-vectors $\{u_k\}_{k=1}^N$, and let

$$\left(\underline{\lambda}_m(x), \overline{\lambda}_m(x)\right) = (\min, \max)\left\{\lambda_k, k \notin \{k_l(x)\}_{l=1}^m\right\} \quad (8)$$

be the extremal remaining eigen-values after m steps, $0 \leq m \leq N-1$. The optimal schedule $\{k_m(x)\}_{m=1}^N$ is characterized by

$$\lambda_{k_m(x)} = \arg \max_{\lambda \in \left\{\underline{\lambda}_{m-1}(x), \overline{\lambda}_{m-1}(x)\right\}} \mathcal{I}\Big(0, \lambda, p_{m-1}(x)\Big) \quad (9)$$

with

$$p_{m-1}(x) = \mathcal{P}\left(Y = 0 \middle| \left\{ \langle X, u_{k_l(x)}\rangle = \langle x, u_{k_l(x)}\rangle \right\}_{l=1}^{m-1}\right) \quad (10)$$

the *a posteriori* distribution of Y after $m-1$ observations. The proof is in the appendix.

This lemma shows that the first index k_1 is actually constant, because $\underline{\lambda}_0$ and $\overline{\lambda}_0$ are independent of x, as well as the *a priori* distribution p_0 of Y. Hence, after the first observation $\langle X, u_{k_1}\rangle = \langle x, u_{k_1}\rangle$, the values $\underline{\lambda}_1$ and $\overline{\lambda}_1$ are still independent of x. However the *a posteriori* distribution $p_1(x)$ now depends on x. The choice of the second index $k_2(x)$ may thus actually depend on x. Hence the AFS has "potentially" the structure of a binary decision tree $\mathcal{T}(\Sigma_0^{-1}\Sigma_1)$:

- Its root is labeled with the first feature u_{k_1}, which is independent of the sample x to be classified.
- A node of depth $m-1$ has at most two children nodes. However if, say, $\mathcal{I}(0, \underline{\lambda}_{m-1}, p) > \mathcal{I}(0, \overline{\lambda}_{m-1}, p)$ for every value of p, there is only one child.

This tree is of depth $N-1$, and we prove in the appendix that the number of its branches is at most $C_{N-1}^{\lfloor (N-1)/2\rfloor} \sim 2^N/\sqrt{2\pi N}$. Even if this is much smaller than a fully-expanded binary tree of depth $N-1$, the storage of such a tree could certainly be a practical burden when N is large. However Lemma 1 eliminates this problem by providing a natural on-line construction of the branch of interest for any given signal x.

To be sure that the algorithm is adaptive, there remains to show that $k_m(x)$ can actually depend on x. The following lemma shows that it is not always the case.

Lemma 2:
- if $\underline{\lambda}_{m-1}(x) \geq 1$, then $\lambda_{k_m(x)} = \overline{\lambda}_{m-1}(x)$.
- if $\overline{\lambda}_{m-1}(x) \leq 1$, then $\lambda_{k_m(x)} = \underline{\lambda}_{m-1}(x)$.

The proof is in the appendix.

As an example, if $\underline{\lambda}_1 \geq 1$, then $\underline{\lambda}_m(x) \geq 1$ for all m, and the AFS —except maybe its first term which may be $\lambda_{k_1} < 1$— is associated with the decreasing rearrangement of the eigen-values. Nevertheless, in the general case, $k_m(x)$ does depend on x. Let us prove it, using the following lemmas.

Lemma 3: For $t > 0$, let $a(t) = t - \log t$ and $b(t) = t^{-1} + \log t = a(t^{-1})$. Let $0 < \underline{\lambda} < 1 < \overline{\lambda}$.
- for p close to 1,

$$\mathcal{I}(0, \overline{\lambda}, p) - \mathcal{I}(0, \underline{\lambda}, p) \sim \frac{1-p}{2}\left(a(\overline{\lambda}) - a(\underline{\lambda})\right); \quad (11)$$

- for p close to 0,

$$\mathcal{I}(0, \overline{\lambda}, p) - \mathcal{I}(0, \underline{\lambda}, p) \sim \frac{p}{2}\left(b(\overline{\lambda}) - b(\underline{\lambda})\right). \quad (12)$$

Lemma 4: For any $\overline{\lambda} > 1$, there exists unique

$$0 < \alpha(\overline{\lambda}) < 1/\overline{\lambda} < \beta(\overline{\lambda}) < 1$$

such that $a(\overline{\lambda}) \geq a(\underline{\lambda}) \Leftrightarrow \underline{\lambda} \geq \alpha(\overline{\lambda})$ and $b(\overline{\lambda}) \geq b(\underline{\lambda}) \Leftrightarrow \underline{\lambda} \geq \beta(\overline{\lambda})$.

The proof of Lemma 3 is somewhat technical and can be found in the appendix. That of Lemma 4 is straightforward and is let to the reader.

Suppose that after $m-1$ steps ($m-1 \geq 1$), $\underline{\lambda}_{m-1}(x)$ is in the open neighborhood $\Big(\alpha\big(\overline{\lambda}_{m-1}(x)\big), \beta\big(\overline{\lambda}_{m-1}(x)\big)\Big)$ of $1/\overline{\lambda}_{m-1}(x)$. Lemma 3 and Lemma 4 combined show that :
- if $p_{m-1}(x)$ is close enough to 0,

$$\mathcal{I}\Big(0, \overline{\lambda}_{m-1}(x), p_{m-1}(x)\Big) < \mathcal{I}\Big(0, \underline{\lambda}_{m-1}(x), p_{m-1}(x)\Big);$$

- if $p_{m-1}(x)$ is close enough to 1,

$$\mathcal{I}\Big(0, \overline{\lambda}_{m-1}(x), p_{m-1}(x)\Big) > \mathcal{I}\Big(0, \underline{\lambda}_{m-1}(x), p_{m-1}(x)\Big).$$

Both of these two "extreme" situations are actually possible, because the *a posteriori* distribution $p_{m-1}(x)$ *does* depend on x. This proves the result we were claiming : the adaptive schedule of observation of the features does depend on x.

