SUBSPACE LEARNING IN MINIMAX DETECTION

Raja Fazliza R. Suleiman, David Mary and André Ferrari

Laboratoire J.-L. Lagrange, UMR7293,
Université de Nice Sophia Antipolis, CNRS, Observatoire de la Côte d'Azur,
Campus Valrose, 06108 Nice Cedex 02, FRANCE
Email: {raja.fazliza, david.mary, andre.ferrari}@unice.fr

ABSTRACT

We consider the problem where a large known library of L alternatives is available and we wish to maximize the detection power in a worst case scenario. The considered minimax detection approach relies on a GLR test allied to a sparsity constraint. This approach conditions the optimization of the target subspaces, in number $r \ll L$. While the exact solution of the minimax optimization problem can be found for $r = 1$, the problem for $r > 1$ is more intricate and we propose two algorithms aimed at finding an approximate solution. The proposed algorithms are illustrated on a face database and on hyperspectral data and are shown to improve on the $r = 1$ case.

***Index Terms*—** Minimax, detection, dictionary learning, sparsity, classification.

1. INTRODUCTION AND PRIOR WORKS

The minimax theory was introduced by Von Neumann [1] in game theory with the objective of minimizing a loss function in a worst case scenario. Minimax principles have led to numerous theoretical results and robust methods in various domains [2, 3] including decision theory [4]. In signal processing, minimax risk arguments paved in the 90's the road of sparsity promoting methods based on thresholding functions for denoising and inverse problems [5].

We consider here a minimax strategy for the following detection problem. Under $\mathcal{H}_1$, one target signature $\mathbf{s}_i$ is activated, among a (possibly very large) set of known alternatives forming a library $\mathbf{S}$. The signature's amplitude and its index i are considered unknown. This problem is instantiated in telecommunications, when detecting an active symbol among a known dictionary of symbols [6], in medical applications for the identification of pathological signatures [7] or in hyperspectral astrophysical data [8].

In [8], we showed that the detection performance of a procedure testing all alternatives may drastically drop for some alternatives as the cardinality of $\mathbf{S}$ increases. For large libraries, a standard approach is indeed to reduce the number of target subspaces by subspace learning (*e.g.* via sample mean, SVD, K-SVD, etc., [9]). Another important observation of [8] is that such standard approaches tend to focus on an *average* behavior: consequently, some alternatives may lie quite far from the learned target subspace(s) and the associated detection power will be low.

In this framework, the minimax approach aims at minimizing such effects. This objective can also be stated as maximizing the minimum detection power, which is a *maximin* problem. In the rest of the paper, we shall generically use the term "minimax" to refer to this approach.

This work was supported by the CNRS program MASTODONS. R.F.R. Suleiman is funded by Majlis Amanah Rakyat (MARA).

The minimax detection considered below relies on a GLRT (Generalized Likelihood Ratio test) allied to a sparsity constraint. This approach conditions the target subspace learning which can be viewed as a dictionary learning problem. Sparsity-based dictionary learning techniques are now widely used in the literature. They may outperform generic dictionaries (*e.g.* wavelets) for image denoising [10, 11] and be very efficient for other tasks such as blind source separation [12] or classification for object recognition [13, 14]. As we will show, traditional learning algorithms do not perform well with respect to (w.r.t.) minimax objectives, which lead to specific optimization issues and call for dedicated learning algorithms.

In a first approach presented in [8], we considered the particular case where the learned target subspace is restricted to one dimension (*i.e.*, the learned dictionary has a single atom). While this approach was shown to improve the minimax power w.r.t. classical learning algorithms such as K-Means [15] or K-SVD [10], such an extreme reduction in dimension might be exaggerated w.r.t. the intrinsic diversity of the library. We thus expect that learning a few, instead of one subspace may improve the minimax power. Besides, learning more atoms should provide a better average representation of the alternatives, and thus might also increase the average performance.

The goal of this paper is to investigate how to derive algorithms for subspace learning with minimax objectives and to evaluate their performances. Sec.2 formalizes the considered detection problem and the corresponding GLRT. The considered reduced dimension model, associated GLR and optimization criterion are studied in Sec.3. Because the optimized minimax dictionary cannot be obtained by standard optimization techniques when more than one subspace are to be learned, we propose two algorithms (Sec.4) aimed at finding an approximate solution. The first is an adaptation of the K-SVD method where the dictionary update stage is replaced by the exact solution of the 1-dimensional minimax problem. The second uses the same dictionary update stage, but samples the distribution in a greedy manner to open new classes. These algorithms are illustrated by numerical results on images and spectra in Sec.5.

