Detection tests for worst-case scenarios with optimized dictionaries. Applications to hyperspectral data

Raja Fazliza Raja Suleiman

UNIVERSITÉ DE NICE SOPHIA ANTIPOLIS - UFR Sciences

ÉCOLE DOCTORALE EN SCIENCES FONDAMENTALES ET APPLIQUÉES

Ph.D Dissertation

to obtain the title of

Docteur en Sciences

of Université de Nice Sophia Antipolis (UNS), France.
Specialization : Science of the Universe

Presented by

Raja Fazliza RAJA SULEIMAN

Detection tests for worst-case scenarios with optimized dictionaries. Applications to hyperspectral data.

Méthodes de détection robustes avec apprentissage de dictionnaires. Applications à des données hyperspectrales.

Defended on 16 December, 2014

Jury :

Reviewers:	Christophe Collet	Professor	-	Université de Strasbourg
	David Brie	Professor	-	Université de Lorraine
Examinator:	Pierre Chainais	MCF HDR	-	Ecole Centrale de Lille
	Thomas Rodet	Professor	-	ENS Cachan
Advisors:	David Mary	Professor	-	Université de Nice Sophia Antipolis
	André Ferrari	Professor	-	Université de Nice Sophia Antipolis

Detection tests for worst-case scenarios with optimized dictionaries. Applications to hyperspectral data.

Abstract: This Ph.D dissertation deals with a "one among many" detection problem, where one has to discriminate between pure noise under $\mathcal{H}_0$ and one among L alternatives that are known up to an amplitude factor under $\mathcal{H}_1$. This work focuses on the study and implementation of robust reduced dimension detection tests using optimized dictionaries. The proposed approaches are principally assessed on hyperspectral data.

In the first part of this dissertation, several technical topics associated to the framework of this dissertation are presented. These topics are hypothesis testing, minimax strategy, detection in Hyperspectral Imaging, dimension reduction techniques and sparse representations.

The second part of this work highlights the theoretical and algorithmic aspects of the proposed methods. Two issues linked to the large number of alternatives arise in this framework. First, the computational complexity associated to the Generalized Likelihood Ratio (GLR) test with the constraint of sparsity-one inflates linearly with L, which can be an obstacle when multiple data sets have to be tested. Second, standard procedures based on dictionary learning aimed at reducing the dimensionality may suffer from severe power losses for some alternatives, thus suggesting a worst-case scenario strategy.

In the case where the learned dictionary has $K = 1$ column, we show that the exact solution of the resulting detection problem, which can be formulated as a minimax problem, can be obtained by Quadratic Programming.

Furthermore, the case $K > 1$ allows a better sampling of the intrinsic diversity of the alternatives, but it is much more complex to implement. The worst-case analysis of this case, which is more involved, leads us to propose three minimax learning algorithms.

Finally, the third part of this manuscript presents applications of the proposed reduced dimension detection tests using optimized dictionaries. The principal application regards astrophysical hyperspectral data of the Multi Unit Spectroscopic Explorer (MUSE) instrument. Numerical results show that the proposed algorithms are indeed robust and in the case $K > 1$ they allow to increase performances over the $K = 1$ case. The resulting performances are in fact comparable to the GLR using the full set of alternatives, while being computationally simpler. Other possible applications are also taken into account such as learning minimax faces and worst-case recognition of handwritten digits.

Keywords: Statistical Signal Processing, Detection, Minimax, Dimensionality Reduction, Dictionary Learning, Sparsity, Hyperspectral Imaging.

Méthodes de détection robustes avec apprentissage de dictionnaires. Applications à des données hyperspectrales.

Résumé : Le travail dans cette thèse porte sur le problème de détection «one among many» où l'on doit distinguer entre un bruit sous $\mathcal{H}_0$ et une parmi L alternatives connues à un facteur d'amplitude près sous $\mathcal{H}_1$. Ce travail se concentre sur l'étude et la mise en œuvre de méthodes de détection robustes de dimension réduite utilisant des dictionnaires optimisés. Les approches proposées sont principalement évaluées sur des données hyperspectrales.

Dans la première partie de cette thèse, plusieurs sujets techniques associés à cette thèse sont présentés: les tests d'hypothèse, la stratégie minimax, la détection en imagerie hyperspectrale, les techniques de réduction de dimension et les représentations parcimonieuses.

La deuxième partie de ce travail met en évidence les aspects théoriques et algorithmiques des méthodes proposées. Deux inconvénients liés à un grand nombre d'alternatives L se posent. Tout d'abord, la complexité de calcul associée au test de Rapport de Vraisemblance Généralisé (GLR) sous contrainte de 1-parcimonie augmente linéairement avec L, ce qui constitue une limitation dans le cas où le test doit être répété sur plusieurs jeux de données. Ensuite, les approches standards basées sur l'apprentissage d'un dictionnaire visant à réduire la dimensionnalité peuvent conduire à une perte de puissance élevée pour certaines alternatives.

Nous proposons dans ce cadre des techniques d'apprentissage de dictionnaire basées sur un critère robuste de type minimax. Dans le cas où l'on cherche un dictionnaire à $K = 1$ atome, nous montrons que la solution exacte peut être obtenue par Programmation Quadratique.

Par ailleurs, nous montrons que le cas $K > 1$ permet un meilleur échantillonnage de la diversité intrinsèque des alternatives, mais que la résolution exacte du problème minimax dans ce cas est aussi beaucoup plus difficile à obtenir. Nous proposons dans ce cas trois algorithmes d'apprentissage minimax qui permettent d'approcher cette solution.

Finalement, la troisième partie de ce manuscrit présente plusieurs applications. L'application principale concerne les données astrophysiques hyperspectrales de l'instrument *Multi Unit Spectroscopic Explorer* (MUSE). Les résultats numériques montrent que les méthodes proposées sont robustes et que le cas $K > 1$ permet d'augmenter les performances de détection minimax par rapport au cas $K = 1$. Les méthodes proposées sont comparables au test GLR utilisant la bibliothèque complète, tout en réduisant considérablement le coût de calcul. D'autres applications possibles sont également prises en compte telles que l'apprentissage minimax de visages et la reconnaissance de chiffres manuscrits dans le pire cas.

Mots clés : Traitement Statistique du Signal, Détection, Robustesse, Réduction de Dimension, Apprentissage de Dictionnaire, Parcimonie, Imagerie Hyperspectrale.

Acknowledgements

Growing up in a developing country located in Southeast Asia, my childhood dream was to see the other part of the world: especially Europe. This dream was realized when I was awarded a scholarship from the Malaysian government after graduating high school to pursue higher education abroad. In 2004, a small group of young and ambitious students arrived in south of France: Nice (including me, of course). After 5 years of blood, sweat and tears, my friends and I graduated with a Master's degree in 2009 and we went back to our hometown. During the study in Nice, we have met many great lecturers including Mr. David Mary and Mr. André Ferrari. Little that I know at that time that they will be my future Ph.D advisors.

In 2011, my husband and I decided to pursue our studies abroad. We managed to obtain a scholarship from a Malaysian government agency, MARA (Majlis Amanah Rakyat). Lucky for me, Mr. André Ferrari accepted to be my advisor, along with Mr. David Mary.

During the course of my Ph.D studies, I worked closely with Mr. David Mary. I thank him for every piece of advices and discussions related to this research work, not to forget his particular attention in the importance of explaining well our works for publications. I would like to thank Mr. André Ferrari for his supervision, especially for his expertise in the theory of statistical signal processing. I feel blessed to have them as my advisors, for their continuous supports, and for the opportunities to participate in both national and international conferences. I have gained ample knowledge in my research area thanks to them.

I would like to thank the members of my dissertation committee: Prof. Brie, Prof. Collet, Prof. Rodet and Dr. Chainais for their time and insightful comments.

My sincere thanks goes to Mr. Roland Bacon, the principal investigator of the MUSE instrument and the MUSE consortium for providing spectral data. To Antony Schutz, thanks for providing the code to visualize the 3D atoms. The administration team of Laboratoire J.-L. Lagrange has been a great help in preparing paperworks, claims, etc. I thank Caroline, Delphine and Jocelyne for this.

A special thanks to my ex-labmates: Silvia, Jie Chen and Nguyen for sharing their knowledge and experiences. To Gao Wei, Rita and Roula, thanks for the fun, encouragements and your sincere friendship. Rita and Roula are the best listener to all my stories: work related or not. I would like to thank Norazah, a personal friend for hanging out with me during stressful time.

Finally, my most heartfelt thanks to my dear husband for his patience and emotional support, my parents, my late grandparents, my parents-in-law, and my siblings for their constant care and presence.

Contents

List of Figures

List of Tables

Notations and Definitions

Notations

a	Scalar
$\boldsymbol{a}$	Vector
$\boldsymbol{A}$	Matrix
$\boldsymbol{A}^\top$	Transpose of a matrix A
$\boldsymbol{A}_K^{\mathrm{M}}$	Matrix A containing K columns learned from method M
$\mathbf{0}$	Vector of zeros
$\mathbf{I}$	Identity Matrix
$\boldsymbol{a} \sim N$	$\boldsymbol{a}$ follows distribution N
$\boldsymbol{a} \gtrless \boldsymbol{b}$	$\boldsymbol{a}$ is greater or less than $\boldsymbol{b}$
$\lvert\boldsymbol{a}\rvert$	Absolute value of $\boldsymbol{a}$
$\lVert\boldsymbol{a}\rVert_0$	ℓ_0 pseudo-norm:= $\#\{n : a_n \neq 0\}$: total number of non-zero elements in vector $\boldsymbol{a}$
$\lVert\boldsymbol{a}\rVert_p$	ℓ_p norm:= $\left(\sum_{n=1}^{N} \lvert a_n\rvert^p\right)^{1/p}$
$\lVert\boldsymbol{a}\rVert_\infty$	ℓ_∞ norm:= $\max_n \lvert a_n\rvert$
$\lVert\boldsymbol{a}-\boldsymbol{b}\rVert_2$	Euclidean distance $(\boldsymbol{a}, \boldsymbol{b})$: $= \sqrt{\sum_{n=1}^{N}(a_n - b_n)^2} = \sqrt{(\boldsymbol{a}-\boldsymbol{b})^\top(\boldsymbol{a}-\boldsymbol{b})}$
$\overline{\boldsymbol{a}}$	Average of $\boldsymbol{a}$
$\mathcal{H}_0, \mathcal{H}_1$	Hypothesis 0 (null), Hypothesis 1 (alternative)
$\mathcal{N}(\mu, \sigma^2)$	Normal distribution of mean μ and variance σ^2
$p(\boldsymbol{a})$	Probability density function of $\boldsymbol{a}$
$p(\boldsymbol{a}, \boldsymbol{\theta})$	Joint probability density function of $\boldsymbol{a}$, $\boldsymbol{\theta}$
$p(\boldsymbol{a}; \boldsymbol{\theta})$	Probability density function of $\boldsymbol{a}$ with $\boldsymbol{\theta}$ as parameter
$p(\boldsymbol{a}; \mathcal{H}_i)$	Probability density function of $\boldsymbol{a}$ when $\mathcal{H}_i$ is true
$p(\boldsymbol{a}\lvert\boldsymbol{\theta})$	Conditional probability density function of $\boldsymbol{a}$ conditioned on $\boldsymbol{\theta}$
$p(\boldsymbol{a}\lvert\mathcal{H}_i)$	Conditional probability density function of $\boldsymbol{a}$ conditioned on $\mathcal{H}_i$ being true
$\mathbb{P}$	Probability
$\mathbb{R}^N$	N-dimensional Euclidean space
Φ	Cumulative distribution function

Abbreviations

i.i.d.	independent and identically distributed
inf	infimum: greatest lower bound
$\mathrm{P_{CR}}$	Probability of correct rejection

P_{Det}	Probability of detection
P_E	Probability of error
P_{FA}	Probability of false alarm
P_M	Probability of miss
sup	supremum: least upper bound
vs.	versus
w.r.t.	with respect to

Acronyms

1D	One-dimensional
3D	Three-dimensional
ASD	Adaptive Subspace Detectors
AVIRIS	Airborne Visible/Infrared Imaging Spectrometer
AUC	Area Under Curve
CDF	Cumulative Distribution Function
CFAR	Constant False Alarm Rate
ESO	European Southern Observatory
GAP	Greedy Analysis Pursuit
GLR	Generalized Likelihood Ratio
HSI	Hyperspectral Imaging
IFU	Integral Field Unit
JPL	Jet Propulsion Laboratory
KKT	Karush-Kuhn-Tucker
LASSO	Least Absolute Shrinkage and Selection Operator
LR	Likelihood Ratio
ML	Maximum Likelihood
MOD	Method of Optimal Directions
MP	Matching Pursuit
MSD	Matched Subspace Detector
MSE	Mean Square Error
MUSE	Multi Unit Spectroscopic Explorer
NN	Nearest Neighbor
NP	Neyman-Pearson
NPD	Neyman-Pearson Detector
OMP	Orthogonal Matching Pursuit
ONPD	Oracle Neyman-Pearson Detector
PCA	Principal Component Analysis

PDF	Probability Density Function
PSF	Point Spread Function
QP	Quadratic Programming
RD	Reduced Dimension
ROC	Receiver Operating Characteristic
RX	Reed-Xiaoli
SDT	Signal Detection Theory
SKM	Spherical K-Means
SMF	Spectral Matched Filter
SNR	Signal to Noise Ratio
SVD	Singular Value Decomposition
SVDD	Support Vector Data Description
SVM	Support Vector Machine
VQ	Vector Quantization

General Overview

Introduction

Imagine that a chef is cooking *fugu* (i.e., a possibly highly toxic japanese puffer fish) for 100 guests, and that to avoid a long wait, the chef has to respect a time constraint: each fugu has to be prepared within ten minutes time frame. The chef has to make sure that each fugu is well-prepared. If one of the fugu is not prepared well, the guest who eats it will die. If the chef prepares the fishes fairly well, they are non-poisonous and taste quite good, but not great. If the preparation is really well, then they taste very delicious.

In such a situation, it is likely that the chef will follow a *minimax* strategy, that is, he will optimize the preparation of the fish with the aim that no fish is badly prepared, so that none of the guests will be poisoned to death (i.e., the chef will minimize the maximum risk). If he applies an "average performance strategy", this could lead to a large fraction of perfectly prepared fugus, but risking a small number of fatal fugus. In this setting, clearly the chef cannot follow an average performance strategy, although on average, the fugus would presumably be better prepared.

In connection to this illustrative example, this dissertation is about detection tests which are designed to minimize some maximum loss instead of focusing on good average detection. Signal Detection Theory (SDT) assesses mathematically the ability to distinguish between correct and wrong decisions when facing uncertainty. From the point of view of SDT, signal is an informative data plus noise, and noise is a random unwanted disturbance. Various fields apply detection theory to analyze different kinds of decision problems. For instance, in astrophysics, some scientists are interested in the study of very distant galaxies (known as Lyman-α emitters) characterized by very faint spectral lines. Once the data (typically high dimension) is acquired by a dedicated instrument, the first task is to detect the presence of these targeted Lyman-α lines (at random unknown position) in the data. This is not an easy task, as the data generally contains high noise. Moreover the spectral lines are very weak with respect to (w.r.t.) noise and may present very different signatures. For this particular application, focusing on average detection performance may lead to leave "atypical" spectral profiles undetected. Since such atypical profiles may also be the most interesting ones, a minimax approach appears relevant. This specific application case clearly motivated the minimax detection study developed in this dissertation.

Other examples of possible applications for the proposed minimax detection strategy are worst handwritten digits recognition, detection of cancerous cell or mines detection. For the first example, minimax strategy would maintain correct recognition rate of handwritten digits in worst-case scenarios, i.e., when the digits are written "badly" (hardly recognizable). For the second example, cancer screening may involve the analysis of a tissue sample taken from a patient. This sample contains a large number of cells and the cancerous cells have specific signatures. If we miss the detection of even one cancerous cell, a wrong diagnostic would be given to the patient. In regard of the last example, mines detection is very important for civilian security. Some countries are contaminated with a large area of buried

mines, left from the wars. This case requires a *robust* detector aiming at minimizing the miss detection of all types of mines.

When dealing with a general detection problem, we are deciding whether a target is present or is absent. This can be translated as forming statistical hypothesis models where one assumes that only noise is present under the null hypothesis ($\mathcal{H}_0$), and the signal of interest along with noise is present under the alternative hypothesis ($\mathcal{H}_1$). When there are more than two uncertain events, one can also form multiple hypothesis models. The observed data under each hypothesis are characterized by probability density functions (PDFs), which are known entirely or just partially. A figure of merit in hypothesis testing is the probability of detection (also termed power of a test) and its complement is called the probability of miss. Following the minimax strategy, we will seek to minimize the maximum probability of miss for a given set of possible alternatives.

The number of possible alternatives can be arbitrary, ranging from small sets (i.e., in the tens or hundreds, in telecommunications symbols for instance) to very large sets (in the hundred of thousands or more; e.g., in genomics or for samples drawn from numerical models as we will see for hyperspectral application). In this dissertation, we focus on the most interesting setting where we have a large number of possible target signals. This case often arises when there are few known target signals, but those are registered with systematic disturbances that can be modeled and sampled, leading to a large set of alternatives. In other applications, the unknown target signal can have arbitrary variations. Numerical simulations can then provide numerous possible templates of the target signal, yielding again a large number of alternatives.

With regard to the large data set that we will consider, the detection problem dealt with in this dissertation involves *dictionary learning* techniques. A dictionary is a smaller representation of a large known data set (often called a library). The columns of this dictionary are called atoms. In dimension reduction techniques, a dictionary is learned from the reference library w.r.t. an optimization criterion. In line with the strategy mentioned above, we shall apply worst-case (minimax) criteria to build the dictionary, instead of using more common criteria such as the minimum *Mean Squared Error* (MSE).

Motivation

The main objective of this Ph.D thesis is to study techniques for detecting the presence of very faint targets in highly noisy large data set. The spectral profiles are assumed known, but the amplitude and position of the possible targets in the data, activated one at a time under $\mathcal{H}_1$, are however *unknown*. Thus, our hypothesis model is composite.

These targets exhibit some diversity in their signatures. Some of them may share similar features, which define an *average* or *typical* behavior. Some others may show very specific or unique signatures. Those represent a *dissimilar* or *atypical* behavior. As a simple illustration of this, assume our library is a a basket of mix fruits, where most of them are apples, but there is also one banana in the basket. In this case, the average, typical shape of the known fruits is roughly a sphere, while the atypical shape is a long curved-cylinder.

A classical detection test such as the Generalized Likelihood Ratio (GLR) test consists

in this context of testing all the possible positions of the target signatures in the data with amplitudes estimated by maximum likelihood. As a result, this test leads to a prohibitive computation complexity if the size of the library is large.

Generally, when having a large library, a standard approach is to operate in subspaces of reduced dimension (RD). There exists many reduction dimension methods such as the Principal Component Analysis (PCA), the low rank approximation by SVD (Singular Value Decomposition) or dictionary learning algorithms (e.g., K-SVD). An extreme example of dimension reduction is by representing the large reference library by its sample mean vector. As we will see, such an RD approach is indeed good for detecting most target signatures, because it tends to represent the average behavior of the considered data. Reducing the dimension in such a way leads to low detection power for atypical signatures because these dissimilar targets may lie quite apart from the learned subspace (e.g., in the illustration above, testing for a spherical shape when a "banana" is present under $\mathcal{H}_1$ yields a poor matching). These effects are actually common to most classical dimension reduction techniques, and will be illustrated in this dissertation.

Missing the target whilst it is present is a loss of interesting and important information, for instance in the case of the detection of spectral lines emitted by galaxies in astrophysics. Furthermore, for some applications such as mine detection, missing a single target may be highly damageable.

We therefore choose to focus here on minimax RD approach relying on the classical GLR test, where the subspaces are learned from the known library, with the aim to minimize (over all alternatives) the maximum probability of miss of the GLR (i.e., a minimax strategy), while controlling the probability of false alarm. We restrict the studies in this dissertation to the GLR test as it can provide implementable testing procedures and further benefits from important properties. It also allows to keep the complexity low. The identification of the corresponding subspace is tackled as a dictionary learning problem for worst-case detection scenarios. An important highlight in our framework is that conventional learning algorithms do not perform well w.r.t. worst-case objectives, which call for specific learning algorithms.

The optimization problem can equivalently be defined as maximizing the minimum detection power i.e., a *maximin* problem. In the rest of this dissertation, we shall generically use the term minimax.

Contributions

After a detailed study of the detection problems induced by classical detection methods when testing a very large reference library, we propose reduced dimension models using adapted (minimax) dictionaries that are learned from the reference library.

i. The first approach solves a one-dimensional (1D) subspace minimax learning problem, in the form of quadratic programming (QP). Connections between this 1D optimization problem and One-class classifiers of Support Vector Machine (SVM) type are also investigated.

ii. We find that the minimax optimization problem for general K-dimensional subspaces ($K > 1$) is intricate. As a consequence, we turn to algorithms that solve approximately the minimax optimization problem. The first algorithm which we call the Greedy minimax algorithm functions in a greedy approach. The second algorithm is a variant of the classical K-SVD technique, where the dictionary learning step is replaced by the 1D exact minimax solution (of i.), and the sparsity of the representation vectors is set to one. We name this method K-minimax. Apart from these two minimax algorithms, we also combine the 1D exact minimax solution with several clustering methods taken from literature. This is a two-step approach. In the first step, we apply a chosen clustering technique on the unit sphere to partition the known alternatives into K clusters. Then, for each cluster, the minimax atom is generated. The concatenation of the resulting K minimax atoms (each representing one cluster) forms the final dictionary of K columns. The proposed algorithms (of i. and ii.) yield detection performances as good as testing using the full library $\boldsymbol{S}$, yet with lower complexity.

iii. Applications:

- The main motivation for elaborating the proposed methods is for the detection of very faint and noisy Lyman-α spectral profiles in the MUSE (Multi Unit Spectroscopic Explorer) data cube. Comparing with K-SVD technique, we show that RD tests using the proposed approaches yield higher detection power in worst-case scenarios.
- Apart from astrophysics, we also illustrate in this dissertation minimax dictionary for learning faces and consider a possible machine learning application for worst handwritten digits recognition.

Several scientific publications and communications were presented in the framework of this dissertation. The list of related publications is:

Journal article and International conference proceedings

1. Raja Fazliza Raja Suleiman, David Mary and André Ferrari. "Dimension reduction for hypothesis testing in worst-case scenarios". IEEE Transactions on Signal Processing, vol. 62, no. 22, pp. 5973-5968, November 2014. [Suleiman *et al.* 2014a]

2. Raja Fazliza Raja Suleiman, David Mary and André Ferrari. "Subspace learning in minimax detection". IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), May 2014, pp. 3062-3066. [Suleiman *et al.* 2014b]

3. Raja Fazliza Raja Suleiman, David Mary and André Ferrari. "Minimax sparse detection based on one-class classifiers". IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), May 2013, pp. 5553-5557. [Suleiman *et al.* 2013a]

4. Silvia Paris, Raja Fazliza Raja Suleiman, David Mary and André Ferrari. "Constrained likelihood ratios for detecting sparse signals in highly noisy 3D data". IEEE

International Conference on Acoustics, Speech, and Signal Processing (ICASSP), May 2013, pp. 3947-3951. [Paris *et al.* 2013b]

Francophone conference proceeding

5. Raja Fazliza Raja Suleiman, David Mary and André Ferrari. "Parcimonie, apprentissage et classification pour une approche minimax en détection". Actes du 24e Colloque GRETSI sur le Traitement du Signal et Images, September 2013. [Suleiman *et al.* 2013b]

Communication without proceeding

6. David Mary, André Ferrari and Raja Fazliza Raja Suleiman. "Détection de sources astrophysiques dans les données du spectographe intégral de champ MUSE". Actes du 3e colloque scientifique SFPT-GH sur l'Imagerie Hyperspectrale, May 2014. [Mary *et al.* 2014]

Organization of the dissertation

This Ph.D dissertation consists of three main Parts. The first Part introduces some fundamentals in detection theory and the principles of several techniques in fields related to the framework of this dissertation (minimax criterion, spectral matching, dimensionality reduction, sparsity). Chapter 1 deals with detection problems, associated tests and with the minimax approach. Chapter 2 deals with target detection in Hyperspectral Imaging (HSI), which is the primary application of this dissertation. The final Chapter in the first Part, Chapter 3, describes basic principles in sparse modeling and RD learning techniques.

The second Part presents the main contributions of this research work. Chapter 4 investigates issues induced by classical detection methods which test exhaustively very large reference libraries. In Chapter 5, two types of detection tests for RD models using minimax dictionary learning methods are proposed. The first is a constrained GLR test, and the second test is an unconstrained GLR. Next, Chapter 6 is dedicated to algorithmic aspects based on the theoretical analysis in preceding Chapter. The proposed minimax (worst-case) learning techniques for an arbitrary size dictionary are presented in this Chapter.

The third Part depicts some applications of the proposed worst-case detection tests. Chapter 7 evaluates the proposed methods for the detection of spectral profiles in astrophysical hyperspectral data of the MUSE instrument. Chapter 8 illustrates two applications of the considered worst-case scenario from the viewpoint of machine learning: the first concerns minimax learning of faces, and the second involves the recognition of handwritten digits.

Part I

Introduction to technical topics connected to the dissertation

The first Part of this dissertation discusses basic aspects of several topics connected to the principal detection problem investigated in this dissertation. These topics will be further addressed in the following Parts II and III.

Chapter 1 conveys fundamental principles of statistical hypothesis testing. The associated probabilities characterizing a detection test are presented. It summarizes various aspects to be considered when building a detection test and assessing its performances. The Chapter outlines a comparison between frequentist and Bayesian approaches to hypothesis testing.

Furthermore, we discuss in Chapter 1 some of the earlier works related to minimax hypothesis testing. Generally, these approaches attempt to establish conditions with data size and/or the Signal to Noise Ratio allowing to distinguish the null hypothesis from the alternatives. We contrast this view with the definition of minimax (worst-case) criterion retained in this dissertation. In our work, we aim to minimize the maximum probability of miss, among a known and countable (large) set of possible alternatives, while controlling the probability of false alarm.

Hyperspectral images are the main application of our proposed approaches in detection testing. In line with this, **Chapter 2** presents general knowledge on target detection in Hyperspectral Imaging. Two categories of detectors often used in Hyperspectral Imaging, namely Spectral Matching detectors and Anomaly detectors, are introduced.

The final Chapter in the first Part of this dissertation, **Chapter 3**, is dedicated to various dimension reduction techniques found in the literature. In particular, sparse learning algorithms are highlighted as they will be closely related to the proposed minimax detection algorithms.