IV. Numerical experiments

Numerical results of active testing using the Hellinger divergence [Li99] suggest that the posterior distribution $p_m(x)$ converges faster towards $p_N(x)$ with active testing

than with a fixed schedule. In the following, we provide similar numerical evidence for our active testing algorithm with the mutual entropy criterion.

From now on, we make the simplifying assumption that Σ_0 and Σ_1 commute, hence the eigen-vectors $\{u_k\}_{k=1}^N$ of $\Sigma_0^{-1}\Sigma_1$ are the common eigen-vectors of Σ_0 and Σ_1.

A. Semi-empirical adaptive algorithm.

After centering the signal x by removing the mean $\vec{\mu}_0 = \vec{\mu}_1$, the active testing algorithm shall go as follows :
1- initialize $\underline{\lambda}_0$, $\overline{\lambda}_0$, and p_0;
2- set $m = 1$;
3- compute and compare $\mathcal{I}(0, \underline{\lambda}_{m-1}(x), p_{m-1}(x))$ and $\mathcal{I}(0, \overline{\lambda}_{m-1}(x), p_{m-1}(x))$ to select $k_m(x)$;
4- set $\underline{\lambda}_m(x)$, $\overline{\lambda}_m(x)$;
5- observe $\langle x, u_{k_m(x)} \rangle$ and compute $p_m(x)$ using Bayes rule (remember that we assume Σ_0 and Σ_1 commute, *i.e.* they are diagonal in the same basis)

$$\frac{1-p_m(x)}{p_m(x)} = \frac{1-p_{m-1}(x)}{p_{m-1}(x)} \frac{\sigma_{0,k_m(x)}}{\sigma_{1,k_m(x)}} \frac{e^{-\frac{|\langle x, u_{k_m(x)} \rangle|^2}{\sigma^2_{1,k_m(x)}}}}{e^{-\frac{|\langle x, u_{k_m(x)} \rangle|^2}{\sigma^2_{0,k_m(x)}}}}; \quad (13)$$

6- increment m and go back to step 2.

The decision rule (step 3) could be implemented by tabulating $\mathcal{I}(0, \lambda_k, s/S), 1 \leq k \leq N, s = 0, 1, \ldots S$ where $1/S$ is some small step. The tabulation would then be part of a learning step, which includes estimating Σ_0 and Σ_1 and diagonalizing $\Sigma_0^{-1}\Sigma_1$, once for all, prior to any identification. However, because of the results expressed in Lemma 3 and Lemma 4 we conjecture that the decision set

$$D(\underline{\lambda}, \overline{\lambda}) := \{p \in [0,1], \mathcal{I}(0, \overline{\lambda}, p) \geq \mathcal{I}(0, \underline{\lambda}, p)\} \quad (14)$$

actually has a simple characterization.

Conjecture 1: For any $0 < \underline{\lambda} < 1 < \overline{\lambda}$:

$$D(\underline{\lambda}, \overline{\lambda}) = [0,1] \cap [p(\underline{\lambda}, \overline{\lambda}), \infty), \quad (15)$$

where $p(\underline{\lambda}, \overline{\lambda})$ is close to the expression

$$\widetilde{p}(\underline{\lambda}, \overline{\lambda}) := \frac{1}{1 - \frac{a(\overline{\lambda}) - a(\underline{\lambda})}{b(\overline{\lambda}) - b(\underline{\lambda})}}. \quad (16)$$

Remark 5: Thanks to the continuity of $p \mapsto \mathcal{I}(0, \lambda, p)$ and to Lemma 3 and Lemma 4, (15) is equivalent to the existence of at most one value $p \in (0,1)$ such that $\mathcal{I}(0, \overline{\lambda}, p) = \mathcal{I}(0, \underline{\lambda}, p)$. The expression (16) gives $D(\underline{\lambda}, \overline{\lambda}) = \emptyset$ for $a(\underline{\lambda}) \geq a(\overline{\lambda})$ and $D(\underline{\lambda}, \overline{\lambda}) = [0,1]$ for $b(\underline{\lambda}) \leq b(\overline{\lambda})$, which is compatible with the results of Lemma 3 and Lemma 4.

In the numerical experiments, the exact active testing algorithm is replaced with a "semi-empirical" one using a decision rule which comes directly from our conjecture. The rule is summarized in table I.

B. Passive algorithm.

By definition, a passive testing algorithm does not take into account any information obtained on x after a given number of steps. We use the following passive testing algorithm : use the features $\{u_k\}_{k=1}^N$ with a fixed schedule $\{k_m^p\}_{m=1}^N$ such that $\{\max(\lambda_{k_m^p}, 1/\lambda_{k_m^p})\}_{m=1}^N$ is non-increasing. Note that $\{k_m^p\}_{m=1}^N$ is, of course, independent of x. It is quite easy to see that this schedule of tests corresponds to following a special branch of the binary decision tree $\mathcal{T}(\Sigma_0^{-1}\Sigma_1)$: the one which is followed by all signals x for which $p_m(x) \approx 1/2$ for every m. Hence the passive schedule $\{k_m^p\}_{m=1}^N$ is good when not much information is available on the class y, and it is reasonable to think that this passive testing algorithm has fair efficiency for getting information about the unknown class y.

$p_{m-1}(x) - \widetilde{p}(\underline{\lambda}_{m-1}(x), \overline{\lambda}_{m-1}(x))$	< 0	≥ 0
$f(\underline{\lambda}_{m-1}(x)) \geq f(\overline{\lambda}_{m-1}(x))$	$\underline{\lambda}_{m-1}(x)$	
$f(\underline{\lambda}_{m-1}(x)) < f(\overline{\lambda}_{m-1}(x))$ and $g(\underline{\lambda}_{m-1}(x)) > g(\overline{\lambda}_{m-1}(x)))$	$\underline{\lambda}_{m-1}(x)$	$\overline{\lambda}_{m-1}(x)$
$g(\underline{\lambda}_{m-1}(x)) \leq g(\overline{\lambda}_{m-1}(x))$	$\overline{\lambda}_{m-1}(x)$	

TABLE I
SEMI-EMPIRICAL DECISION RULE FOR ACTIVE TESTING.