2. EXACT DETECTION MODEL AND ASSOCIATED GLRT

The detection problem discussed in introduction can be written as

$$\begin{cases} \mathcal{H}_0 & : \mathbf{x} = \mathbf{n}, & \mathbf{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \\ \mathcal{H}_1 & : \mathbf{x} = \mathbf{S}\boldsymbol{\alpha} + \mathbf{n}, & ||\boldsymbol{\alpha}||_0 = 1 \end{cases} \tag{1}$$

where $\mathbf{x}$ and $\mathbf{n} \in \mathbb{R}^N$. The known alternatives are here collected as columns of the library $\mathbf{S} \in \mathbb{R}^{N \times L} = [\mathbf{s}_1, \ldots, \mathbf{s}_L]$. We assume that the set $\mathbf{s}_i$ belong to the positive orthant (as for instance the test data considered later on) and that they are normalized ($\|\mathbf{s}_i\|_2^2 = 1$, $i = 1, \ldots, L$). The 1-sparse constraint on the unknown vector

$\boldsymbol{\alpha} \in \mathbb{R}^L$ (only one non-zero element) indicates that under $\mathcal{H}_1$, only one signal $\mathbf{s}_i$ of (unknown) amplitude α_i is activated. This model assumes that the covariance matrix (say, $\mathbf{R}$) is known and equal under both hypotheses, in which case "whitening" the data by $\mathbf{R}^{-\frac{1}{2}}$ leads to a model of the form (1). The GLR for (1) involves the constrained Maximum Likelihood (ML) estimate of $\boldsymbol{\alpha}$:

$$T_{GLR}(\mathbf{x},\mathbf{S}) : \max_{\boldsymbol{\alpha}:||\boldsymbol{\alpha}||_0=1} \frac{p(\mathbf{x}|\mathbf{S}\boldsymbol{\alpha})}{p(\mathbf{x}|\mathbf{0})} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma', \tag{2}$$

with γ' a threshold. Noting $i^* = \arg\max_{i=1,\ldots,L} |\mathbf{s}_i^\top \mathbf{x}|$, (where superscript $^\top$ denotes transposition), the non-zero component of the constrained ML estimate of $\boldsymbol{\alpha}$ is $\alpha_{i^*}^{ML} = \mathbf{s}_{i^*}^\top \mathbf{x}$. Plugging this in (2) yields the *extreme value* or *Max test* (see *e.g.* [16])

$$T_{max}(\mathbf{x},\mathbf{S}) = \max_{i=1,\ldots,L} |\mathbf{s}_i^\top \mathbf{x}| \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma, \tag{3}$$

where $\gamma = \sqrt{2\ln\gamma'}$ (γ can always be computed by Monte Carlo simulations; this is necessary when $\mathbf{S}$ is not orthogonal).

While the GLR is a natural approach benefiting from optimality properties in an asymptotic (w.r.t. to the amount of available data) setting, the GLR may present in the considered framework two undesirable effects for high cardinality ($L \gg N$) libraries $\mathbf{S}$ [8]. First, the detection performance may happen to drastically drop for some alternatives that are dissimilar to the others as L increases. The reason is that a larger number of tested alternatives increases the rate of false alarms in a relatively higher proportion than the detection rate. Second, the computation complexity of test (3) scales as L, which may be prohibitive for very large L [17]. These aspects suggest to devise tests focusing on few target subspaces, which obviously improves the second effect. As for the first effect, reducing the number of target subspaces by classical techniques (like MOD or K-SVD) is not enough, because such optimized dictionaries represent well the core of the distribution of the alternatives $\mathbf{s}_i$, but not the marginal ones − which are critical for minimax detection performance. Hence, the objective of maximizing the minimum detection power should be explicitly accounted for in the learning procedure.

3. CONSTRAINED MODEL OF REDUCED DIMENSION

For problem (1), we investigate a GLR operating on the following constrained model of reduced dimension $r \ll L$ under $\mathcal{H}_1$:

$$\begin{cases} \mathcal{H}_0 & : \mathbf{x} = \mathbf{n}, & \mathbf{n} \sim \mathcal{N}(\mathbf{0},\mathbf{I}) \\ \mathcal{H}_1 & : \mathbf{x} = \mathbf{D}\boldsymbol{\beta} + \mathbf{n}, & ||\boldsymbol{\beta}||_0 = 1 \end{cases}, \tag{4}$$

where $\boldsymbol{\beta}$ is unknown and contains only one non zero element, to encourage the axes of $\mathbf{D}$ to align with the main "modes" (possibly represented by isolated alternatives of $\mathbf{S}$) of the distribution of the set $\mathbf{s}_i$ over the unit sphere. The columns of $\mathbf{D} = [\mathbf{d}_1,\ldots,\mathbf{d}_r]$ are normalized and otherwise unconstrained.