Chapter 1

Testing statistical hypotheses and the minimax strategy

Contents

1.1 Introduction

By means of Signal Detection Theory, one can assess statistically a decision making process when facing uncertainty. The first approach to decision making was initiated in 1908 by W. Gosset for small data sample, introducing the t-test to monitor the quality of beer for his employer (Guinness) [Student 1908]. In the 1920s - 1930s, R. Fisher [Fisher 1925], J. Neyman and E. Pearson [Neyman & Pearson 1933] concentrated their research works in this domain, constituting groundbreaking theories and methods in the probability theory and hypothesis testing. Making a decision (or conclusions) from observational or experimental data is known as *statistical inference*. These data are modeled by probability distribution functions, involving parameters which may be of interest (i.e., related to the signal one wishes to detect) or not (they are then called nuisance parameters).

The parameters of interest may be known or unknown. In the latter case, one have to estimate these parameters w.r.t. a defined criterion such as the minimum MSE or the minimax criteria.

In statistical inference, two main problems are

i. estimating the parameters of interest,

ii. *testing hypotheses* related to the probabilistic model.

The first problem seeks the "best" (w.r.t. a defined criterion) approximation of the parameters, while the second problem provides answers on the "best" match between the stated hypotheses and the data.

The most basic form of hypothesis testing consists of two exclusive hypotheses, and lead to the so-called binary hypothesis tests. Their extension is termed multiple hypothesis tests.

There are basically four steps when making a hypothesis test: stating statistical hypotheses for the problem at hand, identify (derive) the test statistic, select a significance level for the test, and draw conclusions for the data at hand.

There are two major approaches in statistical inference: the *frequentist* and *Bayesian* approaches. The frequentist approach draws conclusions by inferring on a large number of realizations, where the (possibly unknown) parameters remain fixed (i.e., deterministic). The Bayesian approach fixes the data obtained from the realization, but the parameters are not fixed and are described via probabilistic manner. This approach requires consequently the knowledge of prior probabilities for each hypothesis. As we restrict the research work in this dissertation to the GLR test (which is a frequentist method), this Chapter highlights this approach. The Bayesian approach is nevertheless also presented briefly in this Chapter.

In relation to the framework of this dissertation, we present the minimax strategy in hypothesis testing at the end of this Chapter.

1.2 Stating the detection problem into hypothesis model

In essence, a signal detection problem involves deciding whether a signal is *absent* or is *present* in a noisy data set. The first step in hypothesis testing is to cast these *exclusive* events into a pair of hypotheses. A binary hypothesis model can be written as

$$\begin{cases} \mathcal{H}_0 & : \text{The signal is absent. This is the } \textit{null} \text{ hypothesis,} \\ \mathcal{H}_1 & : \text{The signal is present. This is the } \textit{alternative} \text{ hypothesis.} \end{cases} \tag{1.1}$$

If we have more than two exclusive events to be tested, this calls for multiple hypothesis models, where there will be several alternative hypotheses $(\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_N)$. For example, multiple hypothesis testing is widely used in *Genomics*, for simultaneous detection of genes w.r.t. hundreds of markers [Manly *et al.* 2004].

The objective of hypothesis testing is to quantify the decision making process and draw conclusions w.r.t. the considered data set (whether to accept $\mathcal{H}_0$ or to reject $\mathcal{H}_0$,

i.e., accepting the alternative). Owing to the randomness inherent to any data acquisition process, each hypothesis is statistical in nature and described by PDFs, whose parameter(s) may be known or unknown. In the case where the distributions are completely known, we talk about *simple* hypothesis tests. In the contrary case, the tests are called *composite*.

1.3 Test statistics and associated probabilities

1.3.1 Test statistics

The second step in hypothesis testing is the computation of the test statistic, which allows to evaluate $\mathcal{H}_0$ w.r.t. a *decision rule*. The test statistic (denoted by T) is a single mathematical function derived from the data set and allowing to perform the hypothesis test.

Let $\boldsymbol{x} = [x_1, x_2, \ldots, x_N]^\top$ be an experimental data vector (superscript $\top$ denotes a transposition), the test statistic is expressed as a function of $\boldsymbol{x}$

$$T(\boldsymbol{x}) = T(x_1, x_2, \ldots, x_N), \tag{1.2}$$

where here, $T : \mathbb{R}^N \mapsto \mathbb{R}$. $T(.)$ is deterministic but $T(\boldsymbol{x})$ is random because $\boldsymbol{x}$ is the realization of a random process.

Depending on the data PDF under both hypotheses, and on the available knowledge on their parameters, there exists many ways to derive test statistics. For instance, the well-known Likelihood Ratio (LR) test, as presented in Section 1.4.1 is the most powerful test for a given probability of false alarm, but requires perfect knowledge of all parameters. Deriving a statistically controllable test statistic is often not a straightforward task, as the distribution of $T(\boldsymbol{x})$ under $\mathcal{H}_0$ should be known and there should be a clear distinction between its distribution under both hypotheses.

1.3.2 Performance characterization of a test

The significance level of a test (often denoted by α) is the probability of rejecting the null hypothesis by mistake; meaning that we decide $\mathcal{H}_1$ but $\mathcal{H}_0$ is true. This is a so-called type I error and the probability of making such error is known as the probability of false alarm:

$$\mathrm{P_{FA}} = \mathbb{P}(\text{decide } \mathcal{H}_1; \mathcal{H}_0 \text{ is true}). \tag{1.3}$$

The complement of the probability of false alarm is named the probability of correct rejection ($\mathrm{P_{CR}}$), which is the probability of correctly rejecting $\mathcal{H}_1$:

$$\mathrm{P_{CR}} = \mathbb{P}(\text{decide } \mathcal{H}_0; \mathcal{H}_0 \text{ is true}). \tag{1.4}$$

Besides the probability of false alarm, another kind of mistake is deciding $\mathcal{H}_0$ when $\mathcal{H}_1$ is true, namely a type II error. It happens with a probability called the probability of miss

($\mathrm{P_M}$) also denoted by β in the literature:

$$\mathrm{P_M} = \mathbb{P}(\text{decide } \mathcal{H}_0; \mathcal{H}_1 \text{ is true}). \tag{1.5}$$

The complement of the probability of miss is termed the probability of detection ($\mathrm{P_{Det}}$), also known as *power of the test*. This is the probability of correctly rejecting $\mathcal{H}_0$:

$$\mathrm{P_{Det}} = \mathbb{P}(\text{decide } \mathcal{H}_1; \mathcal{H}_1 \text{ is true}). \tag{1.6}$$

Table 1.1 summarizes these probabilities.

Truth \ **Decision**	**Decide $\mathcal{H}_0$**	**Reject $\mathcal{H}_0$**
$\mathcal{H}_0$ **True**	Correct Rejection $(1-\alpha)$ $\mathrm{P_{CR}} = \mathbb{P}(\mathcal{H}_0;\ \mathcal{H}_0)$	**Type I error** (α) $\mathrm{P_{FA}} = \mathbb{P}(\mathcal{H}_1;\ \mathcal{H}_0)$
$\mathcal{H}_0$ **False**	Type II error (β) $\mathrm{P_M} = \mathbb{P}(\mathcal{H}_0;\ \mathcal{H}_1)$	**Hit** $(1-\beta)$ $\mathrm{P_{Det}} = \mathbb{P}(\mathcal{H}_1;\ \mathcal{H}_1)$

Table 1.1: Probabilities associated to the decision making process.

In hypothesis testing, the larger the probability of detection w.r.t. the probability of false alarm, the most powerful the test is. A classical approach in optimizing a detector is to fix one of the error while minimizing the other error (for instance fix the probability of false alarm and calculate the test threshold, and maximize the probability of detection). The Neyman-Pearson (NP, or LR test) detector does this (see Section 1.4.1).

Note finally that a detector is called "optimal" w.r.t. a specific criterion. For instance, the NP maximizes the probability of detection for a fixed probability of false alarm considering deterministic and known parameters. This approach cannot be followed in the case of composite hypotheses. A Bayesian approach (see Section 1.5) aims to minimize the overall probability of errors which is evaluated under a probabilistic distribution of the parameters and of the hypotheses themselves. We will see in this dissertation that under some circumstances, a minimax criterion can be relevant in the case of deterministic parameters for composite hypotheses. This approach aims at minimizing the maximum probability of miss.

On another note, a *post-data* probability called the p-value represents the smallest significance level at which $\mathcal{H}_0$ would be rejected, given observed data. When comparing the p-value to the significance level of a test α (1.3), one will either reject $\mathcal{H}_0$ or accept $\mathcal{H}_0$. Assuming that $\mathcal{H}_0$ is true, the p-value associated to an observed test statistic T_{obs} is the probability of observing the test statistic T as *extreme* as T_{obs}, namely (for a one-sided right-tailed test)

$$p = \mathbb{P}(T(\boldsymbol{x}) \geq T_{\text{obs}} | \mathcal{H}_0). \tag{1.7}$$

For instance, if the significance level of a (one-sided right-tailed) test $\alpha = 0.05$, and $p = 10^{-4}$, this means that one is confident to reject $\mathcal{H}_0$ because $p \ll \alpha$.

1.4 Tests based on likelihood function

This Section presents two hypothesis tests (LR and GLR) following the frequentist approach, and their evaluation procedures. Let us consider the following statistical model

$$\begin{cases} \mathcal{H}_0 & : \boldsymbol{\theta} = \boldsymbol{\theta}_0, \\ \mathcal{H}_1 & : \boldsymbol{\theta} = \boldsymbol{\theta}_1,\ \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_0. \end{cases} \tag{1.8}$$

1.4.1 The Likelihood Ratio test

The LR test statistic, often denoted by Λ, is a simple hypothesis test (i.e., all the parameters are known). Through a *likelihood function*, LR measures the fit of the data w.r.t. the hypotheses. This function represents a probability function $p(\boldsymbol{x}; \boldsymbol{\theta})$ of the observed data set ($\boldsymbol{x} = [x_1, \ldots, x_N]^\top \in \mathbb{R}^N$) parametrized by the parameters $\boldsymbol{\theta}$ of the model. The likelihood ratio for (1.8) under $\mathcal{H}_1$ against $\mathcal{H}_0$ is

$$\Lambda(\boldsymbol{x}) := \frac{p(\boldsymbol{x}; \boldsymbol{\theta}_1)}{p(\boldsymbol{x}; \boldsymbol{\theta}_0)}. \tag{1.9}$$

This LR test statistic is compared to a test threshold γ which tunes the probability of false alarm. This yields the test

$$\text{LR} : \Lambda(\boldsymbol{x}) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma. \tag{1.10}$$

An important lemma addressing the detector that maximizes probability of detection at a fixed probability of false alarm was introduced by Neyman & Pearson [Neyman & Pearson 1933].

Lemma 1. ***The Neyman-Pearson Lemma*** *(see Appendix A.1 for a proof). The likelihood ratio test which rejects $\mathcal{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ in favor of $\mathcal{H}_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_1$ if*

$$\Lambda(\boldsymbol{x}) > \gamma, \tag{1.11}$$

where

$$P_{FA} = \mathbb{P}(\Lambda(\boldsymbol{x}) > \gamma; \mathcal{H}_0) = \alpha, \tag{1.12}$$

is the "Most Powerful test" of size α for a threshold γ.

The NP lemma defines a rejection region of $\mathcal{H}_0$ and maximizes the Lagrangian associated to probability of detection at fixed probability of false alarm. An example of LR testing is given in Appendix A.3.1. Given the above lemma, the Neyman-Pearson Detector (NPD) often acts as the *reference* for composite tests (see the comparisons of Section 4.3 and Chapter 7).

In the case where there are unknown parameters, LR cannot be implemented. In such a case, one can turn to one generalization of LR test, termed the Generalized Likelihood Ratio test.

1.4.2 The Generalized Likelihood Ratio test

The GLR test is a classical approach in composite hypothesis testing, widely applied in decision theory because it is often simple to implement. Moreover, it can be shown that asymptotically (in the number of data), GLR test is Uniformly Most Powerful[1] among all tests that are invariant[2] (see, e.g., [Scharf & Friedlander 1994, Lehmann & Romano 2005]). As its name indicates, GLR test is a generalization of the LR test, in which the unknown parameter $\boldsymbol{\theta}$ is replaced by its Maximum Likelihood (ML) estimate. This yields the test:

$$\text{GLR} : T_{\text{GLR}}(\boldsymbol{x}) := \frac{\max\limits_{\boldsymbol{\theta}_1} \; p(\boldsymbol{x}; \boldsymbol{\theta}_1)}{\max\limits_{\boldsymbol{\theta}_0} \; p(\boldsymbol{x}; \boldsymbol{\theta}_0)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma, \tag{1.13}$$

where the sets over which the maxima are sought are disjoint. Appendix A.3.2 shows an example of GLR test of a model with an unknown mean vector under $\mathcal{H}_1$.

For a GLR test, when the distribution of $T_{\text{GLR}}(\boldsymbol{x})$ under the null hypothesis is known, we can compute the threshold (see for instance (A.22)) to obtain a desired probability of false alarm. A detector that uses fixed threshold w.r.t. probability of false alarm is known as a Constant False Alarm Rate (CFAR) detector.

1.4.3 Decision rule and type of hypothesis tests

To evaluate whether one rejects or accepts a hypothesis via the test statistic T, a decision rule is required. In the Examples A.3.1 and A.3.2, we see that the decision rule is in the form

$$T \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma,$$

where γ is a certain value of threshold. The null hypothesis is rejected if the test statistic is more extreme than the threshold. Decision rule depends on the type (or, direction) of hypothesis tests, which is determined by the alternative hypothesis. There are three types of hypothesis testing: right-tailed, left-tailed, and two-tailed tests.

[1] Uniformly Most Powerful test: the test that maximizes the power of all alternatives against a fixed null hypothesis among all tests of a given size.

[2] Invariant: the test statistic remains unchanged under transformations (rotations, translations and scalings).

Example 1.4.1. One-tailed and two-tailed parameter testing problems

Assume that $\mathcal{H}_0$ is the same for each type of hypothesis test given by the following model

$$\mathcal{H}_0 : \theta = \theta_0.$$

The alternative hypothesis can be defined as

$$\mathcal{H}_1 : \theta, \text{ with } \theta_0 \notin \mathcal{H}_1.$$

The associated models w.r.t. each type of hypothesis test are

i. *right-tailed* test: $\mathcal{H}_1 : \theta > \theta_0$,

ii. *left-tailed* test: $\mathcal{H}_1 : \theta < \theta_0$,

iii. *two-tailed* test: $\mathcal{H}_1 : \theta \neq \theta_0$.

$\mathcal{H}_0$ is rejected if the test statistic T falls within the critical region (i.e., the *rejection region* of $\mathcal{H}_0$) . For the one-tailed tests (i. and ii.), there is only one side of *critical region*. For the two-tailed test, the critical region is on both sides of the data distribution under $\mathcal{H}_0$. ■

1.4.4 Evaluation of the test performances

A common evaluation of detection test's performance in a frequentist approach is by plotting probability of detection w.r.t. probability of false alarm as a function of the test threshold. This curve is called the Receiver Operating Characteristic (ROC) curve, a tool developed in the 1950s. We show an example below.

Example 1.4.2. Associated probabilities for a one sided, right-tailed test.

Assume that we observe a realization of a scalar random variable X, which is distributed according to the Gaussian PDF under both hypotheses, with known parameters θ_0, θ_1 where $\theta_1 > \theta_0$ and known variance σ^2.

$$\begin{cases} \mathcal{H}_0 & : x \sim \mathcal{N}(\theta_0, \sigma^2), \\ \mathcal{H}_1 & : x \sim \mathcal{N}(\theta_1, \sigma^2),\ \theta_1 > \theta_0 \end{cases} . \tag{1.14}$$

The PDF under both hypotheses is given by

$$p(x; \mathcal{H}_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - \theta_i)^2\right),\ i = 0, 1. \tag{1.15}$$

Following the LR approach, we have

$$\begin{aligned}\Lambda(x) &= \frac{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-\theta_1)^2\right)}{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-\theta_0)^2\right)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma' \\ &= \exp\left(-\frac{1}{2\sigma^2}(x^2 - 2\theta_1 x + \theta_1^2 - x^2 + 2\theta_0 x - \theta_0^2)\right) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma' \end{aligned} \tag{1.16}$$

Taking the logarithm of (1.16) conducts to

$$\begin{aligned}\Lambda(x) &= \left(\frac{\theta_1 - \theta_0}{\sigma^2}\right) x - \left(\frac{\theta_1^2 - \theta_0^2}{\sigma^2}\right) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \ln\gamma' \\ &= x \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \frac{\sigma^2}{\theta_1 - \theta_0}\left(\ln\gamma' + \left(\frac{\theta_1^2 - \theta_0^2}{\sigma^2}\right)\right) \\ \Lambda(x) &= x \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma. \end{aligned} \tag{1.17}$$

According to (1.17), we reject $\mathcal{H}_0$ if the test statistic is larger than γ, and since $\theta_1 > \theta_0$, the critical region (where we reject $\mathcal{H}_0$) is on the right side of the data set distribution under $\mathcal{H}_0$. This is a one-sided, right-tailed test as illustrated in Figure 1.1.

Next, we describe the associated probabilities of the test (1.17). The first type of error, which gives the significance level of the test is formulated as

$$\mathrm{P_{FA}} = \mathbb{P}(\Lambda(x) > \gamma; \mathcal{H}_0) = \int_{\gamma}^{+\infty} p(x; \mathcal{H}_0)\, dx, \tag{1.18}$$

and its complement, the probability of correct rejection

$$\mathrm{P_{CR}} = \mathbb{P}(\Lambda(x) < \gamma; \mathcal{H}_0) = \int_{-\infty}^{\gamma} p(x; \mathcal{H}_0)\, dx. \tag{1.19}$$

The second type of error

$$\mathrm{P_{M}} = \mathbb{P}(\Lambda(x) < \gamma; \mathcal{H}_1) = \int_{-\infty}^{\gamma} p(x; \mathcal{H}_1)\, dx, \tag{1.20}$$

which is also the complement of the power of the test

$$\mathrm{P_{Det}} = \mathbb{P}(\Lambda(x) > \gamma; \mathcal{H}_1) = \int_{\gamma}^{\infty} p(x; \mathcal{H}_1)\, dx. \tag{1.21}$$

Denoting by Φ the Cumulative Distribution Function (CDF) of a standard normal distribution

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{u^2}{2}}\, du, \tag{1.22}$$

we can write the probability of false alarm (1.18) and probability of detection (1.21) as

$$\mathrm{P_{FA}} = 1 - \Phi\left(\frac{\gamma - \theta_0}{\sigma}\right), \tag{1.23}$$

$$\mathrm{P_{Det}} = 1 - \Phi\left(\frac{\gamma - \theta_1}{\sigma}\right). \tag{1.24}$$

These probabilities are illustrated in Figure 1.1. As depicted, one cannot minimize simultaneously the probabilities of error (probability of false alarm and probability of miss), because minimizing the probability of false alarm results to maximizing the probability of miss (consequently minimizing as well the probability of detection, $\mathrm{P_{Det}} = 1 - \mathrm{P_M}$).

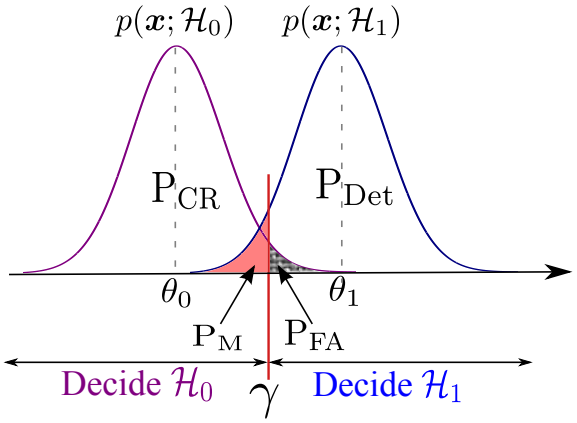


Figure 1.1: Distributions under hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ of model (1.14), and illustration of the probabilities associated to the test statistic (1.17). This test is called one sided and right-tailed test ($\Lambda \gtrless_{\mathcal{H}_0}^{\mathcal{H}_1} \gamma$ and $\theta_1 > \theta_0$).

■

The performance of a detector depends also on the Signal to Noise Ratio (SNR) associated to the test statistic. For instance, if a signal associated to NP detector (e.g., (1.17)) has very low SNR, its PDFs are nearly flat and very wide, as shown in Figure 1.2 (dark purple line under $\mathcal{H}_0$ and blue dash-dots under $\mathcal{H}_1$). In such a case, the probability of detection is approximately equivalent to the probability of false alarm. If the associated SNR is high, its PDFs become sharper and thinner (green line and black dash-dots in Figure 1.2).

The power of any test increases with the probability of false alarm (as also visible in Figure 1.1). The ROC curve can be plotted to evaluate the power of a test at any probability of false alarm. The ROC curve can be used to compare different tests at fixed SNR, or to evaluate the evolutions of power for varying SNR.

Several examples of ROC curves are depicted in Figure 1.3, varying in performances. The probability of detection and probability of false alarm decrease when the threshold increases. An ideal detector would produce a ROC curve with a Γ shape (i.e., $\mathrm{P_{Det}} = 1$ at all probability of false alarm values). However, such detector generally does not exist. For example, NP detector tends towards this ideal detector only when its SNR $\to \infty$. As the

SNR increases, the PDFs in Figure 1.1 become sharper and thinner, as shown in Figure 1.2 (green line under $\mathcal{H}_0$ and black dash-dots under $\mathcal{H}_1$).

On the contrary, if we obtain a diagonal ROC curve (commonly called the "random line") where the detection rate is equal to the false alarm rate, this means that the SNR is too low for the two hypotheses to be statistically distinguished, or that the particular detector under investigation is completely inefficient to distinguish them. To measure precisely the closeness of a curve to the diagonal (or vice versa, to the ideal), one can compute the Area Under Curve (AUC) of the ROC, thus obtaining AUC value between 0.5 to 1 (random line to ideal detector). AUC will be used in Chapter 7).

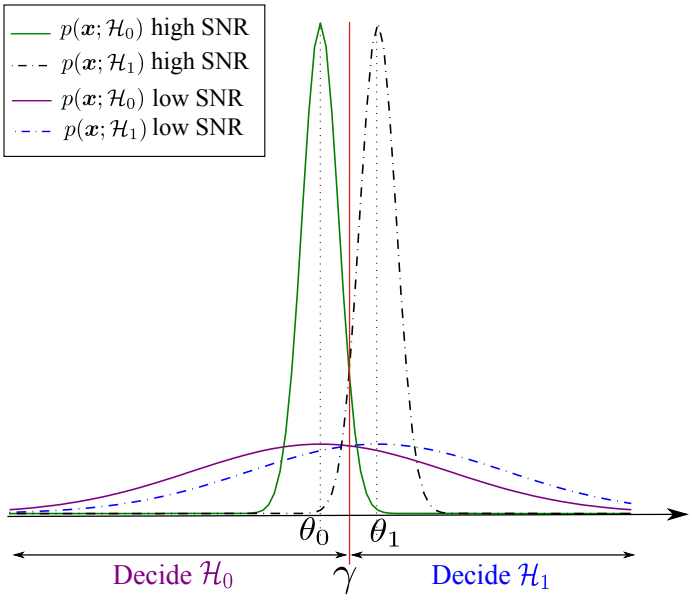


Figure 1.2: Illustration of the probability density functions (under $\mathcal{H}_0$ and $\mathcal{H}_1$) for high SNR and low SNR associated to the test statistic (1.17) for model (1.14). For high SNR, the PDFs become sharper and thinner (green line and black dash-dots). For low SNR, the PDFs become flatter and wider (dark purple line and blue dash-dots), yielding $\mathrm{P}_{\mathrm{Det}} \approx \mathrm{P}_{\mathrm{FA}}$.

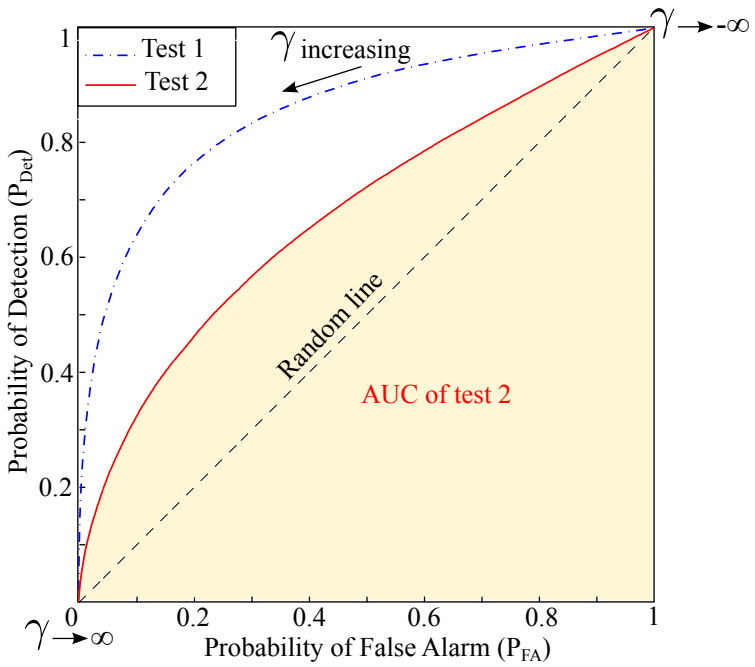


Figure 1.3: Comparison of ROC curves: probability of detection against probability of false alarm for two curves. Test 1 shows better performance than test 2. The yellow area depicts the Area Under Curve of the ROC for test 2. The ROC curve can be used to compare different tests at fixed SNR, or to evaluate the detection power of a same test for varying SNR (i.e., SNR for blue dash dots ROC curve $>$ SNR for solid red ROC curve). The "random line" indicates that $\mathrm{P_{Det}} = \mathrm{P_{FA}}$.