C. Numerical results.

In signal processing, many finite-length signals (such as the voiced parts of speech signals, *e.g. vowels*) are modeled as a realization of a cyclo-stationary Gaussian noise. The covariance operator of such a random process is diagonalized in the discrete Fourier basis (2), and an estimate of the eigen-values $\{\lambda_k\}_{k=1}^N$ is obtained through the estimation of the power spectrum of each class. After a Fast Fourier Transform [BP85], [FJ98] of the signal x, one can proceed to the on-line active identification by adaptively picking frequencies k where to measure the power spectrum $|\widehat{x}[k]|^2$.

Let us consider two noises : a high-frequency one ($Y = 0$)

$$\sigma^2_{0,k} = \begin{cases} c^{-1}, & 1 \leq k \leq N/2 \\ c, & N/2+1 \leq k \leq N \end{cases} \quad (17)$$

and a low-frequency one ($Y = 1$)

$$\sigma^2_{1,k} = \begin{cases} c, & 1 \leq k \leq N/2 \\ c^{-1}, & N/2+1 \leq k \leq N \end{cases} \quad (18)$$

where $c > 1$. For low frequencies $1 \leq k \leq N/2$, $\lambda_k = c^2 > 1$, while for high frequencies $N/2+1 \leq k \leq N$, $\lambda_k = c^{-2} < 1$. Lemma 3 and Lemma 4 show that when the high-frequency noise is more likely ($p \approx 1$), the most informative measures are on the low-frequency components $|\widehat{x}[k]|^2, 1 \leq k \leq N/2$ while, when the low-frequency hypothesis is more likely ($p \approx 0$), the high frequency components give more information. Hence the active testing

strategy is very different from an approximation strategy : on the one hand, for approximation (either linear or non-linear), one would try to select frequencies where the components are *as large as possible*; on the other hand, active testing selects the frequencies in the opposite order, looking for the *least energetic* components first.

There is an intuitive explanation for this behavior : when the low-frequency noise is likely, if we make a measure at a high frequency, the variance of the measure is likely to be small. Hence, if ever we do observe a large high-frequency component $|\hat{x}[k]|^2 \gg 1$, $k \geq N/2+1$, it will give us quite a strong indication that our assumption (low-frequency, $Y=1$) is false. On the contrary, a measurement at a low frequency is not likely to make us revise our estimation of the class.

With the two Gaussian classes (17) and (18), in dimension $N = 16$, using $c = 2.5$, $L = 10000$ samples (x_j, y_j) of the mixture model were drawn with an equal *a priori* probability $p_0 = 1/2$. The active and the passive testing algorithm where run on each sample. The MATLAB code for the active and the passive testing algorithms and these numerical experiments is available at http://www.math.sc.edu/~remi/Preprints/activetesting/

Figure 1 shows some typical active schedules $\{k_m(x_l)\}$. Let us consider, for example, the upper figure, which corresponds to active schedules when $y_l = 0$. The first measurement is fixed, $k_1(x) = 1$. In the most typical schedule $\{k_m^t(y=0)\}_{m=1}^N$, which is plotted with circles, the true class $Y = 0$ immediately gets a high enough likelihood, hence the following tests are done at the low frequencies, until there is no longer any low frequency available. The last measurements are thus made at high frequencies. An example of a less frequent schedule is displayed with crosses : after three tests on the signal x_l, the likelihood of the true class has become too small. Hence the active testing switches to measurements at high frequencies, that seem more adapted. However, after five more tests, the "mistake" is corrected and the algorithm switches back to measurements at low frequencies. When the pool of low frequencies is empty, the last measurements are done at high frequencies.

The lower figure displays an opposite behavior when $y_l = 1$. The first measurement, which is not adaptive, is on a low frequency. The most typical schedule $\{k_m^t(y=1)\}_{m=1}^N$ (circles) immediately switches to measurements at high frequencies. However, this first test is less reliable than a measure at a high frequency, hence it happens from time to time (crosses) that its result is misleading. In such a case, the algorithm goes on measuring low frequencies, until it realizes its mistake.

In this example, for every k, $\max(\lambda_k, 1/\lambda_k) = c^2$, hence the passive schedule might be any permutation of $\{1..N\}$. However, not all of them are equivalent for the fast estimation of $p_N(x)$ with $p_m(x)$, $m \ll N$. Actually, if the true class of x_l is $y_l = 0$, the schedule $\{k_m^t(y=0)\}_{m=1}^N$ should be good, but if $y_l = 1$ it will have a poor performance in terms of how fast $p_m(x_l)$ goes to $p_N(x_l)$. The schedule $\{k_m^t(y=1)\}_{m=1}^N$ will have the opposite behavior. In order to have a balanced performance that does not depend on the value of y_l, it is better to use a permutation that alternates measurements at low frequencies and at high-frequencies. We used