Similarly to (2)-(3), the GLR for (4) and for a given $\mathbf{D}$ leads to

$$T_{\mathbf{D}}(\mathbf{x}) : \max_{j=1,\ldots,r} |\mathbf{d}_j^\top \mathbf{x}| \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \xi \Leftrightarrow \max_{j=1,\ldots,r} (\mathbf{d}_j^\top \mathbf{x})^2 \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \xi^2 \tag{5}$$

where ξ is a threshold. The question is now to optimize $\mathbf{D}$ of size $N \times r$ for test (5), so as to maximize the minimum probability of detection $\mathrm{P_{Det}}$ at a fixed probability of false alarm ($\mathrm{P_{FA_0}}$) , that is,

$$\begin{aligned} \max_{\mathbf{D}} \quad & \min_{i=1,\ldots,L} \mathbb{P}\left(\max_{j=1,\ldots,r} (\mathbf{d}_j^\top \mathbf{x})^2 > \xi^2 | \mathcal{H}_1, \mathbf{s}_i\right) \\ \text{subject to} \quad & \mathbb{P}\left(\max_{j=1,\ldots,r} (\mathbf{d}_j^\top \mathbf{x})^2 > \xi^2 | \mathcal{H}_0\right) \le \mathrm{P_{FA_0}}, \\ & \|\mathbf{d}_j\|_2 = 1,\ j = \{1,\ldots,r\}. \end{aligned} \tag{6}$$

In the analysis below, all alternatives are considered of the same amplitude to be comparable in terms of SNR. We assume without loss of generality unit amplitude and we denote by $\mathbf{s}_\ell$ the target signal under $\mathcal{H}_1$ of (1).

3.1. Case $r = 1$: Exact solution to problem (6)

Consider first the case $r = 1$, which was treated in [8]. Here, $\mathbf{d} = \mathbf{D}$ and the GLR (5) becomes $T_{\mathbf{d}}(\mathbf{x}) : (\mathbf{d}^\top \mathbf{x})^2 \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \xi^2$. Let $c = \mathbf{d}^\top \mathbf{x}$. It can be proved that: $c^2|\mathcal{H}_0 \sim \chi_1^2$ and $c^2|\mathcal{H}_1, \mathbf{s}_\ell \sim \chi_{1,\lambda}^2$, which denotes a noncentral chi-squared distribution with noncentrality parameter $\lambda = \mathbf{s}_\ell^\top(\mathbf{d}\mathbf{d}^\top)\mathbf{s}_\ell = (\mathbf{d}^\top \mathbf{s}_\ell)^2$. Hence,

$$\mathrm{P_{FA}} = \mathbb{P}\left(T_{\mathbf{d}} > \xi^2 \,|\mathcal{H}_0\right) = 1 - \Phi_{\chi_1^2}(\xi^2),$$

$$\mathrm{P_{Det}}(\mathbf{s}_\ell,\mathbf{d}) = \mathbb{P}(T_{\mathbf{d}} > \xi^2 | \mathcal{H}_1, \mathbf{s}_\ell) = 1 - \Phi_{\chi^2_{1,\lambda}}(\xi^2) = Q_{\frac{1}{2}}(\sqrt{\lambda}, \xi),$$

where Φ_v denotes the cumulative distribution function of the random variable v and $Q_{\frac{1}{2}}(\sqrt{\lambda},\xi)$ is the generalized Marcum-Q function of order $1/2$. This function is monotonically increasing (for $\sqrt{\lambda} > 0$ and $\xi \ge 0$) in its first argument $\sqrt{\lambda} = |\mathbf{d}^\top \mathbf{s}_\ell|$ [18]. The $\mathrm{P_{FA}}$ is indeed independent of $\mathbf{s}_\ell$ and $\mathbf{d}$ in this case, so maximizing the minimum $\mathrm{P_{Det}}$ under $\mathcal{H}_1$ at fixed $\mathrm{P_{FA}}$ leads to the optimization problem

$$\mathbf{d}^* = \arg\max_{\mathbf{d}:\|\mathbf{d}\|_2=1} \min_{i=1,\ldots,L} (\mathbf{d}^\top \mathbf{s}_i)^2 = \arg\max_{\mathbf{d}:\|\mathbf{d}\|_2=1} \min_{i=1,\ldots,L} |\mathbf{d}^\top \mathbf{s}_i|, \tag{7}$$

which can be solved by a standard QP solver [8].