1.5 A glance at Bayesian approach

The Bayesian method considers the unknown parameters as realizations of random variables and assigns prior PDFs to them. Given a data vector $\boldsymbol{x} = [x_1, \ldots, x_N]^\top \in \mathbb{R}^N$, where $\boldsymbol{x} = \boldsymbol{\theta} + \boldsymbol{n}$, let us consider the following model

$$\begin{cases} \mathcal{H}_0 & : \boldsymbol{\theta} \sim p_0(\boldsymbol{\theta}), \\ \mathcal{H}_1 & : \boldsymbol{\theta} \sim p_1(\boldsymbol{\theta}), \end{cases}$$

where $p_i(\boldsymbol{\theta})$ are the prior probabilities on the random parameter under $\mathcal{H}_i$, $i = 0, 1$. The Bayes' theorem (also known as Bayes' rule) defines the joint probability $p(\boldsymbol{x}, \boldsymbol{\theta})$ in function of posterior probability $p(\boldsymbol{\theta}|\boldsymbol{x})$ and prior probability on the data $p(\boldsymbol{x})$

$$p(\boldsymbol{x}, \boldsymbol{\theta}) = p(\boldsymbol{\theta}|\boldsymbol{x})\, p(\boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}), \tag{1.25}$$

or in function of the likelihood $p(\boldsymbol{x}|\boldsymbol{\theta})$ and the prior probability on the unknown parameters $p(\boldsymbol{\theta})$. The posterior probability $p(\boldsymbol{\theta}|\boldsymbol{x})$ is related to the probability of *event* $\boldsymbol{\theta}$ given *event*

$\boldsymbol{x}$, which can be reformulated from (1.25) to

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p(\boldsymbol{x},\boldsymbol{\theta})}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|\boldsymbol{\theta})\,p(\boldsymbol{\theta})}{p(\boldsymbol{x})}. \tag{1.26}$$

The PDF of the data under both hypotheses ($\mathcal{H}_0$ and $\mathcal{H}_1$) is

$$p(\boldsymbol{x}|\mathcal{H}_i) = \int p(\boldsymbol{x}|\boldsymbol{\theta})\,p_i(\boldsymbol{\theta})\,d\boldsymbol{\theta} \quad \text{for } i = 0, 1. \tag{1.27}$$

The Bayesian *strategy* to obtain an optimal detector is to minimize the probability of error ($\mathrm{P_E}$)

$$\begin{aligned}
\mathrm{P_E} &= \mathbb{P}\{\text{decide } \mathcal{H}_0, \mathcal{H}_1 \text{ is true}\} + \mathbb{P}\{\text{decide } \mathcal{H}_1, \mathcal{H}_0 \text{ is true}\} \\
&= \mathbb{P}(\mathcal{H}_0, \mathcal{H}_1) + \mathbb{P}(\mathcal{H}_1, \mathcal{H}_0) \\
&= \mathbb{P}(\mathcal{H}_0|\mathcal{H}_1)\,\mathbb{P}(\mathcal{H}_1) + \mathbb{P}(\mathcal{H}_1|\mathcal{H}_0)\,\mathbb{P}(\mathcal{H}_0) \\
\mathrm{P_E} &= \mathrm{P_M}\,\mathbb{P}(\mathcal{H}_1) + \mathrm{P_{FA}}\,\mathbb{P}(\mathcal{H}_0).
\end{aligned} \tag{1.28}$$

The probabilities $\mathbb{P}(\mathcal{H}_1)$ and $\mathbb{P}(\mathcal{H}_0)$ represent an *initial belief* in the stated hypotheses, and it can be shown that (see Appendix A.2 for a proof) if one decides $\mathcal{H}_1$ when

$$\frac{p(\boldsymbol{x}|\mathcal{H}_1)}{p(\boldsymbol{x}|\mathcal{H}_0)} > \frac{p(\mathcal{H}_1)}{p(\mathcal{H}_0)} = \gamma, \tag{1.29}$$

then an optimal detector (i.e., one that minimizes $\mathrm{P_E}$) is obtained. This detector is termed the Bayes Factor.

The test statistic minimizing the probability of error is

$$T_{\mathrm{Bayes}}(\boldsymbol{x}) := \frac{p(\boldsymbol{x}|\mathcal{H}_1)}{p(\boldsymbol{x}|\mathcal{H}_0)} = \frac{\int p(\boldsymbol{x}|\boldsymbol{\theta})\,p_1(\boldsymbol{\theta})\,d\boldsymbol{\theta}}{\int p(\boldsymbol{x}|\boldsymbol{\theta})\,p_0(\boldsymbol{\theta})\,d\boldsymbol{\theta}} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma. \tag{1.30}$$

In practice, it can be difficult to choose the prior PDFs and the integration might be complicated. Appendix A.3.3 shows an example of a Bayesian test statistic.

1.6 Minimax approach in detection

The minimax strategy also sometimes called MinMax or MM aims to minimize a loss function in a worst-case scenario, hence the term *minimax*. The concept of minimax was first applied in the game theory by Von Neumann [Neumann 1928] with the objective of minimizing the maximum loss of each player in a *zero-sum* game. It can equivalently be viewed as maximizing the minimum gain, a problem of type *maximin*.

In this dissertation, the detection problem is to maximize the minimum *detection power* (i.e., probability of detection), following a maximin rule. We shall generically use the term "minimax" to refer to this objective.

The first application of minimax strategy to statistical decision theory was introduced by

Abraham Wald in the 1940s [Wald 1945, Wald 1950]. Following the work of Von Neumann, A. Wald interpreted the zero-sum two players game as a statistical inference problem. He restricted the game procedure to the case where the first player is viewed as "Nature", which follows distribution F. This distribution is made known to everybody. On the other hand, the second player acts as a statistician, seeking a decision rule $\mathcal{D}$ to maximize the minimum payoff of the game. The outcome (i.e., the payoff) represents a risk function: $f(F, \mathcal{D})$. The Wald's maximin risk is defined as

$$\mathcal{D}^* = \sup_{\mathcal{D}} \inf f(F, \mathcal{D}). \tag{1.31}$$

In the specific framework of detection, minimaxity has been studied widely in the literature. For instance, in Chapter 8 of [Lehmann & Romano 2005], maximin principle for general decisions problems is discussed.

In Chapter 9 of the same book, maximin algorithmic aspects (of the problem introduced in Chapter 8) are presented where the parameters and data samples are considered to be finite or infinite.

In [Kassam & Poor 1985], minimax detection strategy is viewed as a robust approach in signal processing because the worst-case detection performance is maximized, yielding "stable" overall performances than other strategies that usually rely on minimizing the MSE or minimizing the probability of error.

On another note, minimax arguments related to *estimation risks* paved in the 90's the road of *sparsity* promoting methods based on thresholding functions for denoising and inverse problems [Donoho & Johnstone 1998]. Chapter 3 tackles the related topics on sparsity.

This Section discusses the general strategy to minimax approach, and the particularity of this approach applied in our framework.

1.6.1 General minimax strategy

Assume that we have an unknown parameter $\theta \in \Theta$ and a measurable function $\psi \in \Psi$, associated to a function f. Minimax strategy deals with determining ψ to minimize the function f, whilst computing θ to maximize f

$$\inf_{\psi \in \Psi} \sup_{\theta \in \Theta} f(\psi, \theta). \tag{1.32}$$

Consequently, a maximin problem can be defined as

$$\sup_{\psi \in \Psi} \inf_{\theta \in \Theta} f(\psi, \theta). \tag{1.33}$$

The minimax problem (1.32) is equivalent to the maximin problem (1.33) if f has a saddle point [Sion 1958]. This type of problem not only occurs in statistical signal processing for estimation theory or hypothesis testing, but also in machine learning for classification tasks. In a classification task, minimax strategy consists in searching a decision boundary that separates maximally the distance between this boundary and the data set [Boser

et al. 1992].

In statistical decision theory, we often take $f(\psi, \theta)$ as a risk function, that is the expected value of some loss function L w.r.t. the realizations of a random variable X

$$f(\psi, \theta) = \int_{\mathrm{dom}(X)} \mathrm{L}(\psi(x), \theta)\, p_\theta(x)\, \mathrm{d}x, \tag{1.34}$$

where $\theta \in \Theta$ is an unknown continuous parameter, ψ is a decision function and $p_\theta(x)$ is the probability distribution function of X (e.g., see [Lehmann & Romano 2005]). We then seek to determine the decision function ψ minimizing f, which usually depends on an unknown parameter $\theta \in \Theta$. In this context, the minimax procedure is the form of (1.32).

1.6.2 Minimax testing

In the literature of (multiple) hypothesis testing, the case where rare and weak deviations are suspected among a large number of null hypotheses has received an increasingly large interest in the last decade, e.g., [Ingster & Suslina 2003, Donoho & Jin 2004, Arias-Castro *et al.* 2010, Mary & Ferrari 2014]. In these works, the minimax approach attempts to establish conditions under which the joint null hypothesis can be asymptotically distinguished from all alternatives of specific sets (typically sparse signals in ℓ_p balls[3], with $0 \leq p \leq 1$) and to derive test statistics that provide such distinguishability.

Example 1.6.1. Minimax testing: distinguishing $\mathcal{H}_0$ and $\mathcal{H}_1$

In [Ingster & Suslina 2003], the authors considered the following model

$$\begin{cases} \mathcal{H}_0 & : \theta = 0 \\ \mathcal{H}_1 & : \theta \in \Theta_n \end{cases}. \tag{1.35}$$

For a given type I error, $\alpha \in [0, 1]$, the minimax criterion in this case is formulated as minimizing over level-α tests, the maximum type II error (β)

$$T_{\mathrm{MM}}(\alpha, \Theta_n) = \inf_{\psi \in \Psi_\alpha} \sup_{\theta \in \Theta_n} \beta(\psi, \theta), \tag{1.36}$$

where Ψ_α is a set of tests ψ of size smaller than α, $\forall \alpha \in [0, 1]$. Here, $\Theta_n \subset \mathbb{R}^n$ is considered as continuous parameter (ℓ_p ball) and the hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ are said to be

- *distinguishable* if $T_{\mathrm{MM}}(\alpha, \Theta_n) \to 0$, $n \to \infty$.
- *indistinguishable* if $T_{\mathrm{MM}}(\alpha, \Theta_n) \to 1 - \alpha$, $n \to \infty$.

These authors concluded that the condition of distinguishability holds if sets Θ_n contain large enough signals.

[3] A close ℓ_p ball of radius r is defined as the set $\{\boldsymbol{x} \in \mathbb{R}^N : \sum_{n=1}^N |x_n|^p \leq r^p\}$.

Note that a possible approach (e.g., [Jiao *et al.* 2012]) is to associate *priors* under the alternative $\mathcal{H}_1$ in the form of distribution of the parameters $\theta \in \Theta_n$. ■

1.7 Discussion

This Chapter presented some basic aspects of detection theory. We first showed how to model a set of hypotheses given a detection problem, and detailed the four steps in hypothesis testing: stating hypotheses from the problem, identify and derive the test statistic, select a significance level of the test, and finally draw conclusions.

An example to calculate the probability of detection (power of the test) and probability of false alarm once the corresponding test statistic is obtained was shown. These quantities allow to evaluate the performance of a detector, for instance via a ROC curve where the detection rate is plotted against the false alarm as a function of the test threshold. When comparing many tests' performances, we can assess the Area Under Curves of the ROCs.

The research work in this dissertation is based on a GLR test, hence the highlight on the frequentist approach. Nevertheless, the Bayesian approach to hypothesis testing was discussed briefly in this Chapter. Several examples of both approaches are provided in Appendix A.3.1-A.3.3.

The general minimax (and maximin) strategy was also discussed in this Chapter. Minimax hypothesis tests in literature are usually asymptotic in the number of tests (thus in the number of alternatives) and the amplitude under $\mathcal{H}_1$ has to be large (tends to infinity) for a distinguishability between both hypotheses.

In contrast, we consider a "one among many" detection problem where under $\mathcal{H}_1$, only one alternative $\boldsymbol{s}_i$ is activated belonging to a possibly large known library $\boldsymbol{S} \in \mathbb{R}^{N \times L} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_L]$. This model writes

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

where $\boldsymbol{x}$ and $\boldsymbol{n} \in \mathbb{R}^N$. The alternatives of $\boldsymbol{S}$ are ℓ_2 normalized and we consider that N and L are fixed. In our setting, we apply the maximin strategy to maximize the worst probability of detection, at fixed probability of false alarm. Chapter 4 presents observations at the origin of this strategy to our framework of detection. The following Chapter (5) describes in detail the resulting theoretical aspects.

Before going into the detail of our proposed approaches in Part II, the next Chapter talks about target detection in Hyperspectral Imaging and provides important elements to understand the applicative part of this dissertation.

Chapter 2

Target detection in Hyperspectral Imaging

Contents

2.1 Introduction

A picture represents a captured scene, and consists of two *spatial* dimensions (say, x and y). The acquisition of spectral information on every spatial sample (or pixel) of a two-dimensional picture is known as *Spectral Imaging*, which introduces a third dimension, termed *spectral* dimension, represented by wavelength (λ) or frequency. Spectral Imaging combines two scientific branches, namely spectroscopy and imaging (or, photography), allowing the recording of electromagnetic spectrum beyond the range of visible frequencies. Spectroscopy concerns the interaction between an object (or matter) and the radiation energy as a function of the spectral channels. Gathering images from more than one frequency band (i.e., different channels) along the spectrum is known as *Multispectral Imaging*, yielding a three-dimensional image cube. Multispectral scanners deliver several images in discrete and narrow bands, thus the data spectrum of an observed object is non continuous. The advancement in remote sensing and the need to obtain ample spectral information of a target (thousands of channels), led to a new paradigm termed *Hyperspectral Imaging* (HSI).

Hyperspectral Imaging opens a new variety of studies in signal and image processing community as it provides contiguous and very narrow spectral bands that typically span hundreds to thousands wavelength channels (Figure 2.1 depicts an example of a HSI data cube). Thus, a vast portion of the electromagnetic spectrum is collected, providing more spectral information than a classical multispectral image cube.

Note that, according to some authors, the difference between Multispectral and Hyperspectral Imaging is not the number of wavelength channels it spans, but the *narrowness* of each band [Govender *et al.* 2007]. This distinction emphasizes the contiguous aspect of spectral channels obtained from HSI. It means that a scanner with say only 18 bands

but each 10nm wide covering the Near Infrared range is considered as Hyperspectral. In contrast, a scanner with 18 discrete bands, covering the Visible, Near Infrared and Middle Infrared range is considered as Multispectral (the width of each band is not very narrow, hence non continuous spectral bands).

Spectral Imaging in general is used for Earth (ground and sky) and also space monitoring, but it is not limited to only these applications. For instance, military uses this technology to surveil country's frontiers, while astronomers observe planets, stars and galaxies.

With HSI, the obtained data cube expands the horizon of spectral imaging, where the acquisition (as shown in Figure 2.2), is made by spectrometers, that can be airborne (i.e., carried through the air by aircrafts, for example the AVIRIS instrument [Green *et al.* 1998]), on the ground (e.g., integral field spectrographs built for telescope, like the MUSE instrument [Caillier *et al.* 2012], see Section 7.2), or in space attached to satellites (e.g., the Hyperion instrument [Pearlman *et al.* 2011]). Beside these large instruments, handheld HSI sensors, which are far less expensive are also being used for specific applications such as food monitoring (safety and quality) [Huang *et al.* 2014] and the detection of aerosols in air [Hinnrichs *et al.* 2004].

Three core tasks in the processing of Hyperspectral Images are

- *classification* (i.e., assignment) of single pixels to a set of classes defined w.r.t. some chemical or physical characteristics,
- *unmixing* of the pixel spectra to separate the pure spectrum of a certain material (known as *endmember*) from the mixed pixels,
- *detection* of a target spectrum presents in the image cube, or more generally an "object" which can spread on several pixels.

As one can imagine, HSI provides high spectral resolution images (very narrow band, typically $\leq$ 10nm wide), yielding thus huge data cubes. Despite the advantages of HSI, the first difficulty (as far as information extraction in concerned) is due to the cost of high volume data storage, and the second concerns the processing part, where the computation can be very complex (e.g., in astrophysics, the dimensions of the MUSE data cube is $300 \times 300 \times 3600$, which may lead to about 3.24×10^8 elementary mathematical operations for post-processing [Paris *et al.* 2013a]). For such very large data, the computation of sophisticated algorithms might be intractable.

Note that most of the study in this dissertation focuses on techniques that exploit only spectral information. Towards the end (Section 7.4), a numerical *spatio-spectral* simulations are however also presented.

This chapter focuses on classical methods for the detection of target signatures present in noisy hyperspectral image cubes. In the HSI literature, detectors can be categorized in two types, namely the *Spectral Matching* and the *Anomaly* detectors [Manolakis *et al.* 2009].

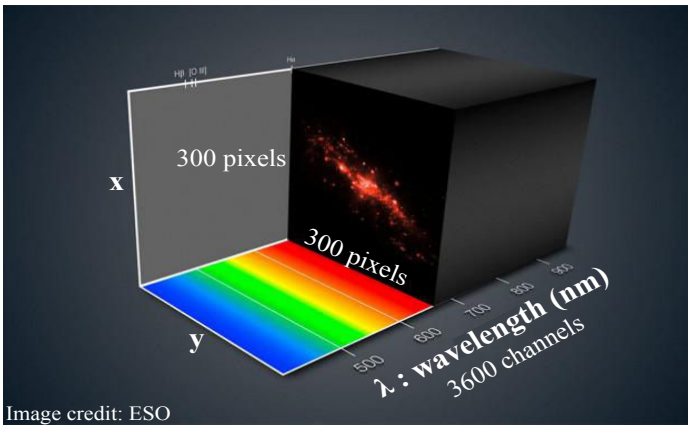


Figure 2.1: An example of a Hyperspectral image cube illustrated for the Multi Unit Spectroscopic Explorer (MUSE) instrument, of dimension 300×300 pixels at 3600 wavelength channels [Caillier *et al.* 2012]. Image credit to the European Southern Observatory (ESO).

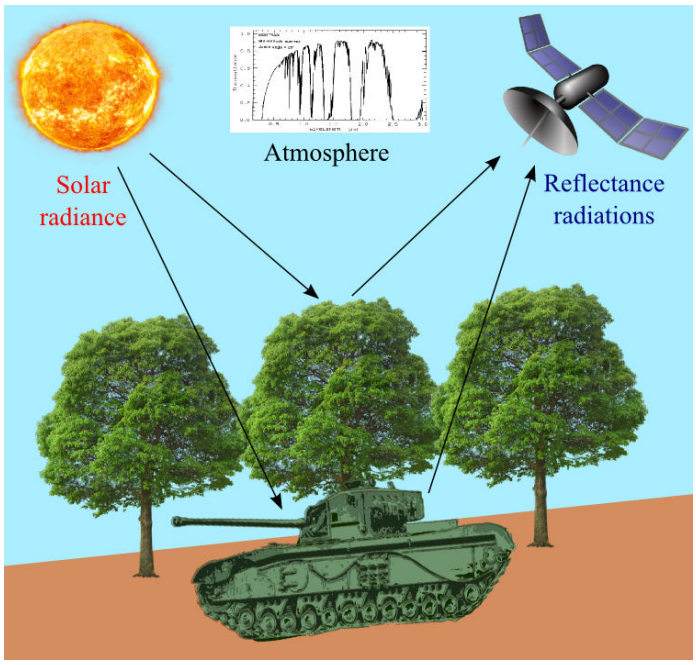


Figure 2.2: Illustration of the Hyperspectral image acquisition by satellite. The instrument mounted on the satellite retrieves the reflectance radiations from all objects in the scene: trees, soil and military tank, along with the atmospheric noise.

2.2 Target detectors

The similarity between target detection and classification in HSI is that both of these tasks use spectral signatures to process the data pixels. Figure 2.3 from JPL-NASA shows the concept of HSI for the AVIRIS (Airborne Visible InfraRed Imaging Spectrometer) instrument. Each type of matter or objects (e.g., atmosphere, soil, water and vegetation) has unique pure spectral signature and are known as the "endmembers". Often, in *classification*, the main objective is to separate the background from the object of interest (binary classifier). A good classifier presents a low probability of misclassification.

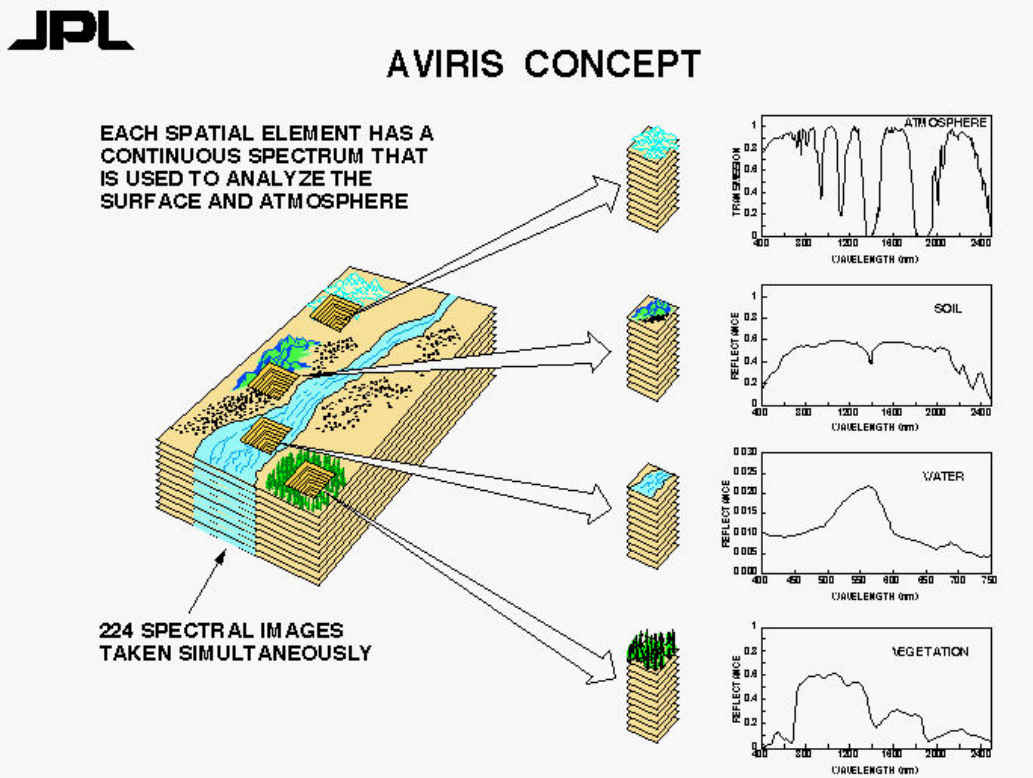


Figure 2.3: Each pixel of the image cube represents a contiguous reflectance spectrum along the wavelength channels. The knowledge on spectral signatures of each objects (e.g., atmosphere, soil, water, and vegetation) concedes the processing of Hyperspectral images. These pure signatures are called "endmembers". Image credit to the Jet Propulsion Laboratory (JPL) and NASA.

In the framework of target *detection*, the aim is to identify the target signature in the noisy signals of the data cube. This is no difference with the classical binary detection problem described in Chapter 1 (1.1), where one distinguishes the alternative hypothesis (target present in the data) against the null hypothesis (target absent). The null is in this case often considered as the background of the observed scene.

In the case where there are more than one target to be detected, it becomes a multiple hypothesis detection problem. In classification, this conducts to a multiclass problem.

A general strategy is then to treat multiclass problems as multiple binary classifications problems. In this type of problems, there exists two approaches: *one against all* and *one against one*. In the first approach, the distinction is made between each class and the rest of the noisy signals. While for the one against one approach, the distinction is between one class and another class (through a pairwise evaluation).

Geologists were the earliest who benefited from HSI for the detection of minerals and vegetations in parks [Kokaly *et al.* 1998]. Not long after that, this technology (specifically in detection) received a lot of interests in other communities such as military (e.g, for civilian rescue [Eismann *et al.* 2009]), astrophysicists (detection of galaxies [Caillier *et al.* 2012, Paris *et al.* 2013a]) and arborists (detection of infected trees in natural forests [Lee & Cho 2006]). Beside the computation complexity of dealing with large data, other challenges of HSI target detectors are low SNR of the image cube due to many factors such as atmospheric noise, instrument's noise: increased spatio-spectral resolution leading to low photon counts, or inherently low amplitude target signatures. Furthermore, due to the high dimensionality of a HSI data cube, testing a large number of pixels increases substantially the probability of false alarm (see Section 4.3). To control the false alarm rate, the threshold level has to be raised, resulting in the non detection of weak signals.

HSI target detection algorithms of the literature can be designed using statistical, physical or heuristic approaches. The performances of any detector are always evaluated using statistical inference. Typically, an *ideal* detector is the one that maximizes the probability of detection at a fixed probability of false alarm (i.e, the Neyman-Pearson criterion). Alternatively, one can sometimes opt for a *minimax* criterion, as presented in the previous Chapter 1.

For future reference, detailed comparison of the existing target detection algorithms for HSI can be found in [Manolakis *et al.* 2009, Manolakis *et al.* 2014, Nasrabadi 2014]. The following Sections discuss the particularity of the two detectors mentioned earlier, namely Spectral Matching and Anomaly detectors.

2.2.1 Spectral Matching

The major difference between a Spectral Matching detector and an Anomaly detector is that Spectral Matching requires a *prior knowledge* of the target spectral. This type of detectors relies on Matched Filter principles, where one matches (or correlates) a known target signal with the data spectrum under test to make a decision on its presence or not in the data. Some examples of the existing methods in this category are Spectral Matched Filter (SMF) [Robey *et al.* 1992], Matched Subspace Detector (MSD) [Scharf & Friedlander 1994] and Adaptive Subspace Detector (ASD) [Kraut *et al.* 2001].

Example 2.2.1. Spectral Matched Filter.

Let us consider the following detection problem, where we assume that target is absent under the null hypothesis, and target is present under the alternative hypothesis in the form

$$\begin{cases} \mathcal{H}_0 & : \boldsymbol{x} = \boldsymbol{n}, & \boldsymbol{n} \sim \mathcal{N}(\boldsymbol{0}, \Sigma), \\ \mathcal{H}_1 & : \boldsymbol{x} = \alpha \boldsymbol{s} + \boldsymbol{n}, & \alpha > 0, \end{cases} \tag{2.1}$$

where α is an unknown amplitude, $\boldsymbol{s} \in \mathbb{R}^N$ is the known target spectral and $\boldsymbol{n} \in \mathbb{R}^N$ is Gaussian random (background) noise, with zero mean, known and same covariance matrix Σ under both hypotheses. The GLR for (2.1) is

$$T_{\mathrm{GLR}}(\boldsymbol{x}) := \frac{\max\limits_{\alpha>0} p(\boldsymbol{x}; \alpha \boldsymbol{s})}{p(\boldsymbol{x}; \boldsymbol{0})} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma', \tag{2.2}$$

where γ' is a threshold. Next, we calculate the ML estimate of α

$$\begin{aligned} \alpha^{\mathrm{ML}} &= \arg\max_{\alpha>0} p(\boldsymbol{x}; \boldsymbol{\theta}) = \arg\min_{\alpha>0} \frac{1}{2}(\boldsymbol{x} - \alpha \boldsymbol{s})^\top \Sigma^{-1} (\boldsymbol{x} - \alpha \boldsymbol{s}) \\ &\frac{\partial}{\partial \alpha} \left(\frac{1}{2} (\boldsymbol{x}^\top \Sigma^{-1} \boldsymbol{x} - 2 \boldsymbol{s}^\top \alpha \Sigma^{-1} \boldsymbol{x} + \alpha^2 \boldsymbol{s}^\top \Sigma^{-1} \boldsymbol{s}) \right) = 0 \;\Rightarrow\; \alpha^{\mathrm{ML}} = \frac{\boldsymbol{s}^\top \boldsymbol{x}}{\boldsymbol{s}^\top \boldsymbol{s}}. \end{aligned} \tag{2.3}$$

Injecting α_{ML} in (2.2) and taking the logarithm leads to

$$\begin{aligned} T_{\mathrm{GLR}}(\boldsymbol{x}) &= -\frac{1}{2}(\boldsymbol{x} - \alpha^{\mathrm{ML}} \boldsymbol{s})^\top \Sigma^{-1} (\boldsymbol{x} - \alpha^{\mathrm{ML}} \boldsymbol{s}) + \frac{1}{2} \boldsymbol{x}^\top \Sigma^{-1} \boldsymbol{x} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \ln \gamma' \\ T_{\mathrm{SMF}}(\boldsymbol{x}) &= \frac{\boldsymbol{s}^\top \Sigma^{-1} \boldsymbol{x}}{\sqrt{\boldsymbol{s}^\top \Sigma^{-1} \boldsymbol{s}}} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma \;\Leftrightarrow\; \frac{\boldsymbol{s}^\top (\Sigma^{-\frac{1}{2}})^\top \mathbf{I} \Sigma^{-\frac{1}{2}} \boldsymbol{x}}{\| \Sigma^{-\frac{1}{2}} \boldsymbol{s} \|_2} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma, \end{aligned} \tag{2.4}$$

where $\gamma = \sqrt{2 \ln \gamma'}$.