$$\{k_m^p\}_{m=1}^N := \{1, N, 2, N-1, 3, \ldots, N/2, N/2+1\}. \quad (19)$$

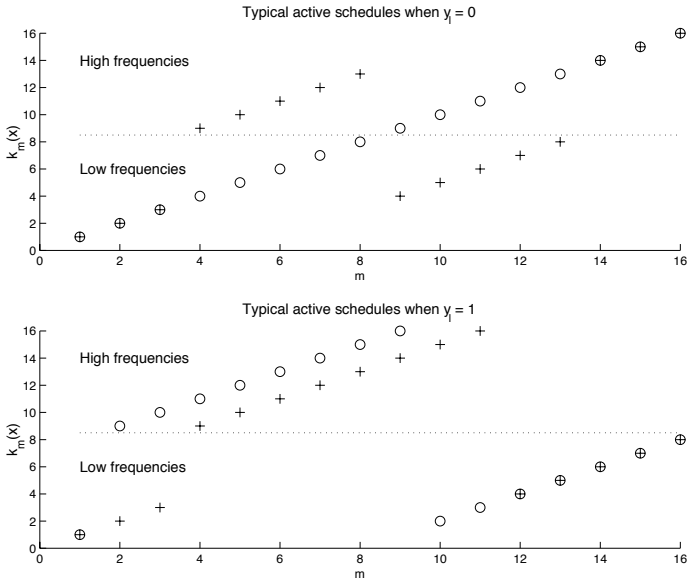

Fig. 1. Location $k_m(x)$ of the test as a function of m, with the active testing algorithm. (Top) when the underlying class is $y_l = 0$; (Bottom) : when $y_l = 1$. The dotted line separate low frequencies from high frequencies. On each figure, the most typical active schedule $\{k_m^t(y_l)\}_{m=1}^N$ is plotted with circles. In both cases, the first measurement is $k_1(x) = 1$; then if $y_l = 0$ (resp. $y_l = 1$), all low frequencies (resp. high frequencies) are measured; eventually, the remaining frequencies are measured. Examples of less typical schedules are displayed with crosses. They correspond to the fact that some measurements may be misleading.

Figure 2 compares the speed at which the average value of $|p_m(x) - p_N(x)|$ converges to zero with the passive schedule $\{k_m^p\}_{m=1}^N$ (plain line) and with the active testing algorithm (dashed line). After a first test $k_1 = 1$ which is common to both algorithms, the active testing algorithm provides in few steps quite a better estimate of $p_N(x)$ than the passive one. At the end, both algorithms give the same exact value $p_N(x) = \mathcal{P}(Y=0|X=x)$.

V. Conclusion

In this paper we have made a detailed study of active testing for an identification problem involving two Gaussian classes.

In the case of two Gaussian classes entirely characterized by their mean (*i.e.* with $\Sigma_0 = \Sigma_1$), we showed that active testing is indeed passive and corresponds to an approximation strategy similar to the Orthogonal Matching Pursuit. Our result is similar to that of Li [Li99] (see p. 44). There remains to study whether the "active" testing is still passive when $\Sigma_1 = \rho\Sigma_0$, $\rho \neq 1$ (see [Li99], Theorem 4.12, p. 60).

In the case of classes entirely characterized by their covariance structure, we obtained somewhat surprising re-

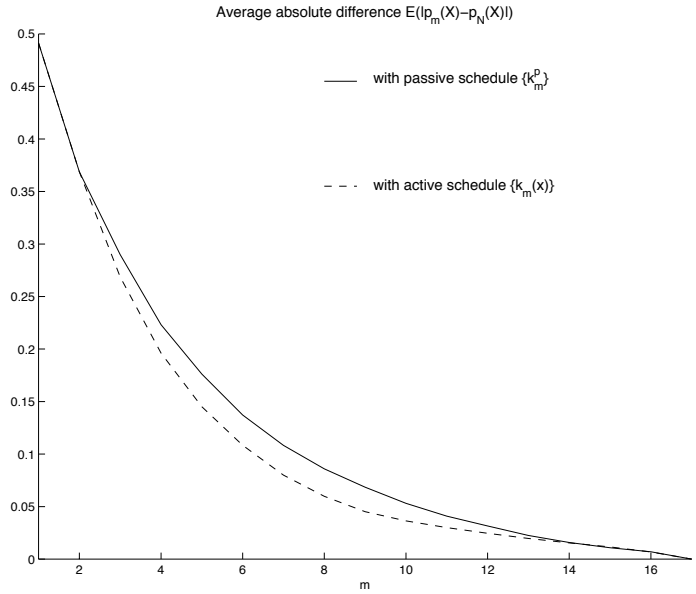

Fig. 2. Average value of $|p_m(x) - p_N(x)|$ with the passive testing algorithm (plain line) and the active testing algorithm (dashed line).

sults. First, the active testing is actually dependent on the analyzed data, taking the form of a binary decision tree. Moreover, the optimal schedule of observations can be seen as selecting the *smallest* components first, which is the opposite of the adaptive selection of components for approximation.

We provided numerical evidence that, with a small number of tests, the likelihood of each class can be estimated more accurately with an adaptive schedule of tests than with a fixed one. A suitable modification of the Sequential Probability Ratio Test [Wal45], [Wal49], [Fu68] is still needed to take full advantage of this desirable property.

We are currently trying to extend Theorem 2 to the selection of linear features $\{g_m(x)\}_{m=1}^N$ in a *limited dictionary* which does not contain the desired unit eigen-vectors. Such an extension will not only provide fast identification techniques, using some fast transform algorithm given by Computational Harmonic Analysis (*e.g.* wavepackets). It will also deal with the fact that, in practical applications, the true covariances Σ_0 and Σ_1 and the eigen-vectors $\{u_k\}$ are only estimated, hence innacurate. Such an extension will show how robust is active testing in a real pattern recognition process, when one has to take into account the effect of the learning stage.

Acknowledgments

The author wants to thank Stéphane Mallat and Emmanuel Bacry for their motivating interest in this research, and Donald Geman for his valuable support and remarks.

Appendix

For any unit vector g, the posterior distribution of $\langle X, g\rangle$ (after conditioning with $\{\langle X, g_l\rangle\}_{l=1}^m$) is a mixture of two one-dimensional Gaussian classes $\mathcal{N}(\mu_{m,y}[g], \sigma^2_{m,y}[g])$, $y = 0, 1$, with mixture parameter $p_m(X)$ (the *a posteriori* probability of $Y = 0$). In order to simplify the notations, we will usually not write the dependence of $\mu_y[g]$, $\sigma_y^2[g]$ and p on X and/or m. Using the invariance of mutual entropy [CT91] with respect to translations and dilations of X, it is easy to show that its maximization is equivalent to that of

$$\mathcal{I}\left(\left|\frac{\mu_1[g] - \mu_0[g]}{\sigma_0[g]}\right|, \frac{\sigma_1^2[g]}{\sigma_0^2[g]}, p\right) \tag{20}$$

(see definition 1). We study the variations of $\lambda \mapsto \mathcal{I}(0, \lambda, p)$ and $\eta \mapsto \mathcal{I}(\eta, 1, p)$, then give the expression of $\mu_y[g]$ and $\sigma_y^2[g]$ as a function of g and $\{g_l\}_{l=1}^m$. Eventually we proceed to the proofs of our theorems.