3.2. Case $r > 1$: Approximation of problem (6)

In this case, denote by c_j^2 the variables $c_j^2 = (\mathbf{d}_j^\top \mathbf{x})^2, j = 1,\ldots,r$, with under $\mathcal{H}_1$: $c_j^2 \sim \chi^2_{1,\lambda=(\mathbf{d}_j^\top \mathbf{s}_\ell)^2}$. The GLR (5) leads now to

$$\begin{aligned} \mathrm{P_{FA}}(\mathbf{D}) &= \mathbb{P}(\max_{j=1,\ldots,r} c_j^2 > \xi^2 | \mathcal{H}_0, \mathbf{D}), \\ \mathrm{P_{Det}}(\mathbf{s}_\ell,\mathbf{D}) &= \mathbb{P}(\max_{j=1,\ldots,r} c_j^2 > \xi^2 | \mathcal{H}_1, \mathbf{s}_\ell, \mathbf{D}). \end{aligned} \tag{8}$$

It can however be checked that, under $\mathcal{H}_1$,

$$\mathrm{cov}(c_j^2, c_k^2) = 4(\mathbf{d}_j^\top \mathbf{d}_k)(\mathbf{d}_j^\top \mathbf{s}_\ell)(\mathbf{d}_k^\top \mathbf{s}_\ell) + 2(\mathbf{d}_j^\top \mathbf{d}_k)^2, \tag{9}$$

so in contrast to the case $r = 1$, the $\mathrm{P_{Det}}$ (and also the $\mathrm{P_{FA}}$: use $\mathbf{s}_\ell = \mathbf{0}$ in (9)) involves the distributions of the maximum of correlated variables. Consequently, finding the exact solution to the minimax problem (6) is much more difficult than for the case where $r = 1$.

To overcome this difficulty, we propose to use bounds for the probabilities in (8). It can be shown that

$$\mathrm{P_{FA}}(\mathbf{D}) \le 1 - \Phi_{\chi_1^2}^r(\xi^2), \tag{10}$$

$$\mathrm{P_{Det}}(\mathbf{s}_{i^*},\mathbf{D}) \ge Q_{\frac{1}{2}}\left(\max_{j=1,\ldots,r} |\mathbf{d}_j^\top \mathbf{s}_{i^*}|, \xi\right) \ge Q_{\frac{1}{2}}\left(\rho^{(r)}(\mathbf{D}), \xi\right), \tag{11}$$

where $\mathbf{s}_{i^*} = \arg\min_{\mathbf{s}_i} \max_{j=1,\ldots,r} |\mathbf{d}_j^\top \mathbf{s}_i|$ denotes (one of) the alternative in $\mathbf{S}$ that is the most poorly represented by the dictionary $\mathbf{D}$ and

$$\rho^{(r)}(\mathbf{D}) = \min_{i=1,\ldots,L} \max_{j=1,\ldots,r} |\mathbf{d}_j{}^\top \mathbf{s}_i| \tag{12}$$

is the "minimax correlation" of $\mathbf{D}$ with $\mathbf{S}$. These bounds indicate that $\mathrm{P_{FA}}(\mathbf{D})$ can be controlled by ξ. To maximize $\mathrm{P_{Det}}(\mathbf{s}_{i^*},\mathbf{D})$, we shall maximize the rightmost term in (11). We thus use $\rho^{(r)}(\mathbf{D})$ as a proxy to the optimization problem (6). The learning algorithm (see Sec.4.2) should produce a value $\rho^{(r)}$ that increases rapidly with r (see Fig.5(a) for an example with $r = 21$ atoms). Note that if $r = 1$, $\rho^{(1)} = \min_i |\mathbf{d}^\top \mathbf{s}_i|$, so maximizing $\rho^{(1)}$ over $\mathbf{d}$ is equivalent to (7).

4. MINIMAX LEARNING

Dictionary learning techniques most often iterate two main steps: the sparse coding step where we fix the dictionary and calculate the unknown representation matrix, say $\mathbf{Y}$, and the dictionary update step. In the literature, sparse coding can be divided into three categories: greedy approaches (*e.g.* OMP [19], SP [20], IHT [21]), global approaches (*e.g.* ℓ_1 [22,23]) and Bayesian approaches (*e.g.* RVM [24], BCS [25]). For dictionary update, there exist also many strategies (MOD [26], K-SVD [10], etc.). This stage relies most often on a Mean Square Error criterion. As we will see, this criterion tends to represent well an average behavior of the atoms within an identified class and may consequently not be appropriate for minimax objectives. We thus propose below a modification of the dictionary update stage instantiated in the K-SVD algorithm (Sec.4.1) to suit the minimax strategy. In Sec.4.2, we describe another minimax learning algorithm based on the analysis of Sec.3.2.

4.1. K-minimax: K-SVD and minimax

For a library $\mathbf{S}$, K-SVD optimizes the dictionary by finding an approximate solution of

$$\min_{\mathbf{D},\mathbf{Y}} \|\mathbf{S}-\mathbf{D}\mathbf{Y}\|_F^2 \text{ subject to } \forall i, \|\mathbf{y}_i\|_0 \leq T.$$

We set $T=1$, in agreement with the unit ℓ_0 pseudonorm considered in the test (4)-(5). This encourages each $\mathbf{s}_i$ to be well represented by at least one column of the learned dictionary.