T_{SMF} is thus a test that computes the (normalized) correlation between the whitened target $\Sigma^{-\frac{1}{2}} \boldsymbol{s}$ and the whitened data $\Sigma^{-\frac{1}{2}} \boldsymbol{x}$, and compares the result to a threshold γ. ■

Example 2.2.2. Matched Subspace Detector.

MSD deals with the problem of detecting under $\mathcal{H}_1$ a signal $\boldsymbol{s} = \boldsymbol{S}\boldsymbol{y} \in \mathbb{R}^N$ that lies in the subspace spanned by $\boldsymbol{S}$ (with $\boldsymbol{S} \in \mathbb{R}^{N \times p}$ and presents in some independent noise subspace (characterized by $\boldsymbol{B} \in \mathbb{R}^{N \times r}$). Let us consider the following model

$$\begin{cases} \mathcal{H}_0 & : \boldsymbol{x} = \boldsymbol{B}\boldsymbol{\theta}_0 + \boldsymbol{n}, & \boldsymbol{n} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \mathbf{I}), \\ \mathcal{H}_1 & : \boldsymbol{x} = \boldsymbol{S}\boldsymbol{y} + \boldsymbol{B}\boldsymbol{\theta}_1 + \boldsymbol{n}, & \|\boldsymbol{y}\|_2^2 > 0. \end{cases} \tag{2.5}$$

$\boldsymbol{B}\boldsymbol{\theta}_i$, $i = 0, 1$ lies in the subspace spanned by $\boldsymbol{B}$, where $\boldsymbol{\theta}_i \in \mathbb{R}^r$, $r < N - p$, is considered unknown. Matrices $\boldsymbol{B}$ and $\boldsymbol{S}$ are considered known, full-rank and linearly independent of each other. The covariance matrix $\sigma^2 \mathbf{I}$ is considered known and equal under both hypotheses.

The GLR for (2.5) is

$$T_{\mathrm{GLR}}(\boldsymbol{x}) := \frac{\max\limits_{\boldsymbol{\theta}_1} p(\boldsymbol{x}; \boldsymbol{S}\boldsymbol{y} + \boldsymbol{B}\boldsymbol{\theta}_1)}{\max\limits_{\boldsymbol{\theta}_0} p(\boldsymbol{x}; \boldsymbol{B}\boldsymbol{\theta}_0)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma', \tag{2.6}$$

where γ' is a threshold. The noise $\boldsymbol{n}_i$, $i = 0, 1$ under both hypotheses can be written as

$$\boldsymbol{n}_0 = \boldsymbol{x} - \boldsymbol{B}\boldsymbol{\theta}_0, \tag{2.7}$$

$$\boldsymbol{n}_1 = \boldsymbol{x} - [\boldsymbol{S}\,\boldsymbol{B}] \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{\theta}_1 \end{bmatrix}. \tag{2.8}$$

Taking the logarithm of the GLR (2.6) and defining

$$\boldsymbol{n}_0^{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}_0} \boldsymbol{n}_0, \tag{2.9}$$

$$\boldsymbol{n}_1^{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}_1} \boldsymbol{n}_1. \tag{2.10}$$

leads to

$$\begin{aligned} T_{\mathrm{GLR}}(\boldsymbol{x}) &= -\frac{1}{2\sigma^2}\|\boldsymbol{n}_1^{\mathrm{ML}}\|_2^2 + \frac{1}{2\sigma^2}\|\boldsymbol{n}_0^{\mathrm{ML}}\|_2^2 \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \ln\gamma' \\ &= \frac{1}{\sigma^2}\left(\|\boldsymbol{n}_0^{\mathrm{ML}}\|_2^2 - \|\boldsymbol{n}_1^{\mathrm{ML}}\|_2^2\right) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma = 2\ln\gamma'. \end{aligned} \tag{2.11}$$

Denoting $P_{\boldsymbol{B}} = \boldsymbol{B}(\boldsymbol{B}^\top\boldsymbol{B})^{-1}\boldsymbol{B}^\top$ as the orthogonal projection onto the subspace spanned by $\boldsymbol{B}$, and $P_{\boldsymbol{SB}} = [\boldsymbol{S}\,\boldsymbol{B}]\left([\boldsymbol{S}\,\boldsymbol{B}]^\top[\boldsymbol{S}\,\boldsymbol{B}]\right)^{-1}[\boldsymbol{S}\,\boldsymbol{B}]^\top$ as the orthogonal projection onto the subspace spanned by $\boldsymbol{SB}$, the ML estimates of $\boldsymbol{n}_0$ and $\boldsymbol{n}_1$ are [Scharf & Friedlander 1994]

$$\boldsymbol{n}_0^{\mathrm{ML}} = (\mathbf{I} - P_{\boldsymbol{B}})\boldsymbol{x} = P_{\boldsymbol{B}}^{\perp}\,\boldsymbol{x}, \tag{2.12}$$

$$\boldsymbol{n}_1^{\mathrm{ML}} = (\mathbf{I} - P_{\boldsymbol{SB}})\boldsymbol{x} = P_{\boldsymbol{SB}}^{\perp}\,\boldsymbol{x}, \tag{2.13}$$

where notation $P_{\boldsymbol{M}}^{\perp}$ denotes the orthogonal projection on the subspace orthogonal to $\boldsymbol{M}$ (in (2.12), $\boldsymbol{M} = \boldsymbol{B}$, and in (2.13), $\boldsymbol{M} = \boldsymbol{SB}$). Injecting $\boldsymbol{n}_0^{\mathrm{ML}}$ and $\boldsymbol{n}_1^{\mathrm{ML}}$ in (2.11) and using the fact that a projection matrix is idempotent $P_{\boldsymbol{M}} = P_{\boldsymbol{M}}P_{\boldsymbol{M}}$ and symmetric $P_{\boldsymbol{M}}^\top = P_{\boldsymbol{M}}$ conducts to

$$\begin{aligned} T_{\mathrm{MSD}}(\boldsymbol{x}) &= \frac{1}{\sigma^2}\left(\boldsymbol{x}^\top P_{\boldsymbol{B}}^{\perp}\,\boldsymbol{x} - \boldsymbol{x}^\top P_{\boldsymbol{SB}}^{\perp}\,\boldsymbol{x}\right) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma \\ &= \frac{1}{\sigma^2}\left(\boldsymbol{x}^\top (P_{\boldsymbol{B}}^{\perp} - P_{\boldsymbol{SB}}^{\perp})\,\boldsymbol{x}\right) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma \\ T_{\mathrm{MSD}}(\boldsymbol{x}) &= \frac{1}{\sigma^2}\boldsymbol{x}^\top P_{\boldsymbol{B}}^{\perp}\,P_{\boldsymbol{G}}\,P_{\boldsymbol{B}}^{\perp}\,\boldsymbol{x} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma, \end{aligned} \tag{2.14}$$

where $\boldsymbol{G} = P_{\boldsymbol{B}}^{\perp}\,\boldsymbol{S}$, the orthogonal projection of $\boldsymbol{S}$ on the subspace orthogonal to $\boldsymbol{B}$. $T_{\mathrm{MSD}}(\boldsymbol{x})$

is in the form of a generalized energy detector. The energy $\|P_{\boldsymbol{G}}\, P_{\boldsymbol{B}}^{\perp}\, \boldsymbol{x}\|_2^2$: of $\boldsymbol{x}$ projected on $\boldsymbol{B}^{\perp}$ then on $\boldsymbol{S}$ then on $\boldsymbol{B}^{\perp}$. ■

The SMF (with $\Sigma = \mathbf{I}$) projects the test pixel $\boldsymbol{x}$ onto the direction of the target signal $\boldsymbol{s}$ to detect its presence in the noisy observation. In contrast, MSD projects the test pixel $\boldsymbol{x}$ on $\boldsymbol{B}^{\perp}$, then on $\boldsymbol{S}$ and again on $\boldsymbol{B}^{\perp}$. The ASD [Kraut *et al.* 2001] is a generalization of the MSD for unknown covariance matrices. Those are estimated from the data set.

2.2.2 Anomaly Detection

When there is no prior knowledge of the target signature, a standard approach in HSI is to use Anomaly Detection methods aimed at discriminating the pixels that are significantly *different* from the background exhibited by the scene. The benchmark algorithm in this category is the Reed-Xiaoli (RX) detector [Reed & Yu 1990]. The RX algorithm relies on GLR and it is a CFAR type detector, where a threshold is computed by fixing the probability of false alarm.

Example 2.2.3. Reed-Xiaoli anomaly detector.

To illustrate the RX detector, we assume that the background under $\mathcal{H}_0$ is modeled as a Gaussian distribution with known mean $\boldsymbol{\mu}_0$ and covariance $\boldsymbol{\Sigma}$ estimated from the background scene. Under $\mathcal{H}_1$, the signal has unknown mean $\boldsymbol{\mu}_1$ and the covariance is assumed known and equal to the one under $\mathcal{H}_0$

$$\begin{cases} \mathcal{H}_0 & : \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}), \\ \mathcal{H}_1 & : \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}), \end{cases} \tag{2.15}$$

where $\boldsymbol{x} \in \mathbb{R}^N$. The GLR for (2.15) is

$$T_{\mathrm{GLR}}(\boldsymbol{x}) := \frac{\max\limits_{\boldsymbol{\mu}_1} p(\boldsymbol{x}; \boldsymbol{\mu}_1)}{p(\boldsymbol{x}; \boldsymbol{\mu}_0)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma', \tag{2.16}$$

where γ' is a threshold. Maximizing the numerator of (2.16) gives $\boldsymbol{\mu}_1^{\mathrm{ML}} = \boldsymbol{x}$, and injecting this into the corresponding GLR (and taking the logarithm) yields

$$T_{\mathrm{RX}}(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{\mu}_0)^{\top} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_0) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma, \tag{2.17}$$

where $\gamma = 2 \ln \gamma'$ is a threshold that will be calculated at a fixed $\mathrm{P_{FA}}$. Under $\mathcal{H}_0$, $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$ conducts to $T_{\mathrm{RX}} \sim \chi^2_N$. Then, $\mathrm{P_{FA}} = \mathbb{P}(T_{\mathrm{RX}}(x) > \gamma; \mathcal{H}_0) = 1 - \Phi_{\chi^2_N}(\gamma)$. The corresponding threshold can be calculated according to a desired value of $\mathrm{P_{FA}}$

$$\gamma\,(\mathrm{P_{FA}}) = \Phi^{-1}_{\chi^2_N}(1 - \mathrm{P_{FA}}). \tag{2.18}$$

The probability of detection writes $\mathrm{P_{Det}} = \mathbb{P}(T_{\mathrm{RX}}(\boldsymbol{x}) > \gamma\,(\mathrm{P_{FA}}); \mathcal{H}_1)$. The obtained test statistic (2.17) shows that the RX anomaly detector measures the square of *Mahalanobis distance* [Mahalanobis 1936] between the active pixel $\boldsymbol{x}$ and the background mean distribution $\boldsymbol{\mu}_0$. It is thus an energy detector. Note that, if the background is assumed to be distributed according to the standard normal distribution (i.e, $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$), the test statistic (2.17) amounts to $T_{\mathrm{RX}}(\boldsymbol{x}) = \|\boldsymbol{x}\|_2^2 \gtrless_{\mathcal{H}_0}^{\mathcal{H}_1} \gamma$, which is a classical energy detector. ■

There are many variants of the RX test proposed in the literature. For instance, the kernel RX addresses a non linear model where the original model is projected in a new high dimensional feature space [Kwon & Nasrabadi 2005].

Other than the RX-based approaches, Anomaly detectors often rely on *discriminative* principles, like in the Support Vector Data Description (SVDD) introduced by Tax and Duin [Tax & Duin 2004]. This algorithm aims at finding the minimum volume of a closed boundary sphere containing all data pixels. Banerjee *et al.* proposed for instance, an Anomaly detector based on SVDD [Banerjee *et al.* 2006].

In the paper [Matteoli *et al.* 2007], the authors compare several Anomaly detection strategies for the detection of man-made targets (e.g., car, building) in rural scenes. They concluded that Anomaly detection method based on Orthogonal Subspace Projection (operating with other methods: SVD, highest *kurtosis*[1] and local RX) yielded the best performance, in term of background suppression. This technique identifies the orthogonal subspace w.r.t. the background, and projects the pixel vectors onto this subspace (thus eliminating the background from the target).

2.3 Discussion

This Chapter first introduced briefly the concept of HSI. The main focus was then on the target detection techniques in HSI, which can be categorized in two types, namely Spectral Matching and Anomaly detection. Spectral Matching methods benefit from the known reliable knowledge of the target spectra, hence they have larger detection power than the Anomaly detectors. As we will see, the detection techniques considered in this dissertation (described in Part II), may be seen as particular (i.e. robust, or minimax) cases of subspace Spectral Matching.

When prior knowledge is encapsulated in a spectral data set, and that this data set is large, detection tests traditionally operate in subspaces of reduced dimensions [Ardekani *et al.* 1999, Fernandes 2010] which can be obtained by several techniques. In the framework of this dissertation, we are interested in such data sets (known and high dimensional) and in reduction dimension methods. Recent dimension reduction techniques are based on the concept of *learning* from the samples. These topics, among others, will be studied in the next Chapter.

[1] Kurtosis is a measure of "peakedness" and "tailedness" of a PDF [Pearson 1905, DeCarlo 1997], often w.r.t. the normal distribution (whose kurtosis is $\mu_4/\sigma^4 = 3$, and its excess kurtosis is $\mu_4/\sigma^4 - 3 = 0$. μ_4 denotes the fourth central moment and σ is the standard deviation).

Chapter 3

Dimension reduction and sparse representations

Contents

3.1 Introduction

3.1.1 Dimension reduction

Representing the complexity of data by means of a reduced but significant number of random variables or parameters is the main objective of *dimension reduction*. This procedure is particularly important in the modern age where most systems produce high dimensional data, making the data processing and analysis very complex or even impossible.

In many fields such as astrophysics, medical, or information theory, the redundancy of the data set allows an efficient dimension reduction. This is because the information of interest always presents a particular statistical *structure* that makes it differentiable from pure stochastic fluctuations (i.e., from the "noise"). This characteristic is called *compressibility* in source coding.

For future use, let us assume that we have a known large data set $\boldsymbol{S} \in \mathbb{R}^{N\times L}$, whose column vectors $\boldsymbol{s}_i,\ i = \{1,\dots,L\}$ are the information signals (a.k.a. *features*). In this context, $\boldsymbol{S}$ is a reference *library* (or *training samples*) and dimension reduction on $\boldsymbol{S}$ consists of representing it in lower dimension by statistical means, for instance via *sparse learning*

algorithms. The obtained subspace, say the subspace spanned by matrix $\boldsymbol{D} \in \mathbb{R}^{N \times K}$, $K < L$, is a new mathematical representation of the reference library $\boldsymbol{S}$.

The well-known Principal Component Analysis (PCA) method [Pearson 1901] is an example of statistical dimension reduction. PCA is widely used for linear dimension reduction because of its simplicity and efficiency. It involves an orthogonal transformation of the (possibly) correlated samples of library $\boldsymbol{S}$ into the same or smaller dimension of uncorrelated atoms, namely the *principal components*. PCA relies on *eigen* analysis, where the first principal component is associated to the largest eigenvalue, the second principal component is associated to the second largest eigenvalue and so on. Such computations can be done through the Singular Value Decomposition (SVD) of the library $\boldsymbol{S}$ (see Section 3.2).

We present next (in Sections 3.1.2 - 3.1.4) three families of approaches to dimension reduction: classification and clustering, sparse learning, and other methods.

3.1.2 Classification and Clustering

When a data set is partitioned into a set of groups, there are two approaches used. The first is called Classification, often used in supervised learning. The second approach is Clustering (in an unsupervised learning context), where a set of available data is sorted into K groups, each sample belonging to the same group (also called *cluster*) presents the same degree of similarity w.r.t. a specific criterion.

Classification relies on two sets of data. One set termed *training* data is known. It is used to obtain the learned atoms (which will be the representatives of specific classes) or to find representative parameters through an optimization problem. The second set is a *new* data set, called *test* data. When the new data arrives, it will be *classified* (i.e., assigned) according to the learned atoms or the computed parameters. The algorithms of this type aim to minimize the probability of misclassification.

In clustering, data set is partitioned into K-clusters based on some distance function, and clusters' representatives (termed *centroids*) are computed. The distance function has to be defined between each samples, and for some algorithms, it is also defined between clusters. For instance, this distance function can be the conventional Euclidean distance, or it can be a correlation function. Clustering can be divided in two categories: Nearest Neighbor (NN) clustering and Hierarchical clustering.

i. NN clustering first starts by fixing the number of clusters, then the samples are partitioned into different clusters w.r.t. a distance criterion. The canonical clustering algorithm is the K-Means algorithm [Lyold 1982, MacQueen 1967], which uses the *nearest mean* (i.e., distance of data samples to their corresponding cluster centroids computed as the intra cluster means) as the distance metric (the pseudo-code of K-means algorithm is provided in Appendix A.5.1).

ii. Hierarchical clustering can be divided into two types: Agglomerative and Divisive clusterings. Agglomerative clustering starts with one sample representing one cluster (thus, for L samples, we have L clusters at initialization). Then the closest (as

defined by a distance function) pair of clusters are merged into one cluster. This is done successively, until obtaining the desired number of clusters (e.g., [Gowda & Krishna 1978] proposed a hierarchical procedure using the nearest Euclidean distance between clusters, and the stopping rule is based on the number of distinct patterns they observed from the data set). Agglomerative clustering is often presented by tree diagrams (dendrograms), also known as "bottom up" approach. In contrast, the Divisive clustering is a "top down" approach, where all data samples are grouped in a single cluster at initialization, then the samples are partitioned recursively into many clusters, towards L clusters. The algorithm stops before obtaining L clusters.

3.1.3 Sparse learning

Sparse learning techniques deal with the approximation of $\boldsymbol{S}$ by linear combination of few vectors. The matrix collecting all the learned vectors (also named atoms) is called *dictionary*, which is associated to a sparse representation vector, or matrix.

In regard to Section 3.1.2, sparse learning can be viewed as subspace clustering where the centroids represent the atoms of the learned dictionary.

Sparsity of a signal refers to the number of non-zero elements it contains, e.g., the sparsity of $\boldsymbol{y} = [0.8, 0, 0, 1, 0, 0.9, 0]^\top$ is three. Mathematically, sparsity is expressed by the ℓ_0 pseudo-norm: $\|\boldsymbol{y}\|_0 = \#\{n : y_n \neq 0\}$.

If the dictionary is assumed known, we can find the sparse representation vector(s) using sparse coding algorithms such as the Matching Pursuit (MP) [Mallat & Zhang 1993], the Orthogonal Matching Pursuit (OMP) [Mallat *et al.* 1994, Pati *et al.* 1993] or the Least Absolute Shrinkage and Selection Operator (LASSO) [Tibshirani 1994].

If the dictionary is unknown, we turn to sparse dictionary learning techniques. One of the earliest work in dictionary learning is called the Method of Optimal Directions (MOD) [Engan *et al.* 1999] shown in Appendix A.5.2. A reference algorithm in this category is called K-SVD [Aharon *et al.* 2006] (see Appendix A.5.3 for an example), which is a generalization of the K-means method.

The *Compressed Sensing* theory and technique were introduced in 2004 which exploit the naturally sparse signal present in specific applications and rely on sparse coding algorithm to recover unknown signals [Donoho 2004a, Emmanuel *et al.* 2004]. One of its most successful application is in medical imaging, where the costs in term of time and price are reduced significantly through the the small number of measurements. This is particularly a major advantage when treating young patients that cannot stay still for a long duration (e.g., for an MRI[1] examination).

On other note, around 2007, Elad, Milanfar and Rubinstein were among the first who highlighted two different models in sparse regression problems: synthesis and analysis [Elad *et al.* 2007]. We focus on the synthesis model in our framework. Nevertheless, the analysis sparse model is described in Appendix A.4.

[1] MRI: Magnetic Resonance Imaging

3.1.4 Other methods

The third approach to dimensionality reduction is based on the projection of reference library onto *interesting* directions (as defined by each method), or onto manifolds that best represent the samples. Most of the approaches in this category are non linear. A classical non linear dimension reduction method is termed Sammon projection [Sammon 1969] and involves a gradient descent algorithm to map high dimensional space onto lower dimensional space. As another example, the Locally Linear Embedding technique [Roweis & Saul 2000] uses linear weighted coefficients and seek to maintain the neighborhood structure of the data in the ambient space.

The following Sections of this Chapter present further explanations on dimensionality reduction approaches, emphasizing the sparse learning approach in the search of *data-driven* dictionaries.

3.2 Low rank matrix approximation

Dimensionality reduction exploits the compressibility of the considered data set $\boldsymbol{S}$. A mathematical procedure often used for data reduction is the low rank matrix approximation (through the Eckart-Young theorem, [Eckart & Young 1936]), which can be formulated by the constrained optimization problem

$$\min_{\widehat{\boldsymbol{S}}} \|\boldsymbol{S} - \widehat{\boldsymbol{S}}\|_F^2 \quad \text{subject to } \operatorname{rank}(\widehat{\boldsymbol{S}}) = \operatorname{rank}(\boldsymbol{S}) \tag{3.1}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. This problem can be solved through SVD (of rank $\widehat{\boldsymbol{S}}$, say r). The approximated matrix is

$$\widehat{\boldsymbol{S}} = \mathbf{U}\boldsymbol{\Sigma}_r\mathbf{V}^\top, \tag{3.2}$$

where $\boldsymbol{\Sigma}_r \in \mathbb{R}^{N\times L}$ is a diagonal matrix containing the r largest singular value of $\boldsymbol{S}$. $\mathbf{U} \in \mathbb{R}^{N\times N}$ and $\mathbf{V} \in \mathbb{R}^{L\times L}$ are respectively left-singular and right-singular matrices. $\boldsymbol{\Sigma}_r\mathbf{V}^\top$ is a sparse matrix (by rows, see Figure 3.1) and $\mathbf{U}_r$ (gray columns of $\mathbf{U}$ in the third subfigure of Figure 3.1) is often used in dimension reduction techniques to represent $\boldsymbol{S}$ in lower dimension.

As we will see, the approximation by SVD tends to produce atoms (columns of $\mathbf{U}_r$ here) that capture some *average behavior* of the data set (like most optimization method based on minimizing the MSE). Indeed, it is well known that ℓ_2 norm is sensitive to outliers (the large distance of these outliers being magnified by the ℓ_2 norm). Hence, it seems a bit paradoxical that this criterion represents "well" the average behavior of the data samples. In fact, MSE estimates are *perturbed* by outliers, but not explicitly seek to represent well such samples that are far from the mean. A minimax dimension reduction does this. This observation is an important point of Part II (e.g, see Section 4.3, Examples 5.2.1 and 6.3). In Part III, we will compare, among others, SVD based algorithms with the proposed minimax approaches.

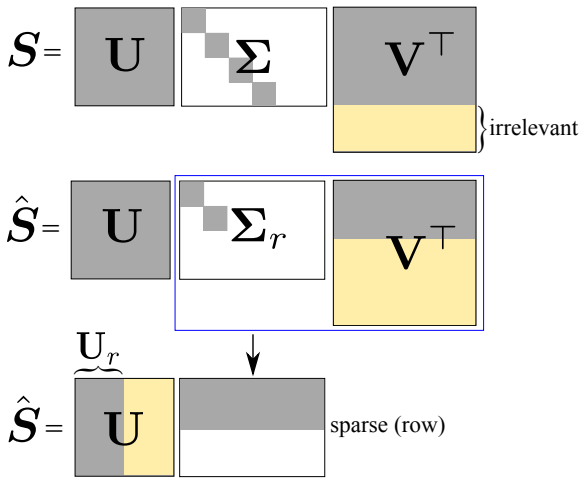


Figure 3.1: Singular Value Decomposition of matrix $\boldsymbol{S} \in \mathbb{R}^{N\times L}$ and its low rank approximation $\widehat{\boldsymbol{S}}$ of the same dimension. White areas represent zero elements and yellow areas indicate *irrelevant* values. The first subfigure illustrates the decomposed matrices $\mathbf{U} \in \mathbb{R}^{N\times N}, \boldsymbol{\Sigma} \in \mathbb{R}^{N\times L}$(diagonal matrix) and $\mathbf{V}^\top \in \mathbb{R}^{L\times L}$ obtained from SVD of $\boldsymbol{S}$. The rows $N+1$ to L of the matrix $\mathbf{V}^\top$ are irrelevant w.r.t. $\boldsymbol{S}$. The second subfigure depicts low rank approximation of $\boldsymbol{S}$, where $\boldsymbol{\Sigma}_r$ contains the r largest singular value of $\boldsymbol{S}$ with the others set to zero. The last subfigure shows that $\boldsymbol{\Sigma}_r\mathbf{V}^\top$ is a (row) sparse matrix. $\mathbf{U}_r$ contains the r atoms representing $\boldsymbol{S}$ in lower dimension.

3.3 A glance at sparsity promoting method in signal processing

3.3.1 Basic model for the sparse representation of vectors

The concept of sparse modeling originated from the problem of recovering a high dimensional signal from a low dimensional signal [Chen *et al.* 1998]. In a viewpoint of linear algebra, it is equivalent to seeking a solution of an underdetermined linear system where the number of unknown variables exceeds the number of equations.