I. Variations of the mutual entropy

Lemma 5: The function $\lambda \mapsto \mathcal{I}(0, \lambda, p)$ is decreasing on $(0, 1]$ and increasing on $[1, \infty)$.

Lemma 6: The even function $\eta \mapsto \mathcal{I}(\eta, 1, p)$ is increasing with $|\eta|$.

Notations 1: Let $\phi(t) = 1/\sqrt{2\pi}e^{-\frac{t^2}{2}}$ be the Gaussian pdf. The entropy [CT91] of $Z \sim \mathcal{N}(\eta, \lambda)$ is $\frac{1}{2}\log 2\pi e\lambda$. So as to simplify future computations, let $\nu = \lambda^{-1/2}$. The pdf of the mixture is $h(y) = p\phi(y - \eta) + (1 - p)\nu\phi(\nu y)$. Let us denote $\psi(x) = x \log x$.

The mutual entropy can be written

$$\mathcal{I}(\eta, \nu^{-2}, p) = -\int \psi\left[h(y)\right] dy - \frac{1}{2}\log 2\pi e + (1 - p)\log\nu \tag{21}$$

Proof of Lemma 5

We compute $\frac{\partial}{\partial\nu}\mathcal{I}(0, \nu^{-2}, p)$

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

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 $\frac{\partial}{\partial\nu}\mathcal{I}(0,\nu^{-2},p)$ goes on

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

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

Proof of Lemma 6

We compute $\frac{\partial}{\partial\eta}\mathcal{I}(\eta,1,p)$

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

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. □.

II. Conditional expectation and variance

Now we want an expression of the arguments of the functional $\mathcal{I}$ in (20) as a function of g and $\{g_l\}_{l=1}^m$.

Lemma 7: The conditional expectation of $\langle X,g\rangle$ is the following random variable

$$\begin{aligned}
\mu_{m,y}[g] &:= \mathbb{E}\Big\{\ \langle X,g\rangle\ \Big|Y=y,\{\langle X,g_l\rangle\}_{l=1}^m\Big\}\\
&= \langle\vec{\mu}_y,g\rangle + \left\langle X-\vec{\mu}_y, P_{m,\Sigma_y}g\right\rangle \qquad (22)
\end{aligned}$$

and its conditional variance is

$$\begin{aligned}
\sigma^2_{m,y}[g] &:= \mathbb{E}\Big\{\left(\langle X,g\rangle - \mu_{m,y}[g]\right)^2\Big|Y=y,\{\langle X,g_l\rangle\}_{l=1}^m\Big\}\\
&= \left\|(Id-P_{m,\Sigma_y})g\right\|^2_{\Sigma_y} \qquad (23)
\end{aligned}$$

where $\mathcal{V}_m = \text{span}\{g_l, 1\le l\le m\}$ and P_{m,Σ_y} is the orthogonal projector onto $\mathcal{V}_m$ with respect to the inner product $\langle .,.\rangle_{\Sigma_y} := \langle .,\Sigma_y.\rangle$, *i.e.* the projector onto $\mathcal{V}_m$ parallel to $\mathcal{V}_m^{\perp_{\Sigma_y}} = (\Sigma_y\mathcal{V}_m)^\perp = \Sigma_y^{-1}\mathcal{V}_m^\perp$.

Proof

By definition, $\mu_{m,y}[g]$ is the orthogonal projection of $\langle X,g\rangle$ onto the space of random variables that are measurable with respect to Y and $\{\langle X,g_l\rangle\}_{l=1}^m$. We can check that the right hand side in (22) is a linear function of the conditioning variables, which implies measurability. Moreover, it is Gaussian and decorrelated from the conditioning variables $\langle X,g_l\rangle$, hence it is independent from them. Hence the first result.

Let $W_y = X - \vec{\mu}_y$. Using (22) we can write

$$\begin{aligned}
\sigma^2_{m,y}[g] &= \mathbb{E}\Big\{\left(\langle W_y,g\rangle - \left\langle W_y, P_{m,\Sigma_y}g\right\rangle\right)^2\Big|\\
&\qquad Y=y,\{\langle X,g_l\rangle\}_{l=1}^m\Big\}.
\end{aligned}$$

Conditionally on $Y=y$, W_y is a centered Gaussian random variable, and it is decorrelated with, thus [Pap84] independent of, $\{\langle W_y,g_l\rangle\}_{l=1}^m$. As a result

$$\begin{aligned}
\sigma^2_{m,y}[g] &= \mathbb{E}\Big\{\ \langle W_y,(Id-P_{m,\Sigma_y})g\rangle^2\ \Big|Y=y\Big\}\\
&= \left\langle(Id-P_{m,\Sigma_y})g,\Sigma_y(Id-P_{m,\Sigma_y})g\right\rangle
\end{aligned}$$

which gives the result. □.

III. Proof of theorem 1

As $\Sigma_1=\Sigma_0=\Sigma$, the projector $P_{m,\Sigma_1}=P_{m,\Sigma_0}=P_m$ is independent of y. Equation (23) can be written as $\sigma^2_y[g] = \langle R_m g, \Sigma R_m g\rangle$ with $R_m = Id - P_m$. As a result $\sigma_1^2/\sigma_0^2 = 1$ for all g, and Lemma 6 shows that $g_{m+1}(x) = \arg\max_{g\in\mathcal{D}}|\eta_m[g]|$ where $\eta_m[g] := |(\mu_1[g]-\mu_0[g])/\sigma_0[g]|$. This choice does not depend on $p_m(x)$, and by induction one easily gets its independence from x. The family $\{g_m\}_{m=1}^N$ can be determined off-line using only $\vec{\mu}_1$, $\vec{\mu}_0$ and Σ. Let us see how this works.