In the K-SVD algorithm, the dictionary update exploits the Eckart-Young theorem: for $T=1$, the rank-one approximation (in Frobenius norm) of the vectors of a class as a product of a representation column vector times a row vector of weights is obtained through the SVD. This criterion may not be relevant in a framework where minimax (instead of Euclidean) distance matters. We thus propose to modify the SVD dictionary update for each class by the minimax approach of (7), where the index i covers the atoms of the considered class. We do not provide the pseudo-code of the resulting algorithm because of space constraint, but we illustrate it in Fig.1:
(i) Start with an initial dictionary of r atoms (green dots; the black dots represent the $\mathbf{s}_i$).
(ii) Sparse coding stage: divide the set $\mathbf{s}_i$ into r clusters $\mathbf{C}_j^{\text{K}*}$, $j=1,\ldots,r$, with nearest neighbor rule (correlation criterion).
(iii) Minimax dictionary update : $\mathbf{d}^*$ is computed for each class $\mathbf{C}_j^{\text{K}*}$, by (7), which we note as $\mathbf{d}_j^{\text{K}*}$. The stages (ii) and (iii) are repeated until convergence or a stopping rule, as in K-SVD. The algorithm can be initialized with r samples randomly chosen from the library $\mathbf{S}$ or, more efficiently, by first computing the global ($r=1$) minimax atom $\mathbf{d}^*$, and then selecting the $r-1$ atoms that are less correlated to $\mathbf{d}^*$ to better sample marginal alternatives. We use the latter in Sec.5.2. The final dictionary is noted $\mathbf{D}_r^{\text{K}*}$.

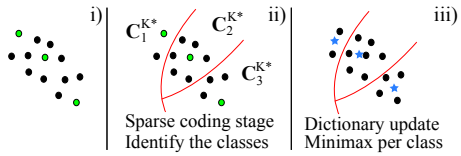


Fig. 1: Illustration of the "K-minimax" algorithm for $r=3$.

4.2. A greedy minimax algorithm

We propose here a heuristic optimization based on the analytical approximation of Sec. 3.2. This is illustrated in Fig.2 for $r=3$:
(i) First, compute through (7) the global minimax atom $\mathbf{d}^*$ representing the whole set of alternatives in $\mathbf{S}$ (blue star).
(ii) Identify the alternative $\mathbf{s}_{i^*}$ that is the most poorly represented by $\mathbf{D}$ (*i.e.*, of minimum correlation) (white dot). Select one at random if they are multiple. The expected result of this step is to obtain subspaces that are well separated, thus producing learned atoms that are discriminative and sample well the diversity of the alternatives. The set $\mathbf{s}_i$ are then classified into $j=2$ classes or clusters $(\mathbf{C}_1^*, \mathbf{C}_2^*)$ by nearest neighbor rule, and one atom $\mathbf{d}^*$ is generated through (7) for each cluster, representing the updated learned dictionary columns (blue stars).
(iii) A new class is opened using the farthest alternative to the current columns. Nearest neighbor rule results in three new classes whose minimax centers constitute the final dictionary $\mathbf{D}_3^*$. The pseudo-code of this Algorithm is given below.

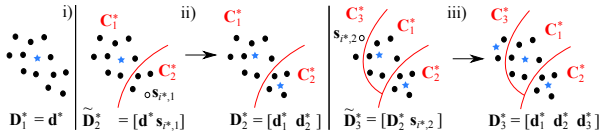


Fig. 2: Illustration of a greedy minimax algorithm, $r=3$.

Algorithm 1 Greedy minimax dictionary learning for $r>1$

Inputs: $\mathbf{S}\in\mathbb{R}^{N\times L}=[\mathbf{s}_1,\ldots,\mathbf{s}_L]$, r.

Initialization: $j=1$, $\mathbf{D}_j^*=\mathbf{d}^*\in\mathbb{R}^N$ as obtained in (7).
Set: $j=2$,
$\mathbf{s}_{i^*,j-1}=\arg\min_{\mathbf{s}_1,\ldots,\mathbf{s}_L}|\mathbf{d}^{*\top}\mathbf{s}_i|$,
$\widetilde{\mathbf{D}}_j^*=[\mathbf{d}^*\ \mathbf{s}_{i^*,j-1}]$,

while $j\leq r$ **do**
- *Classification stage*:
Each signal $\mathbf{s}_i$ is assigned to the class of atom $\widetilde{\mathbf{d}}_l^*$ of $\widetilde{\mathbf{D}}_j^*$ if $|\mathbf{s}_i^\top\widetilde{\mathbf{d}}_j^*|>|\mathbf{s}_i^\top\widetilde{\mathbf{d}}_k^*|, \forall k\neq j$. This yields j clusters $\mathbf{C}_{l=1,\ldots,j}^*$.
- *Dictionary update stage*:
for $l=1:j$
$\mathbf{d}_l^*=\arg\max_{\mathbf{d}:\|\mathbf{d}\|_2=1}\min_{\mathbf{s}_i\in\mathbf{C}_l}|\mathbf{d}^\top\mathbf{s}_i|$.
end
$\mathbf{D}_j^*=[\mathbf{d}_1^*,\ldots,\mathbf{d}_j^*]$,
$\mathbf{s}_{i^*,j}=\arg\min_i\|\mathbf{D}_j^{*\top}\mathbf{s}_i\|_\infty$,
$\widetilde{\mathbf{D}}_{j+1}^*=[\mathbf{D}_j^*\ \mathbf{s}_{i^*,j}]$,
$j=j+1$.
end while