Assume that an informative signal $\boldsymbol{s} \in \mathbb{R}^N$ admits linear relation

$$\boldsymbol{s} = \boldsymbol{D}\boldsymbol{\alpha}, \tag{3.3}$$

where the dictionary $\boldsymbol{D} \in \mathbb{R}^{N\times K}$ is a full rank matrix (typically, $K > N$) and $\boldsymbol{\alpha} \in \mathbb{R}^K$ is an unknown sparse vector. In this setting, $\boldsymbol{D}$ is *fixed* or assumed *known*. Sparse approximation (also termed sparse coding) typically approximates $\boldsymbol{s}$ by the following optimization problem

$$\widehat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}\|_0 \quad \text{subject to } \|\boldsymbol{s} - \boldsymbol{D}\boldsymbol{\alpha}\|_2^2 \le \varepsilon, \tag{3.4}$$

where ε is an error threshold. According to (3.4), we are searching the *sparsest* solution of an underdetermined system. By this, we expect that the sparsity of $\|\boldsymbol{\alpha}\|_0 := k$ is inferior to the dimension N of $\boldsymbol{s}$. In this setting, the estimated vector $\widehat{\boldsymbol{\alpha}}$ is called the *sparse representation*

of signal $\boldsymbol{s}$.

Once we obtain $\widehat{\boldsymbol{\alpha}}$, the signal of interest can be approximated by: $\widehat{\boldsymbol{s}} = \boldsymbol{D}\widehat{\boldsymbol{\alpha}}$. Figure 3.2 illustrates the approximation, where $\widehat{\boldsymbol{\alpha}}$ points out the k-columns of $\boldsymbol{D}$ that best represent $\boldsymbol{s}$ (once weighted by the non-zero amplitudes of $\boldsymbol{\alpha}$). In other words, the signal $\boldsymbol{s}$ can be approximated by linear combination of *few* atoms. Often, this is a good approximate model of initial signals for appropriately chosen $\boldsymbol{D}$.

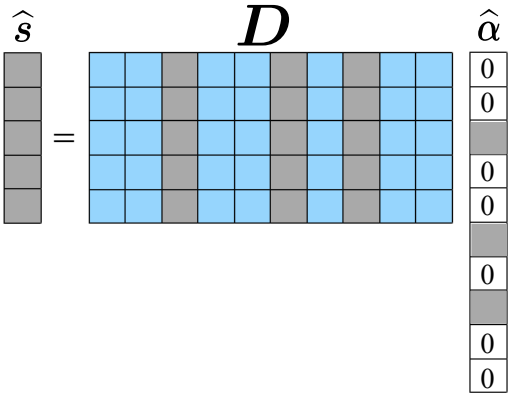


Figure 3.2: The concept of synthesis sparse modeling. The signal $\boldsymbol{s} \in \mathbb{R}^N$ can be approximated by linear combination of few atoms, where $\boldsymbol{D} \in \mathbb{R}^{N \times K}$ is a known dictionary and $\widehat{\boldsymbol{\alpha}} \in \mathbb{R}^K$ is the sparse representation vector. In this example, sparsity of $\widehat{\boldsymbol{\alpha}}$ is $\|\boldsymbol{\alpha}\|_0 = 3$.

3.3.2 Basic approaches for sparse approximation

The optimization problem (3.4) is however combinatorial. In 1993, Mallat and Zhang set a stepping stone in the sparse representation field [Mallat & Zhang 1993]. They proposed a *greedy* method, named Matching Pursuit (MP) to solve approximately this ℓ_0 optimization problem. MP consists of finding a good (if not best) matching projection on the known dictionary $\boldsymbol{D}$, in the sense that the resulting estimate of $\boldsymbol{s}$ is close to yield the smallest possible representation error for a given k. In this setting, representation error is defined by : $\boldsymbol{e} = \boldsymbol{s} - \boldsymbol{D}\widehat{\boldsymbol{\alpha}}$. MP identifies exploratory atoms $\boldsymbol{d}_j$ (columns of $\boldsymbol{D}$) by projection of the current representation error. However, MP only projects the representation residual onto the last identified atom. This decreases the ability of the algorithm to efficiently capture a large part of the signal's energy in the subspace generated by the identified atoms. As a remedy, an extension termed Orthogonal Matching Pursuit (OMP) has been proposed [Mallat *et al.* 1994,Pati *et al.* 1993], where the residual is orthogonalized w.r.t. all previously selected atoms. This assures that the same atom will not be selected again. Many other greedy sparse coding algorithms exist (e.g., SP [Dai & Milenkovic 2009], IHT [Blumensath *et al.* 2007], CoSaMP [Needell & Tropp 2009]).

In a global (instead of greedy) approach, the ℓ_0 pseudo-norm can be relaxed to ℓ_1 norm, such as the problem [Donoho 2004b]

$$\widehat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}\|_1 \quad \text{subject to } \|\boldsymbol{s} - \boldsymbol{D}\boldsymbol{\alpha}\|_2^2 \leq \varepsilon, \tag{3.5}$$

or through ℓ_1 norm regularizer

$$\widehat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \|\boldsymbol{s} - \boldsymbol{D}\boldsymbol{\alpha}\|_2^2 + \mu_1 \|\boldsymbol{\alpha}\|_1, \tag{3.6}$$

a problem termed LASSO [Tibshirani 1994]. ℓ_1 promotes strict sparsity [Donoho 2004b] and benefits from convexity properties.

Problem (3.6) can more generally be written in ℓ_p norm, yielding a regularization term $\mu_p \|\boldsymbol{\alpha}\|_p^p$, where $0 < p \leq 1$. For $p < 1$, the regularization problem maintains its sparsity promoting property, but it is more difficult to solve because of the non-convexity of the objective function.

3.3.3 Basic model for the sparse representation of matrices

If we have many column vectors $\boldsymbol{s}_i$ belonging to a large matrix $\boldsymbol{S} \in \mathbb{R}^{N \times L}$, relation (3.3) becomes

$$\boldsymbol{S} = \boldsymbol{D}\boldsymbol{A}, \tag{3.7}$$

where $\boldsymbol{D} \in \mathbb{R}^{N \times K}$ is a known or fixed dictionary and $\boldsymbol{A} \in \mathbb{R}^{K \times L}$, $(K \leq L)$ is the unknown representation matrix to be estimated. The sparse approximation problem corresponding to (3.4) is

$$\forall i,\ \widehat{\boldsymbol{\alpha}_i} = \arg\min_{\boldsymbol{\alpha}_i} \|\boldsymbol{\alpha}_i\|_0 \quad \text{subject to } \|\boldsymbol{S} - \boldsymbol{D}\boldsymbol{A}\|_F^2 \leq \varepsilon, \tag{3.8}$$

where $\boldsymbol{\alpha}_i$ are the columns of the sparse representation matrix $\boldsymbol{A}$. We can indeed use sparse coding algorithms such as MP or OMP to estimate each sparse vectors $\widehat{\boldsymbol{\alpha}_i}$.

In a different problem, we may be interested in optimizing on $\boldsymbol{D}$ such that the error $\boldsymbol{E} = \boldsymbol{S} - \boldsymbol{D}\boldsymbol{A}$ is minimized (in Frobenius norm) for fixed $\boldsymbol{A}$. This is called dictionary update, see Section 3.4.

In view of the previous discussion, an efficient sparse approximation is possible when we have a "good" dictionary $\boldsymbol{D}$. Image and audio processing communities have been using some generic dictionaries to reconstruct signals (such as the wavelets (localized in time-frequency) [Mallat 2008], and Discrete Cosine Transform (represent oscillations)). Although these pre-defined dictionaries work well, dictionary *adapted* to specific data sets are sometimes more appealing and efficient in some applications. Learned dictionaries may outperform generic dictionaries for image denoising [Aharon *et al.* 2006, Elad & Aharon 2006] and be very efficient for other tasks such as blind source separation [Abolghasemi *et al.* 2012] or classification for object recognition [Zhang & Li 2010, Kong & Wang 2012]. The following Section 3.4 presents the principles underlying sparse dictionary learning techniques, allowing to obtain optimized, data-driven, dictionaries.

3.4 Sparse dictionary learning algorithms

In contrast to Section 3.2, where the dictionary was optimized using SVD, sparse learning methods incorporate in the dictionary learning problem sparsity promoting criteria and build an optimized dictionary from the samples. Here, we seek both the representation vector or matrix, and the dictionary.

In the synthesis sparse linear problem presented in Section 3.3, $\boldsymbol{S} = \boldsymbol{D}\boldsymbol{A}$, the dictionary is assumed known. For sparse learning, we search two variables, the dictionary $\boldsymbol{D}$ and the representation matrix $\boldsymbol{A}$, where each columns $\boldsymbol{\alpha}_i$, $i = 1, \ldots, L$ are k sparse. Extending the problem (3.8) by imposing constraint on $\boldsymbol{A}$, it now becomes a joint optimization problem

$$\min_{\boldsymbol{D},\boldsymbol{A}} \|\boldsymbol{S} - \boldsymbol{D}\boldsymbol{A}\|_F^2 \quad \text{subject to } \forall i = 1, \ldots, L,\ \|\boldsymbol{\alpha}_i\|_0 \leq k. \tag{3.9}$$

With regard to (3.9), sparse learning techniques most often alternate between a sparse coding step and a dictionary update step. In the sparse coding step, the dictionary $\boldsymbol{D}$ is fixed and the unknown representation matrix, $\boldsymbol{A}$, is calculated. In the second stage, the dictionary is updated. In many dictionary learning algorithms, the columns of $\boldsymbol{D}$ are normalized (i.e., $\|\boldsymbol{d}_j\|_2^2 = 1$, $j = 1, \ldots, k$).

The idea of sparse learning originated back in 1996, when Olshausen and Field searched an answer on how to determine a model that best describes the population of simple cells in the primary visual cortex [Olshausen & Field 1996]. According to them, a classical learning method such as PCA is not suitable in describing natural images (e.g., containing curves and edges). Since most of natural images have a sparse structure, they exploited this characteristic. Their investigations were fruitful as they obtained learned dictionaries that possess similar properties resembling the simple cells, suggesting that optical cells are in some sense sensitive to the sparse structure contained in the images.

Following this work, many other algorithms were proposed such as the Method of Optimal Directions [Engan *et al.* 1999] (see pseudo-code in Appendix A.5.2) or the well-known K-SVD algorithm described in Appendix A.5.3. The K-SVD algorithm will be used as comparison to the proposed approaches in the following Parts II and III of this dissertation.

Both of the mentioned algorithms, MOD and K-SVD proceed in two-steps dictionary learning approach. The sparse coding stage is the same: the sparse representation matrix $\boldsymbol{A}$ is computed through pursuit algorithms. However, in K-SVD, the corresponding values of these sparse coefficients are also updated in the dictionary update stage through the decomposition by SVD. As another difference between these two algorithms, MOD uses the general solution to least squares problem to learn its dictionary (i.e., all columns are computed simultaneously). In contrast, K-SVD updates its dictionary one atom at a time by the best rank one approximation (in Frobenius norm) of the restricted residual data $\boldsymbol{E}^R$ (and simultaneously updates the non-zero coefficients in $\boldsymbol{A}$). MOD is an effective learning method, but when the known data is large, it induces a large computation complexity (due to the matrix inversion in the dictionary update stage).

3.5 Discussion

We have reviewed in this Chapter various basic aspects related to dimensionality reduction, in particular clustering and low-rank sparse dictionary learning techniques. The emphasis on sparse learning techniques is related to the principal application of our research work in detection, where under the alternative hypothesis only one signal (in the form of a spectral emission line) from a known library can be activated in the data under the test. In addition, some spectral lines can be very atypical (see Chapter 7), posing the question of a criterion allowing robust detection of such profiles.

In the next Chapter, numerical simulations will show that conventional learning algorithms (based on SVD) do not perform well w.r.t. minimax objectives in a GLR test. These observations will lead to specific optimization issues and will call for dedicated learning algorithms investigated in Chapter 5.

Part II

Subspace Learning in Minimax Detection: proposed methods

The second Part of this dissertation investigates the problem of detecting a target signal belonging to a known and possibly large library $\boldsymbol{S} \in \mathbb{R}^{N \times L} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_L]$. In the considered framework, we will assume that under the alternative $\mathcal{H}_1$, only one $\boldsymbol{s}_i \in \boldsymbol{S}$ can be activated in the test data under the alternative, with some amplitude α, in the presence of known Gaussian noise $\boldsymbol{n} \sim \mathcal{N}(\boldsymbol{0}, \mathbf{I})$. The index i and amplitude α are however both unknown.

This problem can be modeled under $\mathcal{H}_1$ as: $\boldsymbol{x} = \boldsymbol{\alpha}\boldsymbol{S} + \boldsymbol{n}$ with a sparsity constraint $\|\boldsymbol{\alpha}\|_0 = 1$ (imposing that only one $\boldsymbol{s}_i$ is activated). We call this the "exact" model, because it precisely corresponds to our assumptions.

Some of the target signals can however be very atypical w.r.t. the others, in the sense that their geometry may be quite different from the prototyped geometry of the targets $\boldsymbol{S}$. This poses the question of the *robustness* of the detection scheme w.r.t. such profiles.

The specific application that motivates the study of this problematic concerns the detection of faint spectral line of primordial galaxies (also known as the Lyman-α emitters) in HSI data. In the data, each spectral line appears, if present, shifted in wavelengths by an *a priori* unknown amount (which depends on the galaxies' distances), leading to a very large library of possible features under $\mathcal{H}_1$. Other illustration of this problem occur for other large sets of target spectra, e.g., for rare minerals detection or for identification of infected trees.

Chapter 4 examines detection performances based on the 1-sparsity-constrained GLR using the exact model above. A first and immediate issue of this approach when operating with large sets of alternatives is the computation complexity. In this regard, we also study a GLR test using a toy reduced dimension method (SVD) but computationally low and we compare it with the test using the exact model. The investigations' results show that reducing the dimension of the test may lead to a substantial loss of detection power for some (atypical) alternatives. We will see that this effect should not be attributed to the overly simplistic dimension reduction used, but to the very principles used in traditional dimension reduction techniques. This conducts us to propose reduced models w.r.t. a *robust* criterion, as presented in the succeeding Chapter.

We propose in **Chapter 5** reduced dimension detection tests based on the GLR and aimed at maximizing the detection performance in the worst-case scenarios occurring under $\mathcal{H}_1$. The dimensionality reduction is performed by learning from $\boldsymbol{S}$ a low dimension dictionary $\boldsymbol{D}^* \in \mathbb{R}^{N \times K}, (K < L)$. The *minimax* criterion is set up to address the problem of maintaining as much power as possible for the detection of atypical signatures, while reducing the computational complexity of the test w.r.t. a 1-sparse GLR over $\boldsymbol{S}$. The considered reduced dimension model used to implement the test imposes a sparsity constraint on the unknown amplitude vector, similarly as for the exact model.[2]

[2] As a separate research axis, we also propose in Section 5.3 an alternative reduced dimension test, which does not promote sparsity. However, for an arbitrary size dictionary ($K > 1$), the optimization problem for this second reduced model is non-convex. Owing to the satisfying results obtained with the approach proposed in Section 5.2.4, we have not pursued in this direction (by trying to find suboptimal algorithms in the spirit of those of Section 5.3.

Following the analytical analysis of Chapter 5, the succeeding **Chapter 6** describes in detail the proposed minimax learning algorithms. The learned dictionaries $\boldsymbol{D}^*$ are expected to capture all shapes of signatures in $\boldsymbol{S}$, particularly the atypical ones (as those generally induce the worst-case scenarios). The first approach is a greedy type learning algorithm based on the analysis of Section 5.2.4. We call this algorithm "greedy minimax". The second approach arises from the general strategy of injecting minimax objectives in standard dictionary learning algorithms. In this regard, we propose a variant of K-SVD algorithm, where the dictionary update stage is replaced by the exact one-dimensional minimax solution found in Chapter 5. We name it K-minimax algorithm. The last algorithm combines clustering techniques found in the literature (e.g., Spherical K-Means (SKM)) with the exact one-dimensional minimax solution.

The proposed greedy minimax and K-minimax algorithms constitute the most interesting detection tests used for the application to astrophysical data in Part III (Chapter 7).

Chapter 4

Detection test for the exact model with sparsity-constrained

Contents

Some analyses and results presented in this Chapter were published in [Suleiman *et al.* 2013a, Suleiman *et al.* 2014a].

4.1 Introduction

We have seen in Part I the basic aspects of various topics related to our research work in detection. This chapter introduces the principal detection problem that we are dealing with: assume $\boldsymbol{S} \in \mathbb{R}^{N\times L} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_L]$ is a known reference library of L alternatives. Assume further that under the alternative $\mathcal{H}_1$, only one target signature $\boldsymbol{s}_i \in \boldsymbol{S}$ is activated. The amplitude of the active signal and its index i (location in $\boldsymbol{S}$) are however considered unknown.

The number of possible target signatures L can be arbitrary, ranging from small sets (i.e., in the tens or hundreds, in telecommunications symbols for instance [Sklar 2011]) to very large sets (in the hundred of thousands or more; e.g., for samples drawn from numerical models [Berk *et al.* 2005]). In this dissertation, we focus on the most interesting setting where L is very large (but always fixed and *finite*).

In some applications like in HSI, the case where L is very large systematically arises when there are few known target signatures, but those are registered with systematic disturbances than can be modeled and sampled [Berk *et al.* 2005]. This leads to a possibly huge library $\boldsymbol{S}$. In other applications, the unknown target signal can have arbitrary variations. In such cases, numerical simulations can provide numerous possible templates of the target signal yielding again a large reference library. We will see an example of such case in Chapter 7 when dealing with the detection of spectral lines in an astrophysical application.

When facing with this specific detection problem, a straightforward approach is to test with a GLR all the possible alternatives in $\boldsymbol{S}$ (what we call below using the exact model). Consequently, we investigate in this Chapter the detection performances of such a test. As previously mentioned, the framework of this dissertation is restricted to the GLR approach.

In addition, to decrease the computation complexity of using the full library in the detection approach, we also examine in this Chapter the effects of a test which uses a very simple reduced dimension method (SVD). This toy test allows nevertheless to exhibit important effects related to dimension reduction, in particular regarding the power loss that may affect some alternatives.

4.2 Exact detection model and associated GLR test

The detection problem discussed in Introduction can be written using the following composite hypotheses

$$\begin{cases} \mathcal{H}_0 & : \boldsymbol{x} = \boldsymbol{n}, & \boldsymbol{n} \sim \mathcal{N}(\boldsymbol{0}, \mathbf{I}) \\ \mathcal{H}_1 & : \boldsymbol{x} = \boldsymbol{S}\boldsymbol{\alpha} + \boldsymbol{n}, & \|\boldsymbol{\alpha}\|_0 = 1 \end{cases}, \tag{4.1}$$

where $\boldsymbol{x}$ and $\boldsymbol{n} \in \mathbb{R}^N$. Throughout the second Part of this dissertation, the corresponding composite hypotheses models assume no (or perfectly subtracted) background, and a covariance matrix that is known and equal under both hypotheses. In the case of a correlated noise model with covariance matrix $\boldsymbol{\Sigma}$, i.e., $\mathcal{H}_1 : \boldsymbol{x} = \boldsymbol{S}\boldsymbol{\alpha} + \boldsymbol{n}$, $\boldsymbol{n} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, "whitening" the data by $\boldsymbol{\Sigma}^{-\frac{1}{2}}$ indeed yields a model of the form (4.1).

Without loss of generality, the columns of the known library $\boldsymbol{S} \in \mathbb{R}^{N \times L} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_L]$ are normalized[1] : $\|\boldsymbol{s}_i\|_2^2 = 1$, $\{i = 1, \ldots, L\}$. By this, we consider that the finite set $\mathcal{S} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_L\}$ is a collection of points on the unit sphere as a function of their shapes, all having the same energy. This normalization is due to the unknown amplitude α.

The 1-sparse constraint on the unknown vector $\boldsymbol{\alpha} \in \mathbb{R}^L$ imposes that under $\mathcal{H}_1$, only one signal $\boldsymbol{s}_i$ of unknown amplitude α_i is activated. The SNR of $\boldsymbol{x}$ is then controlled by the magnitude of the amplitude α under $\mathcal{H}_1$: $\mathrm{SNR}_{(\mathrm{dB})} = 10 \log_{10} \frac{\alpha^2 \|\boldsymbol{s}\|}{\sigma^2} = 10 \log_{10} (\alpha^2)$. Thus, we compare the detectability of alternatives as a function of their shapes, all having the same energy.

Model 4.1 specifies no distribution under $\mathcal{H}_1$, i.e., we do not specify a discrete distribution reflecting the probability of activation of each possible alternative under $\mathcal{H}_1$; all alternatives are considered equally likely to be activated. As a final remark, in our framework, the dimensions of $\boldsymbol{S}$, set by N and L, together with the SNR, are fixed and finite numbers.

The constrained GLR for model (4.1) is

$$T_{\mathrm{GLR}}(\boldsymbol{x}, \boldsymbol{S}) = \max_{\boldsymbol{\alpha}:\|\boldsymbol{\alpha}\|_0=1} \frac{p(\boldsymbol{x}; \boldsymbol{S}\boldsymbol{\alpha})}{p(\boldsymbol{x}; \boldsymbol{0})} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma', \tag{4.2}$$

[1] If the set $\boldsymbol{s}_i$, $i = 1, \ldots, L$ are not normalized, we can obtain (4.1) as follows. Assume $\boldsymbol{S}^0$ is a non normalized matrix. The normalization writes $\boldsymbol{S}^0 = \boldsymbol{S}\boldsymbol{H}$, where $\boldsymbol{H} \in \mathbb{R}^{L \times L}$ is a diagonal matrix, containing the inverses of the norms of the vectors of $\boldsymbol{S}^0$.
Then under $\mathcal{H}_1$: $\boldsymbol{x} = \boldsymbol{S}^0\boldsymbol{\beta} + \boldsymbol{n}$, $\|\boldsymbol{\beta}\|_0 = 1 \Leftrightarrow \boldsymbol{x} = \boldsymbol{S}\underbrace{\boldsymbol{H}\boldsymbol{\beta}}_{\boldsymbol{\alpha}} + \boldsymbol{n} = \boldsymbol{S}\boldsymbol{\alpha} + \boldsymbol{n}$, $\|\boldsymbol{\alpha}\|_0 = 1$ as in (4.1).

where γ' is a threshold. Under $\mathcal{H}_1$

$$p(\boldsymbol{x}\,;\boldsymbol{S}\boldsymbol{\alpha}) = \frac{1}{(2\pi)^{N/2}} \exp\left(-\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{S}\boldsymbol{\alpha}\|_2^2\right). \tag{4.3}$$

We shall seek the Maximum Likelihood estimate of $\boldsymbol{\alpha}$. The maximization has to be performed over the index i, $\{i = 1,\ldots,L\}$ of the non-zero component, and over the corresponding value α_i. Maximizing the numerator of (4.3) conducts to[2]

$$\boldsymbol{\alpha}^{\mathrm{ML}} = \underset{\boldsymbol{\alpha}:\|\boldsymbol{\alpha}\|_0=1}{\arg\min}\ \|\boldsymbol{x}-\boldsymbol{S}\boldsymbol{\alpha}\|_2^2, \tag{4.4}$$

and fixing i implies

$$\alpha_i^{\mathrm{ML}} = \underset{\alpha_i:\|\alpha_i\|_0=1}{\arg\min}\ (\|\boldsymbol{x}\|_2^2 - 2\alpha_i \boldsymbol{s}_i^\top \boldsymbol{x} + \alpha_i^2\,\|\boldsymbol{s}_i\|_2^2), \tag{4.5}$$

where $\|\boldsymbol{s}_i\|_2^2 = 1$ for $\{i = 1,\ldots,L\}$, which leads to

$$\frac{\partial(-2\alpha_i^\top \boldsymbol{s}_i^\top \boldsymbol{x} + \alpha_i^2)}{\partial \alpha_i} = 0 \Rightarrow\ \alpha_i^{\mathrm{ML}} = \boldsymbol{s}_i^\top \boldsymbol{x}. \tag{4.6}$$

Maximizing over the index i gives

$$\begin{aligned}
\widehat{i} &= \underset{i=1,\ldots,L}{\arg\min}\ (\|\boldsymbol{x}\|_2^2 - 2\alpha_i^{\mathrm{ML}} \boldsymbol{s}_i^\top \boldsymbol{x} + (\alpha_i^{\mathrm{ML}})^2),\\
&= \underset{i=1,\ldots,L}{\arg\min}\ (\|\boldsymbol{x}\|_2^2 - 2(\boldsymbol{s}_i^\top \boldsymbol{x})^2 + (\boldsymbol{s}_i^\top \boldsymbol{x})^2),\\
\widehat{i} &= \underset{i=1,\ldots,L}{\arg\max}\ (\boldsymbol{s}_i^\top \boldsymbol{x})^2.
\end{aligned} \tag{4.7}$$

The non-zero element of the constrained ML estimate of $\boldsymbol{\alpha}$ is $\alpha_{\widehat{i}}^{\mathrm{ML}} = \boldsymbol{s}_{\widehat{i}}^\top \boldsymbol{x}$. Taking the logarithm of (4.2) and injecting $\alpha_{\widehat{i}}^{\mathrm{ML}}$ yield

$$\max_i\ -\frac{1}{2}\left(\|\boldsymbol{x}\|_2^2 - 2(\boldsymbol{s}_i^\top \boldsymbol{x})^2 + (\boldsymbol{s}_i^\top \boldsymbol{x})^2\right) + \frac{1}{2}\|\boldsymbol{x}\|_2^2 \quad \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \quad \ln\gamma'. \tag{4.8}$$

Or equivalently the following test

$$T_{\mathrm{Max}}(\boldsymbol{x},\boldsymbol{S}) = \max_{i=1,\ldots,L}\ |\boldsymbol{s}_i^\top \boldsymbol{x}| \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma, \tag{4.9}$$

where $\gamma = \sqrt{2\ln\gamma'}$ is a threshold controlling the probability of false alarm. This threshold can always be computed by Monte Carlo simulations. This is necessary in the general setting where $\boldsymbol{S}$ is not orthogonal.

This test (4.9) is known as the *scan statistics*, *extreme value test* or *Max* test [Donoho & Jin 2004, Arias-Castro *et al.* 2005, Arias-Castro *et al.* 2010, Arias-Castro *et al.* 2011]. In this

[2] In (4.2) the max is a sup but having $\boldsymbol{\alpha}^{\mathrm{ML}} = 0$ happens with probability zero in practice.

dissertation, we refer to it as *Max* test. The following Section investigates the evolution of the detection performances of $T_{\text{Max}}(\boldsymbol{x}, \boldsymbol{S})$ as the size L of library $\boldsymbol{S}$ grows.

4.3 Complexity and loss of performances

The alternatives under consideration in the illustrative example considered in this Section are signals modeling spectral signatures of distant galaxies obtained by numerical models [Verhamme *et al.* 2012], 100 of which are shown in Figure 4.1(a). We consider these signatures may appear as emission or absorption lines and in this example have all their maximum at the same spectral channel (50).