Let us define on $\mathbb{R}^N$ a new Euclidian structure with the weighted inner product $\langle .,.\rangle_\Sigma = \langle .,\Sigma.\rangle$ and its associated weighted norm $\|.\|_\Sigma$. Equation (22) gives $\mu_1[g]-\mu_0[g] = \langle\vec{\mu}_1-\vec{\mu}_0,R_mg\rangle = \left\langle\Sigma^{-1}(\vec{\mu}_1-\vec{\mu}_0),R_mg\right\rangle_\Sigma$. As a result

$$g_{m+1} = \arg\max_{g\in\mathcal{D}}\left|\left\langle\Sigma^{-1}(\vec{\mu}_1-\vec{\mu}_0),\frac{R_mg}{\|R_mg\|_\Sigma}\right\rangle_\Sigma\right|. \qquad (24)$$

With respect to this Euclidian structure the Gram-Schmidt orthonormalization of $\{g_m\}_{m=1}^N$ is precisely the family $u_m := R_{m-1}g_m/\left\|R_{m-1}g_m\right\|_\Sigma$, hence

$$\left\|P_{m+1}\Sigma^{-1}(\vec{\mu}_1-\vec{\mu}_0)\right\|^2_\Sigma = \sum_{l=1}^{m+1}\left|\left\langle\Sigma^{-1}(\vec{\mu}_1-\vec{\mu}_0),u_l\right\rangle^2_\Sigma\right|. \qquad (25)$$

The greedy choice (24) thus maximizes the increase in the grabbed energy

$$\begin{aligned}
&\left|\left\langle\Sigma^{-1}(\vec{\mu}_1-\vec{\mu}_0),\frac{R_mg}{\|R_mg\|_\Sigma}\right\rangle\right|^2_\Sigma\\
&= \left\|P_{m+1}\Sigma^{-1}(\vec{\mu}_1-\vec{\mu}_0)\right\|^2_\Sigma - \left\|P_m\Sigma^{-1}(\vec{\mu}_1-\vec{\mu}_0)\right\|^2_\Sigma \qquad (26)
\end{aligned}$$

The selected atoms $\{g_m\}_{m=1}^N$ are thus obtained by a variant of the Orthogonal Matching Pursuit [MZ93] [Dav94] [PRK93] on the signal $\Sigma^{-1}(\vec{\mu}_1-\vec{\mu}_0)$. □.

IV. Proof of theorem 2

Let $\{u_k\}_{k=1}^N$ be a basis of unit eigen-vectors for $\Sigma_0^{-1}\Sigma_1$: $\Sigma_0^{-1}\Sigma_1 u_k = \lambda_k u_k$, and $\{g_m(x)\}_{m=1}^N$ a AFS. Let us denote $\mathcal{V}_m = \mathcal{V}_m(x) := \text{span}\{g_l(x), 1 \leq l \leq m\}$, and $\mathcal{V}_0 = \mathcal{V}_0(x) := \{0\}$. We shall prove by induction that there exists $\{k_m(x)\}_{m=1}^N$ such that for $0 \leq m \leq N$

$$\mathcal{V}_m(x) = \text{span}\{u_{k_l(x)}, 1 \leq l \leq m\}. \tag{27}$$

The relation is clearly true for $m = 0$ because both sides are $\{0\}$. Let us show that if (27) holds for $m-1$, then it is also true for m.

The induction hypothesis implies that $\Sigma_0^{-1}\Sigma_1\mathcal{V}_{m-1} = \mathcal{V}_{m-1}$, *i.e.*

$$\Sigma_0\mathcal{V}_{m-1} = \Sigma_1\mathcal{V}_{m-1}. \tag{28}$$

Hence the two projectors P_{m-1,Σ_y}, $y = 0, 1$ onto $\mathcal{V}_{m-1}$ parallel to $(\Sigma_y\mathcal{V}_{m-1})^\perp$ are equal. Using (22) we get $\mu_{m-1,1}[g] - \mu_{m-1,0}[g] = 0$. From now on, let us denote $P_{m-1} := P_{m-1,\Sigma_1} = P_{m-1,\Sigma_0}$ and $R_{m-1} = Id - P_{m-1}$. From (20), Lemma 5 and (23), we know that "the" best atom $g_m(x)$ corresponds to an extremum of $\sigma^2_{m-1,1}[g]/\sigma^2_{m-1,0}[g] = \langle R_{m-1}g, \Sigma_1 R_{m-1}g\rangle \,/\, \langle R_{m-1}g, \Sigma_0 R_{m-1}g\rangle$. Using Lagrange multipliers, such an extremum is obtained when, for some λ,

$$R^*_{m-1}\Sigma_1 R_{m-1}g = \lambda R^*_{m-1}\Sigma_0 R_{m-1}g. \tag{29}$$

Linear algebra shows that condition (29) is equivalent to $(\Sigma_1 - \lambda\Sigma_0)R_{m-1}g \in (\text{Im}R_{m-1})^\perp = \Sigma_y\mathcal{V}_{m-1}, y = 0, 1$, that is to say $(\Sigma_0^{-1}\Sigma_1 - \lambda Id)R_{m-1}g \in \mathcal{V}_{m-1} = \text{Ker}R_{m-1}$.

From (28), we know that the projector R_{m-1} commutes with $\Sigma_0^{-1}\Sigma_1$, because its range $\Sigma_1^{-1}\mathcal{V}^\perp_{m-1}$ and kernel $\mathcal{V}_{m-1}$ are stable under $\Sigma_0^{-1}\Sigma_1$. Thus (29) $\Leftrightarrow$ $R_{m-1}(\Sigma_0^{-1}\Sigma_1 - \lambda Id)g \in \text{Ker}R_{m-1}$ $\Leftrightarrow$ $R_{m-1}(\Sigma_0^{-1}\Sigma_1 - \lambda Id)g = 0$ $\Leftrightarrow$ $(\Sigma_0^{-1}\Sigma_1 - \lambda Id)R_{m-1}g = 0$ $\Leftrightarrow$ $R_{m-1}g$ is either zero or an eigen vector of $\Sigma_0^{-1}\Sigma_1$. But $R_{m-1}g = g - P_{m-1}g$ cannot be zero, for it would mean $g \in \text{Im}P_{m-1} = \mathcal{V}_{m-1}$, and such a g cannot bring *any* additional information on the class. Hence (27) is true at step m.