Output: $\mathbf{D}_r^*=\mathbf{D}_j^*$.

5. NUMERICAL RESULTS

5.1. Minimax learning of faces

To illustrate the behavior of the considered algorithms, we display some results in the case where the library $\mathbf{S}$ is a database of faces. $L=40$ front-facing subjects (Fig.3(a)) were selected from the ORL Database of Faces by AT&T Laboratories Cambridge [27], representing the set of possible alternatives under $\mathcal{H}_1$.

Figs. 3(b) and 3(c) show respectively the learned 1-dimensional atoms for the minimax approach with $r=1$ ($\mathbf{d}^*$) and the classical K-SVD algorithm ($\mathbf{d}^{\text{K-SVD}}$). These images show that while $\mathbf{d}^*$ captures the marginal features (glasses, different eyes, noses, and mouths positions), the $\mathbf{d}^{\text{K-SVD}}$ atom tends to represent an "average face" with shared and smoothed characteristics. Fig.3(d) shows the results for the greedy minimax approach in the case $r=3$. Allowing for more atoms yields a dictionary whose axes dissociate to focus on specific "outliers". In particular, the second atom depicts a woman

face whose features are very different (in a correlation sense) from the others. The first atom is similar to $\mathbf{d}^*$ but with some features removed, features that are trapped by the third atom.

(a) $\mathbf{S}$: database of 40 faces (possible alternatives under $\mathcal{H}_1$)

(b) $\mathbf{d}^*$ (c) $\mathbf{d}^{\text{K-SVD}}$ (d) $\mathbf{D}_3^* = [\mathbf{d}_1^* \, \mathbf{d}_2^* \, \mathbf{d}_3^*]$

Fig. 3: (a) Database of 40 faces. (b) Minimax, $r = 1$; (c) K-SVD, $r = 1$; (d) Greedy-minimax, $r = 3$. K-SVD represents average features while minimax algorithms capture marginal features.

5.2. Minimax detection of spectral profiles

We report numerical experiments in the framework of the hyperspectral data of MUSE, which is a spectrograph built to observe very distant galaxies and will deliver data cubes of 300×300 pixels at 3600 visible wavelength channels. A library of spectral profiles $\mathbf{S} \in \mathbb{R}^{3600\times 9745}$ is available from astrophysical simulations. To compute ROC curves and Area Under the Curves (AUC) for all the tested alternatives, we limit for the present study the library dimension to $\mathbf{S} \in \mathbb{R}^{100\times 100}$ (from 9745 spectral lines, we extracted the first 100 examples, and each spectral line was restricted to an interval of about $N = 100$ contiguous wavelength channels centered around the line's maximum). The minimax detection performances are evaluated through the AUC of ROC curves computed for each alternative $\mathbf{s}_i$ activated one by one ($i = 1, \ldots, 100$) (*cf.* Fig.4).

The performances of five learned dictionaries are evaluated: two of them generated by the K-SVD algorithm for $r = 1$ and $r = 21$ ($\mathbf{d}^{\text{K-SVD}}$ and $\mathbf{D}_{21}^{\text{K-SVD}}$), another two by the greedy minimax approach for the same pair of r values ($\mathbf{d}^*$ and $\mathbf{D}_{21}^*$), and the last dictionary by the K-minimax of Sec.4.1 for $r = 21$ ($\mathbf{D}_{21}^{\text{K}*}$). Note that the case where all alternatives $\mathbf{S} = \mathbf{S}_{100}$ are tested is shown here for comparison purposes (in Table 1 and Fig.4: red-dashes, close to the orange and blue curves). In practice, computing average and minimax powers over all alternatives may not be possible for large L.