In the cases considered below, let us denote by $\boldsymbol{s}_\ell$ the alternative likely activated under $\mathcal{H}_1$. Starting from the alternative $\boldsymbol{s}_\ell$, we build three nested libraries of alternatives of respectively 1, 10^2 and 10^5 spectral lines: $\boldsymbol{S} = \boldsymbol{s}_\ell$, $\boldsymbol{S}_{10^2} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_\ell, \ldots, \boldsymbol{s}_{10^2}]$ and $\boldsymbol{S}_{10^5} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_\ell, \ldots, \boldsymbol{s}_{10^5}]$. It can be noted that for $\boldsymbol{S} = \boldsymbol{s}_\ell$, the test (4.9) becomes $|\boldsymbol{x}^\top \boldsymbol{s}_\ell| \gtrless_{\mathcal{H}_0}^{\mathcal{H}_1} \gamma$, which is a reference test (called "Oracle NPD") as we are testing only the active alternative under $\mathcal{H}_1$. Here, the term "Oracle" is used because the index l is known but the amplitude α is unknown. If both are known, then the reference test is directly the Neyman-Pearson detector (see Section 1.4.1). These different sizes of reference libraries will allow us to examine the performances of $T_{\text{Max}}(\boldsymbol{x}, \boldsymbol{S}_L)$ as a function of L and we will see that T_{Max} is also in function of $\boldsymbol{s}_\ell$.

We are interested in the relative detection performances of

- Oracle NPD: $|\boldsymbol{x}^\top \boldsymbol{s}_\ell| \gtrless_{\mathcal{H}_0}^{\mathcal{H}_1} \gamma$.
- *Max* test (4.9): $\max\limits_{i=1,\ldots,L} |\boldsymbol{s}_i^\top \boldsymbol{x}| \gtrless_{\mathcal{H}_0}^{\mathcal{H}_1} \gamma$, (for $\boldsymbol{S}_{10^2}$ and $\boldsymbol{S}_{10^5}$),
- a simple prototype of reduced dimension (RD) techniques: $|\boldsymbol{x}^\top \boldsymbol{u}| \gtrless_{\mathcal{H}_0}^{\mathcal{H}_1} \gamma$, where $\boldsymbol{u}$ is the eigenvector associated to the largest singular value of $\boldsymbol{S}$ (see Section 3.2). Notice from Figure 4.1(c) that the atom $\boldsymbol{u}$ tends to represent a *common* profile of the alternatives in $\boldsymbol{S}$: compare $\boldsymbol{u}$ in Figure 4.1(c) or 4.1(d) to Figure 4.1(a).

The performances of each test will be evaluated in terms of ROC curves, plotted for $\text{P}_{\text{FA}} \leq 0.1$.

The tests presented above are compared in two configurations of the alternative $\boldsymbol{s}_\ell$ under $\mathcal{H}_1$:

- **Case 1**: $\boldsymbol{s}_\ell$ has a *typical similar* shape w.r.t. the atom $\boldsymbol{u}$ generated from library $\boldsymbol{S}_{10^5}$ ($\boldsymbol{s}_\ell^\top \boldsymbol{u} = 0.93$, see Figure 4.1(c)) and it is thus well correlated to most atoms of $\boldsymbol{S}_{10^5}$.
- **Case 2**: $\boldsymbol{s}_\ell$ has an *atypical dissimilar* shape w.r.t. the atom $\boldsymbol{u}$ generated from library $\boldsymbol{S}_{10^5}$ ($\boldsymbol{s}_\ell^\top \boldsymbol{u} = 0.50$, see Figure 4.1(d)) and it is thus *less correlated* to most atoms of $\boldsymbol{S}_{10^5}$ than in case 1.

This study is divided in two Sections: the first Section (4.3.1) focuses on the performances' comparisons between the *Max* tests and the RD tests, emphasizing the influence of size $\boldsymbol{S}$ on $T_{\text{Max}}(\boldsymbol{x}, \boldsymbol{S})$. The second Section (4.3.2) examines the influence of SNR on the detection performances of these tests.

4.3.1 Detection performances as library S grows

Let us consider Figure 4.1(e). The first interesting observation is that the RD test is more powerful in this setting than the *Max* test (compare blue line to cyan circles, and red crosses to pink diamonds). This comes perhaps as a surprise, since the *Max* test contains the exact alternative under $\mathcal{H}_1$, while the RD test uses only an approximate version of it. But the RD test is close to the Oracle NPD in this case (since $\boldsymbol{u}$ is close to $\boldsymbol{s}_\ell$), while as L grows, the *Max* test correlates increasingly many alternatives to the data. When L is large, this causes substantially increased false alarms with only marginal improvement in the detection rate.

This effect is not easy to address analytically for a general library such as the $\boldsymbol{S}$ considered in this example. It can however be evidenced in closed form in a simplified setting allowing to express the probability of detection and probability of false alarm as a function of L:

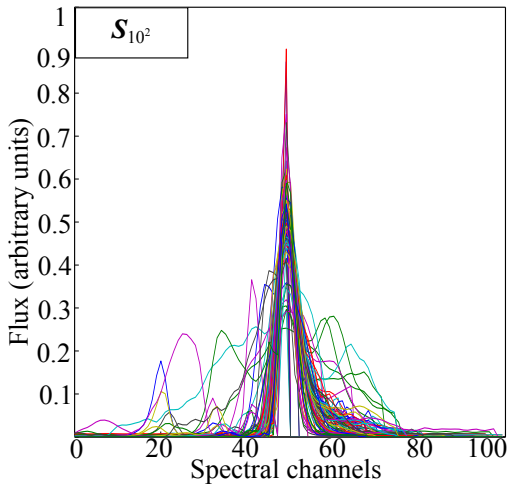
Proposition 1. *(see Appendix B.1 for a proof). Assume that $\boldsymbol{S}$ is orthonormal ($\boldsymbol{s}_i^\top \boldsymbol{s}_j = \delta_{i,j}$) and that the active alternative under $\mathcal{H}_1$ of* (4.1) *is $\alpha_\ell \boldsymbol{s}_\ell$ (without loss of generality, we assume that amplitude $\alpha_\ell = 1, \forall \ell$). Then for a fixed value $0 < P_{FA} < 1$, $P_{Det}(\boldsymbol{s}_\ell, L)$ as obtained by the Max test* (4.9) *is a decreasing function of L, $\forall \ell$. Moreover $\lim_{L \to +\infty} P_{Det}(\boldsymbol{s}_\ell, L) = P_{FA}$, $\forall \ell$.*

Proposition 1 reflects a well known fact in multiple testing: as L increases, the amplitude α_ℓ must be increasingly large (precisely, it must slightly dominate $\sqrt{2 \log N}$, the level of the maximum under the null as $L \to \infty$) for $\mathcal{H}_0$ and $\mathcal{H}_1$ to be distinguishable (see, e.g., Theorem 1.3 of [Donoho & Jin 2004]). Figure 4.1(b) depicts the behavior of the ROC curves of the *Max* test as L grows for an orthonormal matrix $\boldsymbol{S}$. For a fixed probability of false alarm, the probability of detection (B.3) decreases when L increases, and for a very large number of alternatives L, $\mathrm{P_{Det}}(\boldsymbol{s}_\ell, L) \to \mathrm{P_{FA}}$: the GLR test has asymptotically no power in a finite amplitude setting.

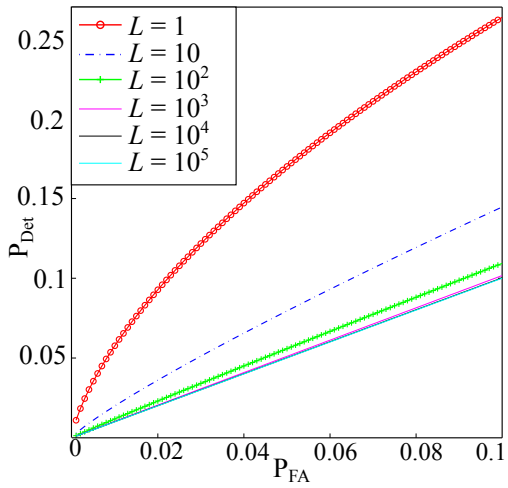
Coming back to Figure 4.1(e), we see that in contrast to the *Max* test, the RD test incurs only slight losses as L grows. An interpretation is that as the number of alternatives increases, the learned atom $\boldsymbol{u}$ represents a larger diversity of alternatives and may thus become dissimilar from $\boldsymbol{s}_\ell$. It can however not be significantly different from $\boldsymbol{s}_\ell$ if the vast majority of alternatives of $\boldsymbol{S}$ are similar to $\boldsymbol{s}_\ell$. In this situation, the detection rate of the RD test is indeed inferior to that of the Oracle NPD, but not significantly so. Thus, in this configuration of $\boldsymbol{s}_\ell$ w.r.t. $\boldsymbol{S}$, there is a clear advantage in operating in reduced dimension.

Figure 4.2 depicts a very different configuration, where $\boldsymbol{s}_\ell$ is less correlated to the atoms of $\boldsymbol{S}$ than in case 1. In this case, the relative behavior of the tests is inverted, with the RD test below the *Max* test. Indeed, testing a RD template obtained from 10^2 or 10^5 alternatives that are mostly dissimilar from $\boldsymbol{s}_\ell$ strongly penalizes the detection rate for the RD test, while the *Max* test still contains (among others) the right alternative.

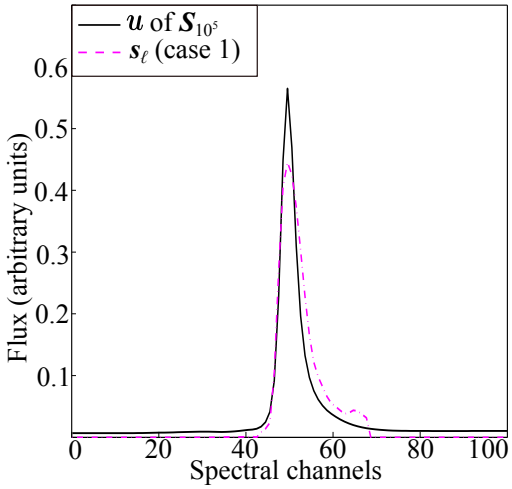
Finally, the lower performance of the *Max* test in case 2 relatively to case 1 can be explained by the fact that the alternatives of $\boldsymbol{S}$ do not help as much in the detection as in case 1 (because they are mostly dissimilar to $\boldsymbol{s}_\ell$), while the effect on the false alarm is comparable.



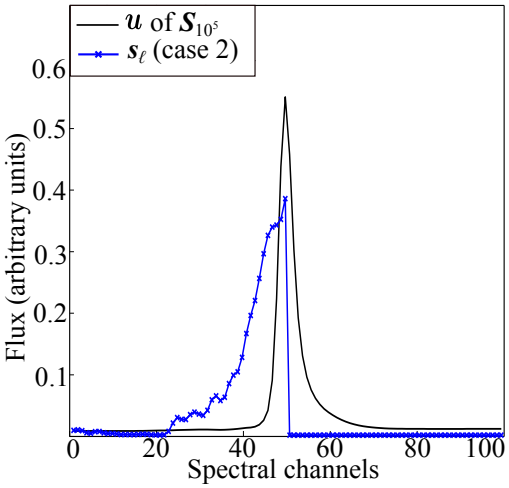
(a) The diversity of spectral lines of $\boldsymbol{S}_{10^2}$.



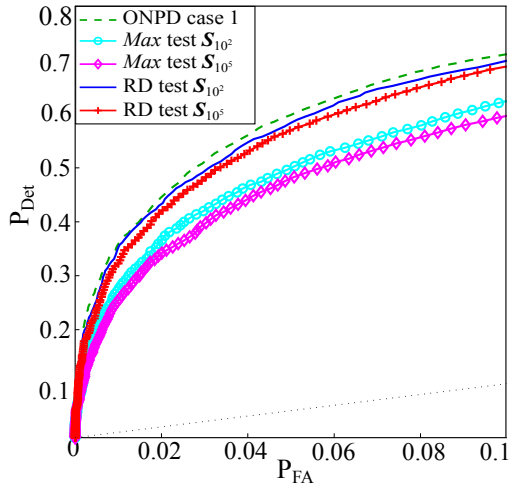
(b) Analytical ROC for (B.4).



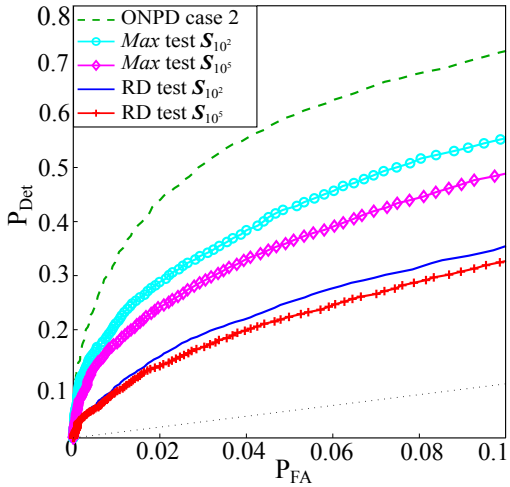
(c) Case 1: active $\boldsymbol{s}_\ell$ has similar shape to $\boldsymbol{u}$.



(d) Case 2: active $\boldsymbol{s}_\ell$ has dissimilar shape to $\boldsymbol{u}$.



(e) ROC of Case 1: RD tests perform better than *Max* tests.



(f) ROC of case 2: RD tests perform poorer than *Max* tests.

Figure 4.1: (a) 100 alternatives (spectral lines) in $\boldsymbol{S}$. (b) ROC curves showing that P_{Det} (B.4) for $\boldsymbol{S}$ orthonormal decreases as L grows, at fixed P_{FA}. Figures (c) and (d): Atom $\boldsymbol{u}$ (the rank-one approximation of library $\boldsymbol{S}_{10^5}$) and the two instances of alternatives $\boldsymbol{s}_\ell$ activated under $\mathcal{H}_1$ in Section 4.3. Figures (e) and (f): ROC curves in each case (ONPD: Oracle NPD).

4.3.2 Influence of SNR on detection performances

In this Section, we study the influence of noise levels on the detection performances of the tests presented earlier in Section 4.3.1. To recall, the noise level for (4.1) is given by

$$\mathrm{SNR}_{(\mathrm{dB})} = 10 \log_{10} (\alpha^2).$$

Here, ROC curves are plotted for different values of α, as depicted in Figure 4.2. In addition, Table 4.1 displays the corresponding values of Area Under Curves (AUCs) of each ROC.

As visible in these numerical results (Figure 4.2 and Table 4.1), the detection performances of all tests indeed depend on the noise level, where low SNR yields lower performances than high SNR. Noticeably, the behavior of *Max* tests and RD tests remains the same in term of relative performances (i.e., RD tests perform better than the *Max* test in case 1, but perform poorly in case 2) regardless of the value of α. The results of Tables 4.1 and 4.2 have to be evaluated w.r.t. the uncertainty caused by estimation noise on the ROC. The uncertainty is evaluated by the difference between the average and the minimum values of AUC of Oracle NPD. The estimation noise is Gaussian and the value of the uncertainty is related to the standard deviation σ by uncertainty $= c\sigma$, where c is a constant (typically $c = 3$ for most numerical simulations shown below).

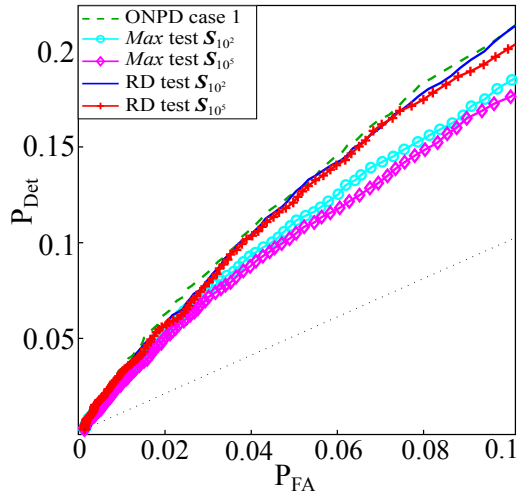
Table 4.2 compares the average and worst-case performances of *Max* test (using reference library $\boldsymbol{S}_{10^2}$) with RD test (using atom $\boldsymbol{u}$ learned from $\boldsymbol{S}_{10^2}$), for various SNR levels.

To compute these quantities, $\boldsymbol{s}_\ell$, $\ell = 1, \ldots, L$ has to be activated one by one under $\mathcal{H}_1$. At each ℓ, a high number of Monte Carlo simulations has to be done in order to evaluate the ROC (e.g., here 2×10^5 realizations). This explains the limitation of this study to $L = 10^2$ for the reference library. Note that performing 2×10^5 Monte Carlo realizations of the *Max* test for all instances $\ell = 1, \ldots, 10^5$ is too complex (in fact, it is intractable) for a standard machine[3]. Following this testing approach in practice also yields the same computation complexity when the index of active alternative is unknown. Lowering the number of realizations is possible, but at the cost of increasing the estimation noise on the ROC (thus, possibility of obtaining inaccurate results because the uncertainty is higher).

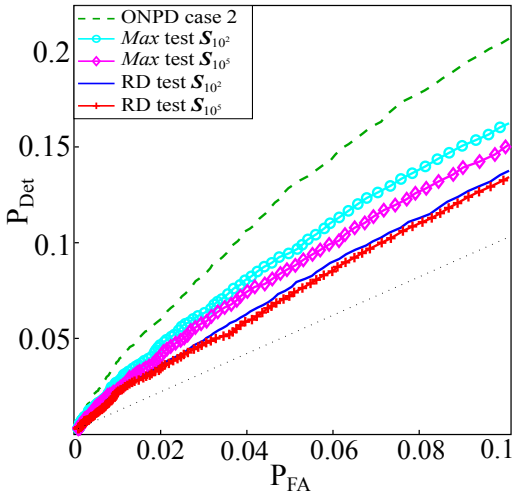
Returning to Table 4.2, the AUC of Oracle NPD is also included in the second column as reference. Numerical simulations show that for different SNR levels, the RD test has better detection performances in most cases (average AUC, compares the fifth column to the third column) than the *Max* test. However, the standard RD test considered here (SVD) is not robust with respect to some marginals alternatives. Its worst-case detection performances (as written in the last column) are inferior to those of the *Max* test using reference library $\boldsymbol{S}_{10^2}$ (see the fourth column). The RD test has better overall detection performances than *Max* test using full library, for low SNR and high SNR because it is close in this case to the Oracle NPD.

To summarize, this Section highlights that the observations of Section 4.3.1 hold true for different SNR levels.

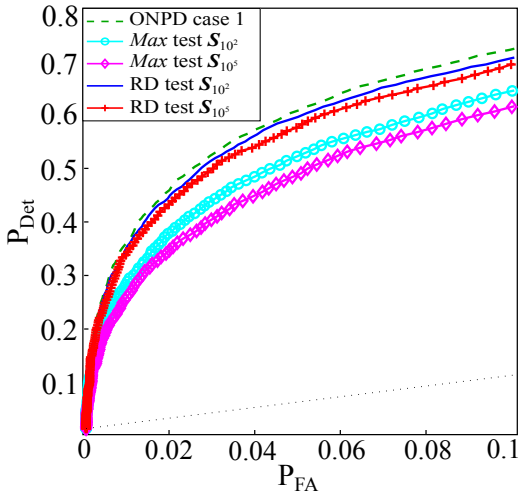
[3] A 2.7GHz processor and 4Go of DDR3 RAM.



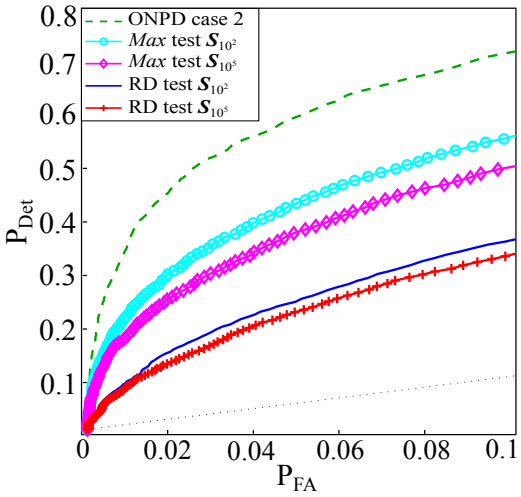
(a) ROC of case 1, $\alpha = 0.8$, SNR $= -1.94$dB.



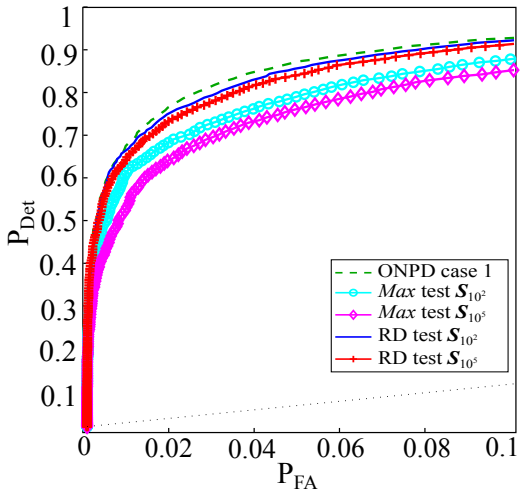
(b) ROC of case 2, $\alpha = 0.8$, SNR $= -1.94$dB.



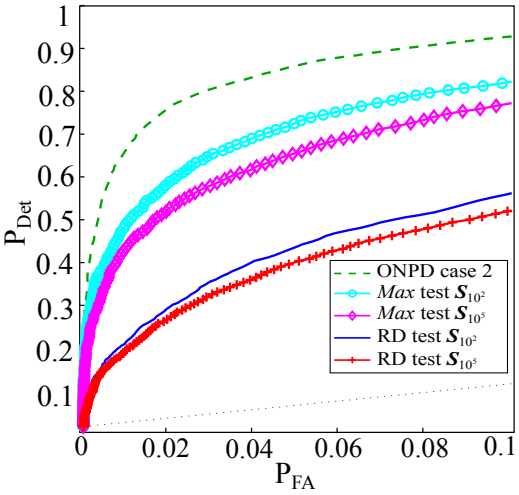
(c) ROC of case 1, $\alpha = 2.2$, SNR $= 6.85$dB.



(d) ROC of case 2, $\alpha = 2.2$, SNR $= 6.85$dB.



(e) ROC of case 1, $\alpha = 3$, SNR $= 9.54$dB.



(f) ROC of case 2, $\alpha = 3$, SNR $= 9.54$dB.

Figure 4.2: Comparison of detection performances via ROC curves, for *Max* tests and RD tests, for different values of α. ONPD denotes the Oracle NPD. The relative behavior of the tests is the same for different noise levels, i.e., in case 1, RD tests perform better than the *Max* test, however they perform poorly in case 2.

AUC of ROC	Case 1			Case 2		
	$\alpha = 0.8$ (-1.94dB)	$\alpha = 2.2$ (6.85dB)	$\alpha = 3$ (9.54dB)	$\alpha = 0.8$ (-1.94dB)	$\alpha = 2.2$ (6.85dB)	$\alpha = 3$ (9.54dB)
Oracle NPD	0.5919	0.8884	0.9669	0.5931	0.8871	0.9660
Max test $\boldsymbol{S}_{10^2}$	0.5763	0.8527	0.9483	0.5624	0.8177	0.9257
Max test $\boldsymbol{S}_{10^5}$	0.5724	0.8383	0.9397	0.5521	0.7884	0.9041
RD test $\boldsymbol{S}_{10^2}$	0.5891	0.8820	0.9640	0.5364	0.6961	0.7997
RD test $\boldsymbol{S}_{10^5}$	0.5871	0.8764	0.9608	0.5320	0.6793	0.7791

Table 4.1: AUCs corresponding to the ROC curves in Figure 4.2. Uncertainty: ± 0.0011. We compare the AUC of 5 tests, in two cases. In each case, we set three different levels of SNR (by varying α, shown in different columns) in order to study the tests' performances w.r.t. SNR. The third row shows the AUC values of the Oracle NPD as reference. For all the other tests (fourth until the last row), we can see that, in both cases, the detection performances of each test clearly depend on the noise level (low SNR yields lower performance than those for high SNR). The behavior however, remains the same as seen in Section 4.3.1 regardless of the noise levels (i.e., RD tests perform better than the *Max* tests in case 1, but perform poorly in case 2).

α	(SNR)	AUC of Oracle NPD	AUC of *Max* test $\boldsymbol{S}_{10^2}$		AUC of RD test $\boldsymbol{S}_{10^2}$	
			Average	Worst-case	Average	Worst-case
0.5	-6dB	0.5379	0.5290	0.5229	0.5348	0.5136
1	0dB	0.6371	0.6105	0.5905	0.6218	0.5503
1.5	3.5dB	0.7521	0.7112	0.6793	0.7312	0.6056
2	6dB	0.8563	0.8135	0.7780	0.8304	0.6722
2.5	8dB	0.9262	0.8919	0.8623	0.9035	0.7389

Table 4.2: AUCs for different values of α (SNR levels, shown by rows). Uncertainty: ± 0.0013. By performing *Max* test $\boldsymbol{S}_{10^2}$ and RD test $\boldsymbol{S}_{10^2}$ over all alternatives $\boldsymbol{s}_\ell$ activated one by one under $\mathcal{H}_1$, where $\ell = 1, \ldots, 10^2$, we compute the average and worst-case (i.e., minimum AUC) performances of each test, at different SNR levels. Second column shows the AUC of Oracle NPD as reference. Comparing the third and the fifth columns, we can see that RD test performs better on average than the *Max* test. However, the worst-case performance of RD test is inferior than the *Max* test (i.e., compare the last column with the fourth column). Both of these observations hold true for various SNR levels.

4.4 Discussion

In conclusions, these experiments show that in the "one among many" detection problem set by (4.1), the "curse of dimensionality" [Bellman 1957] has two effects. First, the computational complexity of this test grows linearly with L, which can be far too demanding for applications involving large L and multiple data sets [Paris *et al.* 2013b]. Second, the performances of the *Max* test (i.e., the constrained GLR using the full library) degrade as L grows.

This study also shows that GLR testing in learned subspaces of reduced dimension (exemplified using the first eigenvector here) is indeed advantageous w.r.t. *Max* test computationally-wise and also performance-wise as far as the active alternative is well correlated to the learned template. For alternatives for which this is not the case, dimensionality reduction results in weak detection power. These observations suggest that desirable properties in this framework may be to devise tests of low complexity while uniformly controlling the worst-case performances over the alternatives.

The next Chapter will do so. We will further investigate how to devise a test that could perform as well as the constrained GLR of exact model, but without testing all the possible alternatives, while still being robust for all of them. With this in mind, we will formulate the optimization problem as maximizing the worst probability of detection occurring for all $\boldsymbol{s}_i$, $\{i = 1, \ldots, L\}$.