□.

V. Proof of Lemmas 1 and 2

We keep the previous notations. We know that $k_m(x) \notin \{k_l(x)\}_{l=1}^{m-1}$ and (29) shows that for $k \notin \{k_l(x)\}_{l=1}^{m-1}$, $\sigma^2_{m-1,1}[u_k]/\sigma^2_{m-1,0}[u_k] = \lambda_k$, hence

$$k_m(x) = \arg \max_{k \notin \{k_l(x)\}_{l=1}^{m-1}} \mathcal{I}\left(0, \lambda_k, p_{m-1}(x)\right) \tag{30}$$

We can thus derive (9) from Lemma 5. Lemma 2 is another immediate consequence of Lemma 5. □.

VI. Proof of Lemma 3

We start by a technical lemma.

Lemma 8: The mutual entropy can be developed as

$$\mathcal{I}(0, \nu^{-2}, p) = \frac{1-p}{2}\left\{-1 + \frac{1}{\nu^2} + \log\nu^2\right\} + o(1-p) \tag{31}$$

Proof

We use the notations of section A. Let us denote $r(y) = \frac{1-p}{p}\frac{\nu\phi(\nu y)}{\phi(y)}$, which enables us to write $\int \psi[h] = \int h\log h$ as $\int h\log p\phi + \int h\log[1+r]$

$$\begin{aligned} &= \int h(y)[\log p - \frac{\log 2\pi}{2} - \frac{y^2}{2}]dy + \int h\log[1+r] \\ &= \log p - \frac{\log 2\pi}{2} - \int h(y)\frac{y^2}{2}dy + \int h\log[1+r] \end{aligned} \tag{32}$$

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

$$\int h(y)y^2 = p\int \phi(y)y^2 + (1-p)\int \nu\phi(\nu y)y^2 = p + \frac{(1-p)}{\nu^2}. \tag{33}$$

Collecting (21), (32) and (33), we get the estimate

$$\begin{aligned} \mathcal{I}(0, \nu^{-2}, p) &= -\log p + \frac{\log 2\pi}{2} + \frac{p}{2} + \frac{1-p}{2\nu^2} - \frac{\log 2\pi e}{2} \\ &\quad + (1-p)\log\nu - \int h\log[1+r] \\ &= \frac{1-p}{2}\left\{-1 + \frac{1}{\nu^2} + \log\nu^2\right\} \\ &\quad - \log p - \int h\log[1+r] \end{aligned} \tag{34}$$

Let us now estimate the remaining integral term. The Dominated Convergence Theorem shows that

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

hence

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

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

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

which leads to

$$\int h\log[1+r] = (1-p) + o(1-p). \tag{35}$$

Combined with the development $\log p = \log(1-(1-p)) = -(1-p) + o(1-p)$, equations (34) and (35) finally lead to (31). □.

Lemma 3 is actually a corollary of Lemma 8. Let $\underline{\lambda} < 1 < \overline{\lambda}$ and define

$$\begin{aligned} \Delta_1(\underline{\lambda}, \overline{\lambda}) &:= \lim_{p\to 1}\frac{2}{1-p}\left(\mathcal{I}(0, \overline{\lambda}, p) - \mathcal{I}(0, \underline{\lambda}, p)\right) \\ \Delta_0(\underline{\lambda}, \overline{\lambda}) &:= \lim_{p\to 0}\frac{2}{p}\left(\mathcal{I}(0, \overline{\lambda}, p) - \mathcal{I}(0, \underline{\lambda}, p)\right). \end{aligned}$$

It is easy to show by a change of variables that $\mathcal{I}(0, 1/\lambda, 1-p) = \mathcal{I}(0, \lambda, p)$, hence using Lemma 8 we get

$$\Delta_1(\underline{\lambda}, \overline{\lambda}) = a(\overline{\lambda}) - a(\underline{\lambda}) \tag{36}$$

$$\Delta_0(\underline{\lambda}, \overline{\lambda}) = b(\overline{\lambda}) - b(\underline{\lambda}). \tag{37}$$

which gives (12) and (11) □.

VII. Size of AFS tree

The branches of the AFS tree $\mathcal{T}(\Sigma_0^{-1}\Sigma_1)$ correspond to permutations of the eigenvalues of $\Sigma_0^{-1}\Sigma_1$ which share the same first term λ_{k_1}. Along each branch, the subsequence of eigenvalues larger than one (resp. smaller than one) is decreasing (resp. increasing). Let p (resp. q) the number of eigenvalues of $\Sigma_0^{-1}\Sigma_1$ that are larger (resp. smaller) than one. Using a classical result of combinatorics , the total number of such sequences cannot exceed $C_{N-1}^{p-1} = C_{N-1}^{q}$ (resp. $C_{N-1}^{q-1} = C_{N-1}^{p}$) if $\lambda_{k_1} > 1$ (resp. $\lambda_{k_1} < 1$). Because some nodes in the tree may have only one child, these actually give upper bounds on the total number of branches. The worst cases are $p - 1 = \lfloor (N-1)/2 \rfloor$ (resp. $p = \lfloor (N-1)/2 \rfloor$). Combined with Stirling's formula they give the overall upper bound

$$C_{N-1}^{\lfloor (N-1)/2 \rfloor} \sim \frac{2^N}{\sqrt{2\pi N}}. \tag{38}$$

References

[AG97] Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. *Neural Computation*, 9:1545–1588, 1997.

[AGW97] Y. Amit, D. Geman, and K. Wilder. Joint induction of shape features and tree classifiers. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(11):1300–1305, November 1997.

[AM99] Y. Amit and A. Murua. Speech recognition using randomized relational decision trees. Technical Report 487, Department of Statistics, University of Chicago, April 1999.