If we first consider AUC results averaged over all alternatives (*i.e.*, the usual criterion, but indeed not the one under focus), the second column of Table 1 and Fig.4 show that the best performances are obtained by the classical K-SVD: the first is $\mathbf{d}^{\text{K-SVD}}$ corresponding to $r = 1$ (pink, dash-dot), which is nearly as good as the reference (*i.e.*, the "Oracle" Neyman-Pearson (NP), to which the index of the active alternative, but not its amplitude, is known, black dots), and the second is $\mathbf{D}_{21}^{\text{K-SVD}}$ corresponding to $r = 21$ (green solid). Note, from Fig.4, that K-SVD represents well most alternatives, but not all of them (*e.g.* $i = 10, 60, 90$). Using K-SVD learned subspaces with sparsity-constrained GLR testing results in a relatively lower detection power for the corresponding spectral lines. We also see that adding more atoms to the K-SVD dictionary decreases the overall performances (compare $\mathbf{d}^{\text{K-SVD}}$ with $\mathbf{D}_{21}^{\text{K-SVD}}$) because this tends to increase more significantly the false alarm than the detection rate.

In contrast, the proposed dictionaries $\mathbf{d}^*, \mathbf{D}_{21}^*$ and $\mathbf{D}_{21}^{\text{K}*}$ perform better than the K-SVDs in term of minimax performances. The overall performances is more stable (*e.g.* limited loss at $i = 60, 90$). The learned spectral profiles for $\mathbf{D}_{21}^{\text{K-SVD}}$ and $\mathbf{D}_{21}^*$ are displayed respectively in Fig.5(b) and 5(c), showing again more diversity in the minimax learning compared to classical K-SVD.

Comparing now the proposed optimization approaches for $r > 1$ (greedy-minimax $\mathbf{D}_{21}^*$: blue, circles; K-minimax $\mathbf{D}_{21}^{\text{K}*}$: orange, crosses) to the minimax dictionary for $r = 1$ ($\mathbf{d}^*$: cyan, diamonds), we see that the minimax (and also the average) performances are improved w.r.t. $r = 1$, which was the main objective of this study.

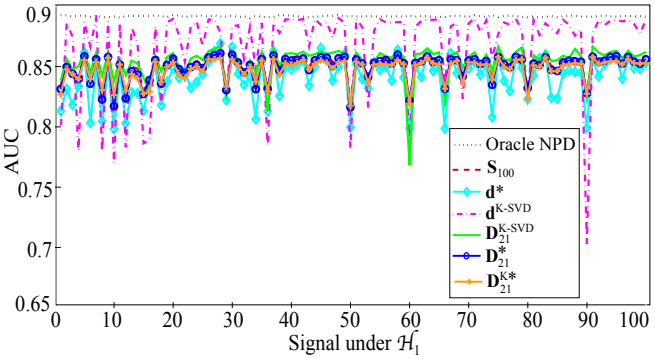
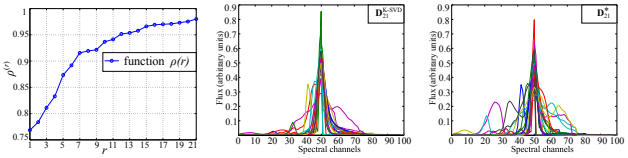


Fig. 4: AUC for 100 instances under $\mathcal{H}_1$.



(a) Function $\rho^{(r)}$ (b) $\mathbf{D}_{21}^{\text{K-SVD}}$ (c) $\mathbf{D}_{21}^*$

Fig. 5: (a): Evolution of $\rho^{(r)}$ (*cf.* (12)). (b) and (c): learned atoms.

Dictionary	Ranking Criterion	
	Min AUC (minimax)	Average AUC
Atom under $\mathcal{H}_1$	Ref : 0.886	Ref : 0.887
$\mathbf{S}_{100}$	1^{st} : 0.813	3^{rd} : 0.847
$\mathbf{d}^*$	2^{nd} : 0.794	6^{th} : 0.836
$\mathbf{d}^{\text{K-SVD}}$	4^{th} : 0.699	1^{st} : 0.863
$\mathbf{D}_{21}^{\text{K-SVD}}$	3^{rd} : 0.764	2^{nd} : 0.849
$\mathbf{D}_{21}^*$	1^{st} : 0.812	4^{th} : 0.845
$\mathbf{D}_{21}^{\text{K}*}$	1^{st} : 0.813	5^{th} : 0.843

Table 1: Results over 100 alternatives (uncertainty: ± 0.001).

6. CONCLUSIONS

The proposed algorithms were shown to improve the minimax (and as a side-effect, the average) detection performances w.r.t. to the $r = 1$ case. We caution, however, that dictionaries learned from large libraries (*e.g.* with $L \propto 10^4$ or more), while keeping r still in the tens, may not yield much improvement w.r.t. $\mathbf{d}^*$, as too low values of r (w.r.t. to L) may not be enough for the dictionary to sample well the diversity of such large amounts of alternatives. This poses the question, left unanswered here, of devising approaches to find the best (w.r.t. minimax criteria) value of r. On this question, we note that this value generally depends on the SNR. We observed in particular that lower values of r may be preferable at lower SNR regimes: when the noise level is high, testing additional target subspaces may produce higher false alarm rates with only marginal improvement of the detection rate. The interplay between the intrinsic separation of the main subpopulations of the alternatives w.r.t. to the scatter caused by noise in the data suggests a library-dependent compromise to be found in the number of target subspaces that are learned and tested.