Chapter 5

Detection tests for reduced dimension models

Contents

Some analyses and results presented in this chapter were published in [Suleiman *et al.* 2013a, Suleiman *et al.* 2013b, Paris *et al.* 2013b, Suleiman *et al.* 2014a, Suleiman *et al.* 2014b].

5.1 Introduction

In the preceding Chapter, we have concluded that the performance of the GLR for the "exact" model

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

(i.e., testing the full reference library $\boldsymbol{S} \in \mathbb{R}^{N \times L}$) degrades as L increases, and the computation complexity increases w.r.t. L. On the contrary, reducing the dimension not only saves computing power, but may also result in power gain for typical alternatives and power loss for the marginal ones.

Based on these observations, this Chapter is dedicated to devise a RD test that is robust for all alternatives. Robustness may be essential in some applications such as detection of very faint spectral lines or detection of gas leakage in pipelines.

Given these points, the optimization problem of the RD test is formulated as maximizing the worst probability of detection occurring for all $\boldsymbol{s}_i$, $\{i = 1, \ldots, L\}$ (which is a maximin problem). Equivalently, it can be viewed as minimizing the maximum detection power loss (which is a minimax problem). We will use below the first formulation leading to an

optimized dictionary $\boldsymbol{D}^* \in \mathbb{R}^{N\times K}$, where $K < L$. We shall generically refer to the resulting optimization as minimax optimization (superscript * will denote "minimax optimized").

It may be useful for clarity to recall that the minimax optimization problem addressed in our framework differs with the usual setting considered in the literature (see Section 1.6). Some of the important differences are

- We consider the case where the set of all alternatives in the reference library $\boldsymbol{S} \in \mathbb{R}^{N\times L}$ is countable.

- The alternatives are known up to finite amplitude factor, are in fixed number (L) and fixed size (N).

- No distribution is specified under $\mathcal{H}_1$ of the exact model (4.1) (i.e., we do not attempt to specify a discrete distribution reflecting the probability of activation of each possible alternative under $\mathcal{H}_1$; equivalently, all alternatives are considered equally likely).

- There will be no degree of freedom in the choice of the test that will support the worst-case testing procedure. The test is based on constrained and unconstrained GLR. It only depends on the dictionary to be optimized $\boldsymbol{D} \in \mathbb{R}^{N\times K}$.

In line with this setting, we propose two RD models on the basis of which the dictionary will be optimized to obtain minimax tests.

i. The first RD model imposes a 1-sparsity constraint on the unknown vector, similar to the exact model (4.1) but with $\boldsymbol{D}$ instead of $\boldsymbol{S}$. We will consider first the mono dimensional case ($K = 1$, the dictionary $\boldsymbol{D}$ has only one atom that will be optimized). This particular case will be the base for constructing algorithms in more general case $K > 1$ proposed in Chapter 6.

ii. The second RD model (presented in Section 5.3) does not impose any sparsity constraint under $\mathcal{H}_1$. The sparsity is controlled by the number of columns in the learned dictionary.

5.2 Reduced model with sparsity constraint

5.2.1 The model and associated GLR test

We propose to replace the exact model (4.1) by

$$\begin{cases} \mathcal{H}_0 & : \boldsymbol{x} = \boldsymbol{n}, & \boldsymbol{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \\ \mathcal{H}_1 & : \boldsymbol{x} = \boldsymbol{D}\boldsymbol{\beta} + \boldsymbol{n}, & \|\boldsymbol{\beta}\|_0 = 1 \end{cases} . \tag{5.1}$$

Here $\boldsymbol{D} \in \mathbb{R}^{N \times K} = [\boldsymbol{d}_1, \ldots, \boldsymbol{d}_K]$ is a low dimension ($K < L$) dictionary with ℓ_2 normalized columns, $\|\boldsymbol{d}_j\|_2^2 = 1$, $\{j = 1, \ldots, K\}$. The 1-sparse constraint on unknown vector $\boldsymbol{\beta}$ is imposed to encourage the axes of $\boldsymbol{D}$ to align, when optimized, with the main "modes" (possibly represented by isolated alternatives) of the distribution of $\mathcal{S}$ over the unit sphere.

For the analysis that follows, we consider that all alternatives have the same amplitude when activated under $\mathcal{H}_1$. As explained above, this makes them comparable in terms of SNR and we assume without loss of generality unit amplitude. The dictionary $\boldsymbol{D}$ will be optimized to maximize the worst-case detection performance.

The constrained GLR for model (5.1) involves the constrained Maximum Likelihood estimate of $\boldsymbol{\beta}$

$$T_{\mathrm{GLR}}(\boldsymbol{x}, \boldsymbol{D}) = \max_{\boldsymbol{\beta}:\|\boldsymbol{\beta}\|_0=1} \frac{p(\boldsymbol{x}; \boldsymbol{D}\boldsymbol{\beta})}{p(\boldsymbol{x}; \mathbf{0})} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \xi', \tag{5.2}$$

where ξ' is a threshold. Following the steps (4.2)-(4.9), the GLR for (5.1) and for a given $\boldsymbol{D} = [\boldsymbol{d}_1, \ldots, \boldsymbol{d}_K]$ leads to

$$T_{\boldsymbol{D}}(\boldsymbol{x}) = \max_{j=1,\ldots,K} |\boldsymbol{d}_j^\top \boldsymbol{x}| \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \xi \Leftrightarrow \max_{j=1,\ldots,K} (\boldsymbol{d}_j^\top \boldsymbol{x})^2 \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \xi^2, \tag{5.3}$$

where ξ is a threshold.

The corresponding probability of false alarm and probability of detection of (5.3) write

$$\mathrm{P}_{\mathrm{FA}}(\boldsymbol{D}) = \mathbb{P}\left(\max_{j=1,\ldots,K} (\boldsymbol{d}_j^\top \boldsymbol{x})^2 > \xi^2; \mathcal{H}_0, \boldsymbol{D} \right), \tag{5.4}$$

$$\mathrm{P}_{\mathrm{Det}}(\boldsymbol{s}_\ell, \boldsymbol{D}) = \mathbb{P}\left(\max_{j=1,\ldots,K} (\boldsymbol{d}_j^\top \boldsymbol{x})^2 > \xi^2; \mathcal{H}_1, \boldsymbol{s}_\ell, \boldsymbol{D} \right). \tag{5.5}$$

Following a minimax strategy, we seek to optimize dictionary $\boldsymbol{D}$, so as to maximizes the minimum probability of detection occurring for all alternatives $\boldsymbol{s}_i$, $\{i = 1, \ldots, L\}$, at a fixed probability of false alarm ($\mathrm{P}_{\mathrm{FA}_0}$). Hence we define

$$\begin{aligned} \boldsymbol{D}^* := \arg\max_{\boldsymbol{D}} \quad & \min_{i=1,\ldots,L} \mathrm{P}_{\mathrm{Det}}(\boldsymbol{s}_i, \boldsymbol{D}) \\ \text{subject to} \quad & \mathrm{P}_{\mathrm{FA}}(\boldsymbol{D}) \leq \mathrm{P}_{\mathrm{FA}_0}, \\ & \|\boldsymbol{d}_j\|_2 = 1, \ j = \{1, \ldots, K\}. \end{aligned} \tag{5.6}$$

We examine first the particular case of one-dimensional minimax problem ($K = 1$), then we extend this study for an arbitrary value of K.

5.2.2 One-dimensional minimax problem ($K = 1$)

In this setting, the one-dimensional dictionary to be optimized is $\boldsymbol{D} = \boldsymbol{d}$. The GLR computed above (5.3) for arbitrary value of K becomes

$$T_{\boldsymbol{d}}(\boldsymbol{x}) = (\boldsymbol{d}^\top \boldsymbol{x})^2 \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \xi^2. \tag{5.7}$$

In this case, the corresponding probabilities of test (5.7) are

$$\mathrm{P}_{\mathrm{FA}}(\boldsymbol{d}) = \mathbb{P}((\boldsymbol{d}^\top \boldsymbol{x})^2 > \xi^2; \mathcal{H}_0), \tag{5.8}$$

$$\mathrm{P}_{\mathrm{Det}}(\boldsymbol{s}_\ell, \boldsymbol{d}) = \mathbb{P}((\boldsymbol{d}^\top \boldsymbol{x})^2 > \xi^2; \mathcal{H}_1, \boldsymbol{s}_\ell). \tag{5.9}$$

The one-dimensional minimax optimization problem can be written as

$$\begin{aligned} \boldsymbol{d}^* := \arg\max_{\boldsymbol{d}} \min_{i=1,\dots,L} & \mathrm{P}_{\mathrm{Det}}(\boldsymbol{s}_i, \boldsymbol{d}) \\ \text{subject to} \quad & \mathrm{P}_{\mathrm{FA}}(\boldsymbol{d}) \leq \mathrm{P}_{\mathrm{FA}_0}, \\ & \|\boldsymbol{d}\|_2 = 1, \end{aligned} \tag{5.10}$$

where $\mathrm{P}_{\mathrm{FA}_0}$ is a fixed alarm rate. The previous Proposition 4 and the result of Proposition 2 below will guide us to resolve this one-dimensional problem (5.10).

Proposition 2. *Generalized Marcum-Q function [Nutall 1974].*
Let $Q_{\frac{k}{2}}(x, y)$ be the generalized Marcum-Q function of order $\frac{k}{2}$, with $x \in \mathbb{R}^{*+}$ and $y \in \mathbb{R}^+$. Then $Q_{\frac{k}{2}}(x, y) = 1 - \Phi_{\chi^2_{k,x^2}}(y^2)$, with $\Phi_{\chi^2_{k,x^2}}$ the CDF of a $\chi^2_{k,\lambda}$ variable, with k degrees of freedom and non-centrality parameter λ. $Q_{\frac{k}{2}}(x, y)$ is monotonically increasing in x.

Assuming that under $\mathcal{H}_1$ of the exact model (4.1), an alternative $\boldsymbol{s}_\ell$ is activated, $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{s}_\ell, \mathbf{I})$. Since $(\boldsymbol{d}^\top \boldsymbol{x})^2 = \boldsymbol{x}^\top (\boldsymbol{d}\boldsymbol{d}^\top)\boldsymbol{x}$, Proposition 4 applied to $\boldsymbol{A} = \boldsymbol{d}\boldsymbol{d}^\top$ gives:

$$\mathrm{P}_{\mathrm{FA}} = \mathbb{P}((\boldsymbol{d}^\top \boldsymbol{x})^2 > \xi^2; \mathcal{H}_0) = 1 - \Phi_{\chi^2_1}(\xi^2), \tag{5.11}$$

$$\begin{aligned} &\mathrm{P}_{\mathrm{Det}}(\boldsymbol{s}_\ell, \boldsymbol{d}) = \mathbb{P}((\boldsymbol{d}^\top \boldsymbol{x})^2 > \xi^2; \mathcal{H}_1, \boldsymbol{s}_\ell) = 1 - \Phi_{\chi^2_{1,\lambda}}(\xi^2) \\ &\mathrm{P}_{\mathrm{Det}}(\boldsymbol{s}_\ell, \boldsymbol{d}) = Q_{\frac{1}{2}}(\sqrt{\lambda}, \xi), \end{aligned} \tag{5.12}$$

where here, $\lambda = \boldsymbol{s}_\ell^\top (\boldsymbol{d}\boldsymbol{d}^\top)\boldsymbol{s}_\ell = (\boldsymbol{d}^\top \boldsymbol{s}_\ell)^2$ and Q is the generalized Marcum-Q function. By Proposition 2, $Q_{\frac{1}{2}}(\sqrt{\lambda}, \xi)$ is increasing in $\sqrt{\lambda} = |\boldsymbol{d}^\top \boldsymbol{s}_\ell|$. Hence, maximizing the probability of detection is equivalent to maximizing λ.

Now, from (5.11), the probability of false alarm for $K = 1$ is independent of $\boldsymbol{d}$. Hence, maximizing at fixed false alarm rate the probability of detection occurring in the worst-case alternative(s) under $\mathcal{H}_1$ requires to solve

$$\boldsymbol{d}^* = \arg\max_{\boldsymbol{d}:\|\boldsymbol{d}\|_2=1} \min_{i=1,\dots,L} (\boldsymbol{d}^\top \boldsymbol{s}_i)^2 = \arg\max_{\boldsymbol{d}:\|\boldsymbol{d}\|_2=1} \min_{i=1,\dots,L} |\boldsymbol{d}^\top \boldsymbol{s}_i|, \tag{5.13}$$

which is non convex because of the non convex constraint $\|\boldsymbol{d}\|_2 = 1$ (see Figure 5.1).

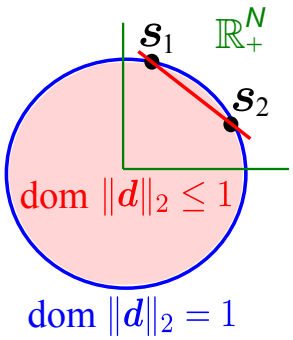


Figure 5.1: Convex and non convex domain. If we draw a line segment between two samples $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$, every point on this (red) line segment does not belong to dom $\|\boldsymbol{d}\|_2 = 1$. Instead, it belongs to dom $\|\boldsymbol{d}\|_2 \leq 1$. This is why the constraint $\|\boldsymbol{d}\|_2 = 1$ of (5.13) is non convex.

Under some conditions, the exact solution of $\boldsymbol{d}^*$ can however be obtained by solving a convex optimization problem in the form of quadratic programming (QP), as evidenced in the following Proposition:

Proposition 3. *(see Appendix B.2 for a proof) The solutions of* (5.13) *are* $\{\boldsymbol{d}^*, -\boldsymbol{d}^*\}$*, and assuming that all* $\mathcal{S} \in \mathbb{R}_+^N$*, then* $\boldsymbol{d}^* \in \mathbb{R}_+^N$ *is the solution of the QP:*

$$\begin{aligned} \boldsymbol{d}^* = \quad & \textit{minimize} \quad -t \\ & \textit{subject to} \quad t - \boldsymbol{d}^\top \boldsymbol{s}_i \leq 0,\ i = \{1, \ldots, L\} \\ & \qquad\qquad \|\boldsymbol{d}\|_2 \leq 1. \end{aligned} \tag{5.14}$$

The main condition for this Proposition to hold is that the elements of the set $\mathcal{S}$ are, up to a common rotation, in $\mathbb{R}_+^N$. This condition is not very restrictive in practice, as many applications in signal or image processing deal with spectra or images contain only positive values. QP problem (5.14) can be solved via a standard toolbox such as the `CVX` [Grant & Boyd 2014].

As a side remark, we also present in Appendix B.3 a possible gradient descent type method to solve one-dimensional minimax optimization problem (5.13). We investigated this approach (gradient descent) for $K = 1$, hoping to generalize it to the case $K > 1$ and to obtain exact minimax solution for such cases. This generalization is however not achieved yet because it poses several important issues (complexity of the coordinate system in $N > 3$-dimensional Euclidean space and the case where the function to be optimized is non-differentiable).

5.2.3 Connections of 1D minimax problem to 1-class classifiers

The minimax optimization problem for $K = 1$ (5.13) entails finding the vector $\boldsymbol{d}^*$ that minimizes the largest angle between $\boldsymbol{s}_i$ and $\boldsymbol{d}$. This is equivalent to finding the smallest circle $\mathcal{C}$ containing the set $\mathcal{S}$, whose center then corresponds to $\boldsymbol{d}^*$. Or finding a plane that maximizes the worst probability of detection, and this plane is perpendicular to $\boldsymbol{d}^*$. As depicted in Figure 5.2, this type of optimization problem can also be viewed as a One-class classifiers problem of SVM type [Boser *et al.* 1992, Cortes & Vapnik 1995].

As discussed in Chapter 3, a conventional classification task is typically a k-ary ($k > 1$) problem where the new sample is to be assigned to one of the k classes built from the available data. In one-class classification however, the problem is to maximally discriminate between the known samples and possible outliers w.r.t. this class [Khan & Madden 2010]. This involves finding the hyperplane of farthest distance from the origin which rejects all training samples aside, termed one-class SVM [Schölkopf 1997]. In addition, it is also equivalent to finding the minimal volume of a closed boundary sphere (Σ in Figure 5.2, of center $\boldsymbol{a}$ and radius R) containing all $\boldsymbol{s}_i$. This is known as SVDD method [Tax & Duin 2004].

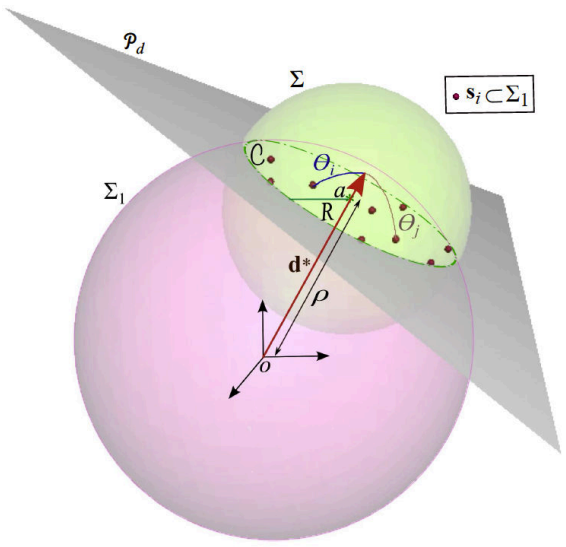


Figure 5.2: Geometrical view of One-dimensional minimax optimization problem (5.13)-(5.14) as One-class classifiers. The set of alternatives $\boldsymbol{s}_i$ lie at the intersection of a unit sphere Σ_1. The problem is equivalent to minimizing the largest angle θ_i between $\boldsymbol{d}$ and $\boldsymbol{s}_i$, to finding the circle $\mathcal{C}$ of minimum radius R that contains all $\boldsymbol{s}_i$, to maximizing the distance ρ of hyperplane $\mathcal{P}_d$ to the origin (i.e., one-class SVM), or to minimizing the volume of an enclosed sphere Σ containing all $\boldsymbol{s}_i$ (i.e., SVDD). In this setting, Σ admits $\mathcal{C}$ as a great circle.

Example 5.2.1. 1D minimax dictionary vs. best rank-one approximation of $\boldsymbol{S}$

A simple numerical experiments using random positives alternatives forming a reference library $\boldsymbol{S} \in \mathbb{R}^{3\times 115}$ is examined, as shown in Figure 5.3. The ℓ_2 normalization implies that all alternatives $\boldsymbol{s}_i$ (black crosses) lie on the unit sphere (in green). From $\boldsymbol{S}$, we learn one-dimensional dictionaries based on two approaches

i. best rank-one approximation (in Frobenius norm) of $\boldsymbol{S}$, yielding atom $\boldsymbol{u}$ (see Section 3.2).

ii. 1D minimax optimization (5.13)-(5.14), yielding atom $\boldsymbol{d}^*$.

Figure 5.3 shows that there are three worst-case alternatives (small red circles) w.r.t. $\boldsymbol{d}^*$ (red star), which belong to the smallest circle $\mathcal{C}$ enclosing set $\mathcal{S}$. These "outliers" (i.e., marginal alternatives) induce the worst probability of detection when using $\boldsymbol{d}^*$ for detection. By definition, no single atom dictionary can achieve best worst-case performances than $\boldsymbol{d}^*$.

Comparing the minimax atom to the eigenvector $\boldsymbol{u}$ associated to the largest eigenvalue $\boldsymbol{S}$ (i.e., best rank-one approximation), we notice that $\boldsymbol{u}$ (blue diamond) lies in the core of the most populated area of alternatives, and thus tends to represent a common behavior of the alternatives. This comes of course at the price of a smaller correlation w.r.t. marginal alternatives, hence lower worst-case detection power. The minimum correlation between $\boldsymbol{u}$ and set $\mathcal{S}$ is 0.47, while it is 0.71 for $\boldsymbol{d}^*$ (this happens for the three worst-case alternatives in $\mathcal{C}$).

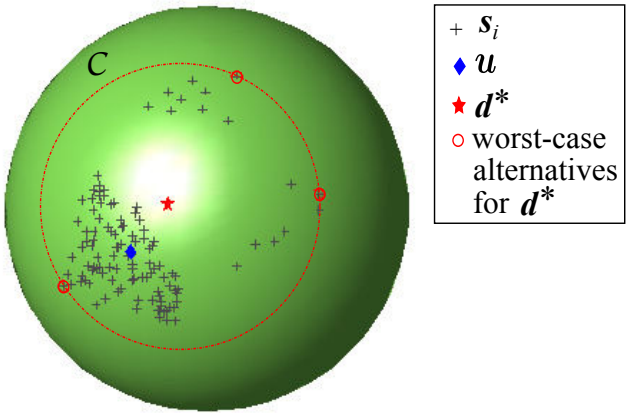


Figure 5.3: $\boldsymbol{d}^*$ is held by three marginal alternatives (at the border of the smallest enclosed circle $\mathcal{C}$). These marginal samples induce the worst probability of detection. For comparison, $\boldsymbol{u}$ represents well the most populated area of the alternatives.

Another obvious observation evidenced by this example is that representing $\boldsymbol{S}$ by only one (even minimax) atom may indeed be insufficient w.r.t. the intrinsic diversity of $\mathcal{S}$. For instance, Figure 5.3 suggests the existence of three subpopulations. In such situations, learning dictionaries with $K > 1$ columns might increase the worst-case power w.r.t. $K = 1$. This is the point of the next analysis. ■

We now turn to an illustrative example in regard to the test statistic (5.7) to compare the cost function of average probability of detection vs. the cost function of minimax probability of detection. Let us assume that $\boldsymbol{s}_{i^*}$ is one of the alternatives in $\boldsymbol{S}$ that is the most poorly correlated to atom $\boldsymbol{d}$.

Example 5.2.2. Study of detection rate: illustration on unit sphere

From (5.9), assuming $\alpha_i = 1$, we define the average probability of detection as

$$\overline{\mathrm{P_{Det}}}(\boldsymbol{S}, \boldsymbol{d}) = \sum_{i=1}^{L} \mathbb{P}(\boldsymbol{s}_i)\, \mathbb{P}(|\boldsymbol{d}^\top(\boldsymbol{s}_i + \boldsymbol{n})| > \xi\,; \boldsymbol{s}_i).$$

If $\mathbb{P}(\boldsymbol{s}_i) = \frac{1}{L}$, $\forall i$ (same probability of activation), $\overline{\mathrm{P_{Det}}}(\boldsymbol{S}, \boldsymbol{d})$ becomes

$$\begin{aligned}
\overline{\mathrm{P_{Det}}}(\boldsymbol{S}, \boldsymbol{d}) &= \frac{1}{L} \sum_{i}^{L} \mathbb{P}(|\mu_i + \varepsilon| > \xi\,; \boldsymbol{s}_i), \quad \text{with } \mu_i = \boldsymbol{d}^\top \boldsymbol{s}_i, \text{ and } \varepsilon = \boldsymbol{d}^\top \boldsymbol{n} \\
&= \frac{1}{L} \sum_{i=1}^{L} \mathbb{P}(|\mathcal{N}(\mu_i, 1)| > \xi\,; \boldsymbol{s}_i). \\
\overline{\mathrm{P_{Det}}}(\boldsymbol{S}, \boldsymbol{d}) &= \frac{1}{L} \sum_{i=1}^{L} [1 - \Phi(\xi - \mu_i) + \Phi(-\xi - \mu_i)], \quad \text{(see Figure 5.4).}
\end{aligned}$$

The CDF: $\Phi(-\xi - \mu_i)$ is equal to $1 - \Phi(\xi + \mu_i)$ i.e., the area to the right from the threshold $\xi + \mu_i$ (marked by red line in the left subfigure of Figure 5.4), hence $\overline{\mathrm{P_{Det}}}(\boldsymbol{S}, \boldsymbol{d})$ also writes

$$\overline{\mathrm{P_{Det}}}(\boldsymbol{S}, \boldsymbol{d}) = \frac{1}{L} \sum_{i=1}^{L} [2 - \Phi(\xi - \mu_i) - \Phi(\xi + \mu_i)]. \tag{5.15}$$

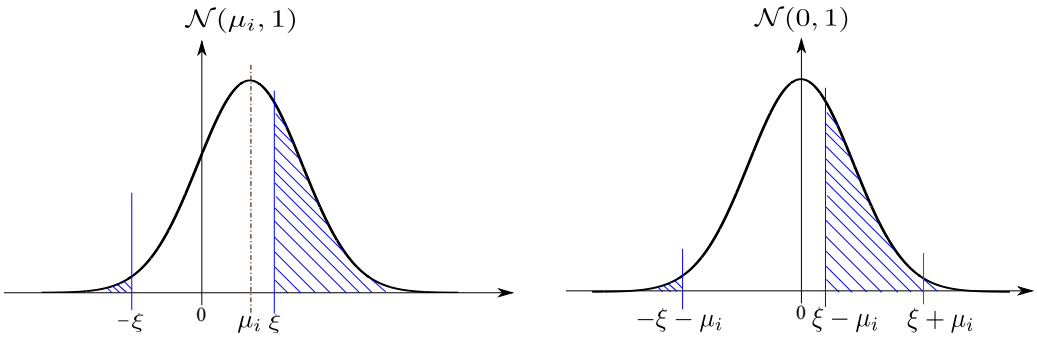


Figure 5.4: Probability distribution under $\mathcal{H}_1$. The left subfigure shows the initial distribution corresponding to $\mathcal{N}(\mu_i, 1)$. The right subfigure depicts the equivalent distribution after subtracting μ_i. The red line marks the equivalent area (to its right) of the CDF $\Phi(-\xi - \mu_i)$.

The minimum probability of detection $\mathrm{P}_{\mathrm{Det}}^{i^*}$ can be defined as

$$\mathrm{P}_{\mathrm{Det}}^{i^*} = \min_i \ \mathrm{P}_{\mathrm{Det}}(\boldsymbol{s}_i, \boldsymbol{d}) = [2 - \Phi(\xi - \boldsymbol{d}^\top \boldsymbol{s}_{i^*}) - \Phi(\xi + \boldsymbol{d}^\top \boldsymbol{s}_{i^*})]. \tag{5.16}$$

This conducts to the minimax detection criterion

$$\max_{\boldsymbol{d}} \mathrm{P}_{\mathrm{Det}}^{i^*} = \max_{\boldsymbol{d}} \ [2 - \Phi(\xi - \boldsymbol{d}^\top \boldsymbol{s}_{i^*}) - \Phi(\xi + \boldsymbol{d}^\top \boldsymbol{s}_{i^*})]. \tag{5.17}$$

We can visualize the cost functions of two distinct criteria: average (5.15) and minimax (5.17) in function of $\boldsymbol{d}$ (see Figure 5.5) on unit sphere, as shown in Figure 5.6 below.