[BFOS84] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification And Regression Trees*. Chapman & Hall, 1984.

[BP85] C.S. Burrus and T.W. Parks. *DFT/FFT and Convolution Algorithms : Theory and Implementation*. John Wiley and Sons, New York, 1985.

[CD99] S. Chen and D.L. Donoho. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, January 1999.

[Com94] P. Comon. Independent component analysis, a new concept ? *Signal Process.*, 36:287–314, 1994.

[CT91] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications. Wiley-Interscience, 1991.

[CW92] R. Coifman and M.V. Wickerhauser. Entropy-based algorithms for best basis selection. *IEEE Trans. Inform. Theory*, 38(2):713–718, March 1992.

[Dau88] I. Daubechies. Orthonormal bases of compactly supported wavelets. *Commun. on Pure and Appl. Math.*, 41:909–996, November 1988.

[Dau92] I. Daubechies. *Ten Lectures on Wavelets*. SIAM, 1992.

[Dav94] G. Davis. *Adaptive Nonlinear Approximations*. PhD thesis, New York University, September 1994.

[DeV98] R. A. DeVore. Nonlinear approximation. *Acta Numerica*, pages 51–150, 1998.

[FJ98] M. Frigo and S.G. Johnson. FFTW: An adaptive software architecture for the FFT. In *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP'98)*, volume 3, page 1381, 1998.

[FS81] J.H. Friedman and W. Stuetzle. Projection pursuit regression. *J. Amer. Stat. Assoc.*, 76:817–823, 1981.

[FT74] J.H. Friedman and J.W. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. on Computers*, C-23:881–889, 1974.

[Fu68] K.S. Fu. *Sequential Methods in Pattern Recognition and Machine Learning*, volume 52 of *Mathematics in Science and Engineering*. Academic Press, 1968.

[Fuk72] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Electrical Science. Academic Press, 1972.

[GDR+96] R. Gribonval, Ph. Depalle, X. Rodet, E. Bacry, and S. Mallat. Sound signals decomposition using a high resolution matching pursuit. In *Proc. Int. Computer Music Conf. (ICMC'96)*, pages 293–296, August 1996.

[GJ96] D. Geman and B. Jedynak. An active testing model for tracking roads in satellite images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 18(1):1–14, January 1996.

[GL00] D. Geman and C. Li. Active testing and bayesian sequential classification. in preparation, 2000.

[Hub85] P. J. Huber. Projection pursuit. *The Annals of Statistics*, 13(2):435–475, 1985.

[JCMW98] S. Jaggi, W.C. Carl, S. Mallat, and A.S. Willsky. High resolution pursuit for feature extraction. *J. Applied and Computational Harmonic Analysis*, 5(4):428–449, October 1998.

[JGL99] Q. Jiang, S.S. Goh, and Z. Lin. Local discriminant time-frequency atoms for signal classification. *Signal Process.*, 72:47–52, 1999.

[Kay93] S.M. Kay. *Fundamentals of Statistical Signal Processing : Estimation Theory*. Signal Processing. Prentice Hall, 1993.

[Li99] Chunming Li. *Classification by Active Testing with Applications to Imaging and Change Detection*. PhD thesis, Univ. of Massachussets Amherst, February 1999.

[LL99] B. Liu and S.-F. Ling. On the selection of informative wavelets for machine diagnosis. *Journal of Mechanical Systems and Signal Processing*, 13(1):145–162, 1999. (ID mssp.1998.0177).

[Mal89] S. Mallat. A theory for multiresolution signal decomposition : the wavelet representation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, pages 674–693, July 1989.

[Mal98] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 1998.

[MC97] M.R. McClure and L. Carin. Matching pursuits with a wave-based dictionary. *IEEE Trans. Signal Process.*, 45(12):2912–2927, December 1997.

[McL92] G. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York, 1992.

[MZ93] S. Mallat and Z. Zhang. Matching pursuit with time-frequency dictionaries. *IEEE Trans. Signal Process.*, 41(12):3397–3415, December 1993.

[Pap84] A. Papoulis. *Probability, Random Variables and Stochastic Processes*. McGraw-Hill, New York, NY, 2nd edition, 1984.

[Phi98] P. Jonathon Phillips. Matching pursuit filters applied to face identification. *IEEE Trans. Image Process.*, 7(8):1150–1164, August 1998.

[PRK93] Y.C. Pati, R. Rezaiifar, and P.S. Krishnaprasad. Orthonormal matching pursuit : recursive function approximation with applications to wavelet decomposition. In *Proceedings of the* 27^{th} *Annual Asilomar Conf. on Signals, Systems and Computers*, November 1993.

[QC94] S. Qian and D. Chen. Signal representation using adaptive normalized Gaussian functions. *Signal Process.*, 36(1):1–11, 1994.

[Sai98] N. Saito. Least statistically-dependent basis and its application to image modeling. In A.F. Laine, M.A. Unser, and A. Aldroubi, editors, *Wavelet Applications in Signal and Image Processing VI*, volume 3458 of *Proc. SPIE*, pages 24–37, San Diego CA., July 1998.

[SC94] N. Saito and R.R. Coifman. Local discriminant bases. In A.F. Laine and M.A. Unser, editors, *Mathematical Imaging : Wavelet Applications in Signal and Image Processing II*, volume 2303 of *Proc. SPIE*, pages 2–14, San Diego, CA., July 1994.

[Tor91] B. Torrésani. Wavelets associated with representations of the affine Weyl-Heisenberg group. *J. Math. Phys.*, 32:1273–1279, May 1991.

[Wal45] A. Wald. Sequential tests of statistical hypothesis. *Annals of Math. Statist.*, 16(2):117–186, jun 1945.

[Wal49] A. Wald. *Sequential Analysis*. John Wiley & Sons, Inc., New York, 1949.

[Wic91] M. V. Wickerhauser. Fast approximate factor analysis. In *Curves and Surfaces in Computer Vision and Graphics II*, volume 1610 of *Proc. SPIE*, pages 23–32, Boston, October 1991.