7. REFERENCES

[1] J. Von Neumann, "Zur Theorie der Gesellschaftsspiele: On the theory of parlor games," *Mathematische Annalenn*, vol. 100, pp. 295–300, 1928.

[2] S. Kim and S. Boyd, "A minimax theorem with applications to machine learning, signal processing and finance," *SIAM Journal on Optimization*, vol. 19, no. 3, pp. 1344–1367, 2008.

[3] S.A. Kassam and H.V. Poor, "Robust techniques for signal processing: A survey," in *Proc. of IEEE*. IEEE, 1985, vol. 73, pp. 433–481.

[4] E.L. Lehmann and J.P. Romano, *Testing statistical Hypotheses*, Springer, 2005.

[5] D.L. Donoho and I.M. Johnstone, "Minimax estimation via wavelet shrinkage," Annals of Statistics, 1998.

[6] B. Sklar, *Digital Communications: Fundamentals & Applications*, Prentice Hall, 2011.

[7] J.L. Whitwell, K.A. Josephs, M.N Rossor, J.M. Stevens, T. Revesz, J.L Holton, S. Al-Sarraj, A.K. Godbolt, N.C. Fox, and J.D. Warren, "Magnetic resonance imaging signatures of tissue pathology in frontotemporal dementia," *Archives of Neurology*, vol. 62, no. 9, pp. 1402–1408, 2005.

[8] R.F.R. Suleiman, D. Mary, and A. Ferrari, "Minimax sparse detection based on one-class classifiers," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 5553–5557.

[9] D. Manolakis, D. Marden, and G.A. Shaw, "Hyperspectral image processing for automatic target detection applications," *Lincoln Lab. J.*, vol. 14, no. 1, pp. 79–116, 2003.

[10] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54(11), pp. 4311–4322, November 2006.

[11] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image Processing,*, vol. 15, no. 12, pp. 3736–3745, 2006.

[12] V. Abolghasemi, S. Ferdowsi, and S. Sanei, "Blind separation of image sources via adaptive dictionary learning," *IEEE Transactions on Image Processing*, vol. 21, no. 6, pp. 2921–2930, 2012.

[13] Q. Zhang and B. Li, "Discriminative K-SVD for dictionary learning in face recognition," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 2010, pp. 2691–2698.

[14] S. Kong and D. Wang, "A dictionary learning approach for classification: Separating the particularity and the commonality," in *Computer Vision - ECCV 2012*, vol. 7572 of *Lecture Notes in Computer Science*, pp. 186–199. 2012.

[15] J.B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*. 1967, vol. 1, pp. 281–297, University of California Press.

[16] E. Arias-Castro, E.J. Candès, and Y. Plan, "Global testing under sparse alternatives: Anova, multiple comparisons and the higher criticism," *Annals of Statistics*, vol. 39(5), pp. 2533–2556, 2010.

[17] S. Paris, R.F.R. Suleiman, D. Mary, and A. Ferrari, "Constrained likelihood ratios for detecting sparse signals in highly noisy 3D data," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 3947–3951.

[18] A.H. Nutall, "Some integrals involving the (Q sub M) - function," Tech. Rep. AD-779846, Naval Underwater Systems Center New London, Connecticut, May 1974.

[19] S. Mallat, G. Davis, and Z. Zhang, "Adaptive time-frequency decompositions," *SPIE Journal of Optical Engineering*, vol. 33, pp. 2183 – 2191, 1994.

[20] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *IEEE Transactions on Information Theory*, vol. 55, no. 5, pp. 2230–2249, 2009.

[21] T. Blumensath, M. Yaghoobi, and M.E. Davies, "Iterative hard thresholding and ℓ_0 regularisation," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2007, vol. 3, pp. III–877–III–880.

[22] S. Chen, D.L. Donoho, and M.A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Review*, vol. 43, no. 1, pp. 129 – 159, 2001.

[23] D.L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization," in *Proceedings of the National Academy of Sciences*, March 2003, vol. 100(5), pp. 2197–2202.

[24] M.E. Tipping, "The relevance vector machine," Advances in Neural Information Processing Systems (NIPS), 2000.

[25] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Transactions on Signal Processing*, vol. 56, no. 6, pp. 2346–2356, 2008.

[26] K. Engan, S.O. Aase, and J.H. Hakon-Husoy, "Method of optimal directions for frame design," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1999, vol. 5, pp. 2443–2446.

[27] F.S. Samaria and A.C. Harter, "Parameterisation of a stochastic model for human face identification," in *Applications of Computer Vision, 1994., Proceedings of the Second IEEE Workshop on*, 1994, pp. 138–142.