Assume that $\boldsymbol{S} \in \mathbb{R}^{3\times L}$, we define an unknown atom $\boldsymbol{d}$

$$\boldsymbol{d} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} r\sin\phi\cos\theta \\ r\sin\phi\sin\theta \\ r\cos\phi \end{pmatrix}, \tag{5.18}$$

where the norm $r = \|\boldsymbol{d}\|_2 = 1$, $\theta \in [0, 2\pi]$ is the azimuth angle and $\phi \in [\frac{\pi}{2}, -\frac{\pi}{2}]$ is the elevation angle (see Figure 5.5).

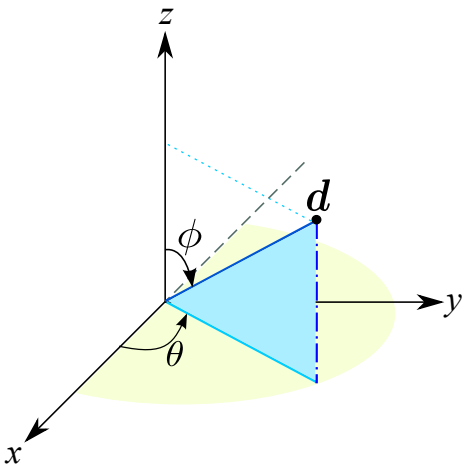


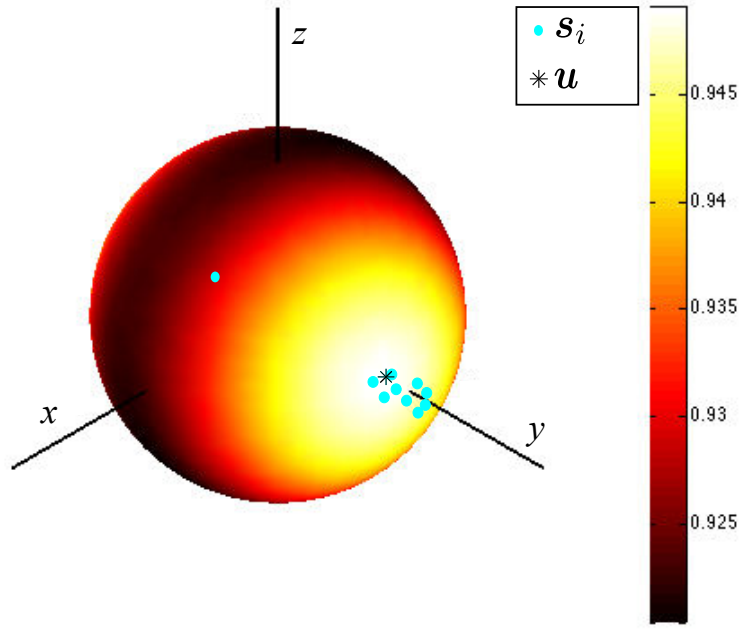
Figure 5.5: Spherical coordinates of the atom $\boldsymbol{d}$ used to represent the mean and minimax cost functions on the sphere.

To illustrate the minimax criterion we proceed as follows. First, we form a random library $\boldsymbol{S}$ of $L = 10$ (cyan dots). These data points $\boldsymbol{s}_i$, $i = 1, \ldots, L$ are distributed on the unit sphere into two groups. The first (isolated) cluster contains only one alternative, and the other nine alternatives are in the second cluster. Then, plotting the average detection criterion of (5.15) and minimax detection criterion of (5.17) as a function of the location θ, ϕ on the sphere yield the shown results (Figure 5.6). The learned SVD atom $\boldsymbol{u}$ and minimax atom $\boldsymbol{d}^*$ of (5.13) are also illustrated in the Figure.

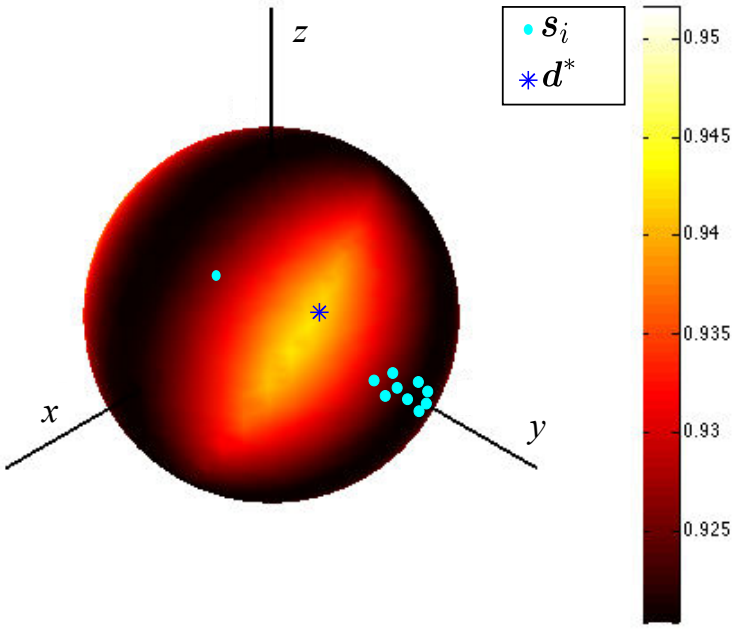
We can see (in Figure 5.6(b)) that even though most of the data points are gathered in the second cluster (near the y-axis), the minimax criterion takes into account the single (isolated) alternative of the first cluster. The minimax detection rate is large somewhere

between these two separated populations of $\boldsymbol{S}$ (see Figure 5.6(b)). The minimax atom $\boldsymbol{d}^*$ (blue star) is in agreement with the minimax cost function, i.e., $\boldsymbol{d}^*$ is situated in the area where the minimum probability of detection is large (precisely at the maximum).

In contrast, the average detection rate is smoothly increasing toward the second cluster where most of the alternatives lie, as depicted in Figure 5.6(a). The rank-one SVD atom $\boldsymbol{u}$ (black star) represents the most populated cluster: $\boldsymbol{u}$ is situated in the area where the average probability of detection is large (but not precisely at this maximum, though).



(a) Average detection rate of (5.15)



(b) Minimax detection rate of (5.17)

Figure 5.6: Cost functions of the two distinct criteria mean and minimax illustrated on the unit sphere. Cyan dots are the alternatives $\boldsymbol{s}_i \in \boldsymbol{S}$. There are one alternative $\boldsymbol{s}_i$ in the first cluster, and nine alternatives in the second cluster which lie near the y-axis of the unit sphere. (a) The average detection criterion (5.15) is smoothly increasing toward the most populated cluster. We also show here the learned SVD atom $\boldsymbol{u}$ (black star). (b) The minimax detection criterion (5.17) is maximum somewhere between the two clusters, taking into account the single alternative, situated far from the other alternatives. The minimax learned atom $\boldsymbol{d}^*$ (blue star) of (5.13) is also shown here and is exactly at the maximum. ■

Now we get back to the minimax optimization problem presented below for $K > 1$.

5.2.4 Optimization for K-dimensional subspaces

Solving the optimization problem (5.6) for $K > 1$ is an extremely intricate task as it involves the distributions of the maximum of correlated variables ($\boldsymbol{d}_j^\top \boldsymbol{x}$, $j = 1, \ldots, K$ are correlated). For the same reason, the computation of $\mathrm{P_{FA}}(\boldsymbol{D})$ (5.4) is analytically intractable. To recall, the corresponding probabilities defined earlier (5.4)-(5.5) of the RD minimax test (5.3) are

$$\mathrm{P_{FA}}(\boldsymbol{D}) = \mathbb{P}\left(\max_{j=1,\ldots,K} (\boldsymbol{d}_j^\top \boldsymbol{x})^2 > \xi^2; \mathcal{H}_0, \boldsymbol{D}\right),$$
$$\mathrm{P_{Det}}(\boldsymbol{s}_\ell, \boldsymbol{D}) = \mathbb{P}\left(\max_{j=1,\ldots,K} (\boldsymbol{d}_j^\top \boldsymbol{x})^2 > \xi^2; \mathcal{H}_1, \boldsymbol{s}_\ell, \boldsymbol{D}\right).$$

To circumvent these difficulties, we propose to replace the exact expressions of probability of false alarm and probability of detection by appropriate upper and lower bounds with simpler expressions. These bounds will be used as proxies to guide the optimization process. We study the bounds for probability of detection in Section 5.2.4.1, and for probability of false alarm in Section 5.2.4.2.

5.2.4.1 Study of detection rate for $K > 1$

We start by analyzing the bound for the probability of detection which is derived using the following result on extrema distributions.

Lemma 2. *CDF of the maximum of variables (e.g. [Nelsen 2006], chap. 2).*
Let $(X_1, \ldots, X_N)$ *be* N *continuous random variables with distribution functions* $(F_1, \ldots, F_N)$. *Then*

$$F_{\max(X_1,\ldots,X_N)}(t) \leq \min\left(F_1(t), \ldots, F_N(t)\right). \tag{5.19}$$

From (5.5),

$$\begin{aligned} \mathrm{P_{Det}}(\boldsymbol{s}_\ell, \boldsymbol{D}) &= \mathbb{P}\left(\max_{j=1,\ldots,K} (\boldsymbol{d}_j^\top \boldsymbol{x})^2 > \xi^2; \mathcal{H}_1, \boldsymbol{s}_\ell, \boldsymbol{D}\right) \\ &= 1 - \mathbb{P}\left(\max_{j=1,\ldots,K} (\boldsymbol{d}_j^\top \boldsymbol{x})^2 < \xi^2; \mathcal{H}_1, \boldsymbol{s}_\ell, \boldsymbol{D}\right) \\ \mathrm{P_{Det}}(\boldsymbol{s}_\ell, \boldsymbol{D}) &= 1 - F_{\max((\boldsymbol{d}_1^\top \boldsymbol{x})^2,\ldots,(\boldsymbol{d}_K^\top \boldsymbol{x})^2)}(\xi^2). \end{aligned} \tag{5.20}$$

Then, applying the inequality (5.19) of Lemma 2 to (5.20) yields:

$$\begin{aligned} 1 - F_{\max((\boldsymbol{d}_1^\top \boldsymbol{x})^2,\ldots,(\boldsymbol{d}_K^\top \boldsymbol{x})^2)}(\xi^2) &\geq 1 - \min(F_{(\boldsymbol{d}_1^\top \boldsymbol{x})^2}(\xi^2), \ldots, F_{(\boldsymbol{d}_K^\top \boldsymbol{x})^2}(\xi^2)) \\ \mathrm{P}_{\mathrm{Det}}(\boldsymbol{s}_\ell, \boldsymbol{D}) &\geq \max(1 - F_{(\boldsymbol{d}_1^\top \boldsymbol{x})^2}(\xi^2), \ldots, 1 - F_{(\boldsymbol{d}_K^\top \boldsymbol{x})^2}(\xi^2)) \\ &\geq \max_{j=1,\ldots,K} \mathbb{P}((\boldsymbol{d}_j^\top \boldsymbol{x})^2 > \xi^2; \mathcal{H}_1, \boldsymbol{s}_\ell, \boldsymbol{D}) \\ &\geq \max_{j=1,\ldots,K} Q_{\frac{1}{2}}(|\boldsymbol{d}_j^\top \boldsymbol{s}_\ell|, \xi). \end{aligned}$$

Based on this computation,

$$\mathrm{P}_{\mathrm{Det}}(\boldsymbol{s}_\ell, \boldsymbol{D}) = \mathbb{P}\left(\max_{j=1,\ldots,K} (\boldsymbol{d}_j^\top \boldsymbol{x})^2 > \xi^2; \mathcal{H}_1, \boldsymbol{s}_\ell, \boldsymbol{D}\right) \geq \max_{j=1,\ldots,K} Q_{\frac{1}{2}}(|\boldsymbol{d}_j^\top \boldsymbol{s}_\ell|, \xi).$$

Similarly to (5.12), $Q_{\frac{1}{2}}(\sqrt{\lambda}, \xi)$ is the generalized Marcum-Q function, which is increasing w.r.t. $\sqrt{\lambda} = |\boldsymbol{d}_j^\top \boldsymbol{s}_\ell|$. We obtain

$$\mathrm{P}_{\mathrm{Det}}(\boldsymbol{s}_\ell, \boldsymbol{D}) \geq Q_{\frac{1}{2}}\left(\max_{j=1,\ldots,K} |\boldsymbol{d}_j^\top \boldsymbol{s}_\ell|, \xi\right).$$

Let $\boldsymbol{s}_{\boldsymbol{D}}^* := \arg\min_{\boldsymbol{s}_\ell} \mathrm{P}_{\mathrm{Det}}(\boldsymbol{s}_\ell, \boldsymbol{D})$ denote one of the alternatives in $\boldsymbol{S}$ that is the most poorly correlated to the columns of dictionary $\boldsymbol{D}$. The associated probability of detection can then be lower bounded by

$$\mathrm{P}_{\mathrm{Det}}(\boldsymbol{s}_{\boldsymbol{D}}^*, \boldsymbol{D}) \geq Q_{\frac{1}{2}}\left(\max_{j=1,\ldots,K} |\boldsymbol{d}_j^\top \boldsymbol{s}_{\boldsymbol{D}}^*|, \xi\right) \geq Q_{\frac{1}{2}}\left(\rho^{(K)}(\boldsymbol{D}), \xi\right) \tag{5.21}$$

where

$$\rho^{(K)}(\boldsymbol{D}) = \min_{i=1,\ldots,L} \max_{j=1,\ldots,K} |\boldsymbol{d}_j{}^\top \boldsymbol{s}_i| \tag{5.22}$$

denotes the minimax correlation of dictionary $\boldsymbol{D}$ with set $\mathcal{S}$.

Instead of solving (5.6) or equivalently maximizing the left term of (5.21) w.r.t. $\boldsymbol{D}$, we propose to maximize the rightmost term, or equivalently $\rho^{(K)}(\boldsymbol{D})$. A possible strategy for a minimax learning algorithm (see Section 6.2) is then to construct the dictionary $\boldsymbol{D} = [\boldsymbol{d}_1, \ldots, \boldsymbol{d}_K]$, where $\|\boldsymbol{d}_j\|_2^2 = 1$, $\{j = 1, \ldots, K\}$, so that the minimum correlation $\rho^{(K)}(\boldsymbol{D})$ is made "as large as possible". The learning algorithm should thus produce a value $\rho^{(K)}$ that increases rapidly with K (see Figure 7.4(d) for an example with $K = 70$ atoms). In this case, the approximation criterion for the optimal minimax dictionary is

$$\boldsymbol{D}^* \approx \arg\max_{\boldsymbol{D}:\|\boldsymbol{d}_j\|_2=1} \min_{i=1,\ldots,L} \max_{j=1,\ldots,K} |\boldsymbol{d}_j{}^\top \boldsymbol{s}_i|. \tag{5.23}$$

Note that if $K = 1$, $\rho^{(1)}(\boldsymbol{d}) = \min_{i=1,\ldots,L} |\boldsymbol{d}^\top \boldsymbol{s}_i|$ and (5.23) is indeed equivalent to (5.13).

5.2.4.2 Study of false alarm rate for $K > 1$

We now study the effect of the dictionary size K on the probability of false alarm. A tight upper bound can be obtained by applying the inequality of the following theorem:

Theorem 1. *CDF of multivariate normals [Khatri 1968].*
$\mathbb{P}(|v_i| \leq c_i, i = 1, \ldots, m) \geq \prod_{i=1}^{m} \mathbb{P}(|v_i| \leq c_i)$ *provide that* $\boldsymbol{v} = (v_1, \ldots, v_m)^\top$ *is distributed as multivariate normal* $\boldsymbol{v} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$.

From (5.4), we can write, noting $\boldsymbol{v} = \boldsymbol{D}^\top \boldsymbol{x}$,

$$\mathrm{P_{FA}}(\boldsymbol{D}) = 1 - \mathbb{P}(|v_1| < \xi, \ldots, |v_K| < \xi), \tag{5.24}$$

where $\boldsymbol{v} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \boldsymbol{D}^\top \boldsymbol{D}$ and $\boldsymbol{\Sigma}_{j,j} = \boldsymbol{d}_j^\top \boldsymbol{d}_j = 1$. Applying Theorem 1 bounds the probability of false alarm (5.24) to

$$\mathrm{P_{FA}}(\boldsymbol{D}) \leq 1 - \prod_{j=1}^{K} \mathbb{P}(|u_j| \leq \xi)$$

which finally yields

$$\mathrm{P_{FA}}(\boldsymbol{D}) \leq 1 - \Phi_{\chi_1^2}^K(\xi^2). \tag{5.25}$$

If $\boldsymbol{D}$ is orthogonal, the upper bound on $\mathrm{P_{FA}}(\boldsymbol{D})$ is the exact $\mathrm{P_{FA}}(\boldsymbol{D})$. In our framework, $\boldsymbol{D}$ is not restricted to be orthogonal. However, the learned minimax atoms represent the outliers of the set $\mathcal{S}$ which, by definition, are marginally correlated. So we expect the columns of $\boldsymbol{D}$ to be weakly correlated to each other, and the bound (5.25) to be reasonably tight.

Based on the study of Section 5.2, we conclude that the minimax learning problem can be solved by the exact solution (5.13) in the case where a single minimax atom $\boldsymbol{d}^*$ has to be obtained. For an arbitrary number of K atoms, analytical solutions do not appear feasible. We will however see that algorithms can be designed according to the approximation criterion (5.23) to obtain a minimax dictionary $\boldsymbol{D}^* \in \mathbb{R}^{N \times K}$. Chapter 6 is dedicated to the design of learning algorithms in this case.

Before moving further, we present below another possible reduced model (5.26), which does not impose a sparsity constraint under $\mathcal{H}_1$.

5.3 An alternative: unconstrained reduced model

$$\begin{cases} \mathcal{H}_0 & : \boldsymbol{x} = \boldsymbol{n}, \qquad\qquad \boldsymbol{n} \sim \mathcal{N}(\boldsymbol{0}, \mathbf{I}) \\ \mathcal{H}_1 & : \boldsymbol{x} = \boldsymbol{D}\boldsymbol{\beta} + \boldsymbol{n}, \end{cases} \tag{5.26}$$

where $\boldsymbol{D} \in \mathbb{R}^{N \times K} = [\boldsymbol{d}_1, \ldots, \boldsymbol{d}_K]$, $\|\boldsymbol{d}_j\|_2^2 = 1$, $\{j = 1, \ldots, K\}$, $K < N$, and $\boldsymbol{\beta} \in \mathbb{R}^K$. The GLR test for (5.26) is :

$$T_{\mathrm{GLR}}(\boldsymbol{x}, \boldsymbol{D}) : \max_{\boldsymbol{\beta}} \frac{p(\boldsymbol{x}; \boldsymbol{D}\boldsymbol{\beta})}{p(\boldsymbol{x}; \boldsymbol{0})} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \xi' \tag{5.27}$$

where under $\mathcal{H}_1$

$$p(\boldsymbol{x}\,; \boldsymbol{D}\boldsymbol{\beta}) = \frac{1}{(2\pi)^{N/2}} \exp\left(-\frac{1}{2}\|\boldsymbol{x} - \boldsymbol{D}\boldsymbol{\beta}\|_2^2\right). \tag{5.28}$$

Maximizing the numerator of (5.27) conducts to

$$\begin{aligned} \boldsymbol{\beta}^{\mathrm{ML}} &= \arg\min_{\boldsymbol{\beta}} \ \|\boldsymbol{x} - \boldsymbol{D}\boldsymbol{\beta}\|_2^2 \\ &= \arg\min_{\boldsymbol{\beta}} \ \|\boldsymbol{x}\|_2^2 - 2\boldsymbol{\beta}^\top \boldsymbol{D}^\top \boldsymbol{x} + \boldsymbol{\beta}^\top \boldsymbol{D}^\top \boldsymbol{D}\boldsymbol{\beta}, \end{aligned} \tag{5.29}$$

which leads to

$$\frac{\partial(-2\boldsymbol{\beta}^\top \boldsymbol{D}^\top \boldsymbol{x} + \boldsymbol{\beta}^\top \boldsymbol{D}^\top \boldsymbol{D}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0 \Rightarrow \ \boldsymbol{\beta}^{\mathrm{ML}} = (\boldsymbol{D}^T \boldsymbol{D})^{-1} \ \boldsymbol{D}^T \boldsymbol{x}. \tag{5.30}$$

Substituting $\boldsymbol{\beta}^{\mathrm{ML}}$ in the GLR test (5.27) and taking the logarithm conduct to

$$\begin{aligned} -\frac{1}{2}\left(\|\boldsymbol{x}\|_2^2 - 2(\boldsymbol{\beta}^{\mathrm{ML}})^\top \boldsymbol{D}^\top \boldsymbol{x} + (\boldsymbol{\beta}^{\mathrm{ML}})^\top \boldsymbol{D}^\top \boldsymbol{D}\boldsymbol{\beta}^{\mathrm{ML}}\right) + \frac{1}{2}\|\boldsymbol{x}\|_2^2 &\underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \ln \xi' \\ 2((\boldsymbol{D}^T \boldsymbol{D})^{-1} \ \boldsymbol{D}^T \boldsymbol{x})^\top \boldsymbol{D}^\top \boldsymbol{x} - ((\boldsymbol{D}^T \boldsymbol{D})^{-1} \ \boldsymbol{D}^T \boldsymbol{x})^\top \boldsymbol{D}^\top \boldsymbol{D}(\boldsymbol{D}^T \boldsymbol{D})^{-1} \ \boldsymbol{D}^T \boldsymbol{x} &\underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} 2\ln \xi' \\ \boldsymbol{x}^\top \boldsymbol{D}(\boldsymbol{D}^T \boldsymbol{D})^{-1} \boldsymbol{D}^T \boldsymbol{x} &\underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} 2\ln \xi'. \end{aligned}$$

This finally leads to

$$T_{\mathrm{Alt}}(\boldsymbol{x}, \boldsymbol{D}) = \boldsymbol{x}^\top \Pi_{\boldsymbol{D}} \, \boldsymbol{x} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \xi \Leftrightarrow \|\Pi_{\boldsymbol{D}} \, \boldsymbol{x}\|_2^2 \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \xi, \tag{5.31}$$

where $\xi = 2\ln \xi'$ and $\Pi_{\boldsymbol{D}} = \boldsymbol{D}(\boldsymbol{D}^\top \boldsymbol{D})^{-1} \ \boldsymbol{D}^\top$ is an orthogonal projection matrix. The test (5.31) is an energy detector where the data vector $\boldsymbol{x}$ is projected on the orthogonal subspace of dictionary $\boldsymbol{D}$.

The probability of false alarm and the probability of detection of test (5.31) writes

$$\mathrm{P}_{\mathrm{FA}}(\boldsymbol{D}) = \mathbb{P}\left(\boldsymbol{x}^\top \Pi_{\boldsymbol{D}}\, \boldsymbol{x} > \xi; \mathcal{H}_0, \boldsymbol{D}\right), \tag{5.32}$$

$$\mathrm{P}_{\mathrm{Det}}(\boldsymbol{s}_\ell, \boldsymbol{D}) = \mathbb{P}\left(\boldsymbol{x}^\top \Pi_{\boldsymbol{D}}\, \boldsymbol{x} > \xi; \mathcal{H}_1, \boldsymbol{s}_\ell, \boldsymbol{D}\right). \tag{5.33}$$

Following the minimax criterion and definitions (5.32)-(5.33), the optimization problem for model (5.26) can be formulated as

$$\begin{aligned} \boldsymbol{D}^{*\mathrm{Alt}} := \arg\max_{\boldsymbol{D}} \quad & \min_{i=1,\ldots,L} \mathrm{P}_{\mathrm{Det}}(\boldsymbol{s}_i, \boldsymbol{D}) \\ \text{subject to} \quad & \mathrm{P}_{\mathrm{FA}}(\boldsymbol{D}) \leq \mathrm{P}_{\mathrm{FA}_0}, \\ & \|\boldsymbol{d}_j\|_2 = 1,\ j = \{1, \ldots, K\}, \end{aligned} \tag{5.34}$$

where $\mathrm{P}_{\mathrm{FA}_0}$ is a fixed probability of false alarm.

Similar to the previous analysis (Section 5.2.2), let us assume that under $\mathcal{H}_1$, the target active signal is $\boldsymbol{s}_\ell$, and the corresponding amplitude $\alpha_\ell = 1$. Then by (4.1), $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{s}_\ell, \mathbf{I})$ under $\mathcal{H}_1$. Since the projection matrix $\Pi_{\boldsymbol{D}}$ is idempotent ($\Pi_{\boldsymbol{D}}\Pi_{\boldsymbol{D}} = \Pi_{\boldsymbol{D}}$), symmetric ($\Pi_{\boldsymbol{D}}^T = \Pi_{\boldsymbol{D}}$) and of rank K, the distributions of the test statistics $\boldsymbol{x}^T \Pi_{\boldsymbol{D}}\, \boldsymbol{x}$ in (5.31) under $\mathcal{H}_1$ can be obtained by Proposition 4, which gives

$$T_{\mathrm{Alt}}(\boldsymbol{x}; \mathcal{H}_1) = (\boldsymbol{s}_\ell + \boldsymbol{n})^\top \Pi_{\boldsymbol{D}}\, (\boldsymbol{s}_\ell + \boldsymbol{n}) \sim \chi^2_{K,\lambda}, \tag{5.35}$$

where the non centrality parameter is here $\lambda(\boldsymbol{s}_\ell, \boldsymbol{D}) = \boldsymbol{s}_\ell^\top \Pi_{\boldsymbol{D}}\, \boldsymbol{s}_\ell = \|\Pi_{\boldsymbol{D}}\ \boldsymbol{s}_\ell\|_2^2$. Thus, Proposition 2 implies

$$\mathrm{P}_{\mathrm{Det}}(\boldsymbol{s}_\ell, \boldsymbol{D}) = \mathbb{P}(T_{\mathrm{Alt}} > \xi; \mathcal{H}_1) = 1 - \Phi_{\chi^2_{K,\lambda}}(\xi) = Q_{\frac{K}{2}}(\sqrt{\lambda}, \sqrt{\xi}), \tag{5.36}$$

where $\Phi_{\chi^2_{K,\lambda}}$ denote the corresponding CDFs, and Q the generalized Marcum Q-function. Since $Q_{\frac{K}{2}}(\sqrt{\lambda}, \sqrt{\xi})$ is an increasing function of λ, maximizing the probability of detection is equivalent to maximizing λ. The minimax optimization (5.34) becomes

$$\boldsymbol{D}^{*\mathrm{Alt}} = \arg\max_{\boldsymbol{D}:\|\boldsymbol{d}_j\|_2^2=1} \min_{i=1,\ldots,L} \mathrm{P}_{\mathrm{Det}}(\boldsymbol{s}_i, \boldsymbol{D}) \Leftrightarrow \arg\max_{\boldsymbol{D}:\|\boldsymbol{d}_j\|_2^2=1} \min_{i=1,\ldots,L} \|\Pi_{\boldsymbol{D}}\ \boldsymbol{s}_i\|_2^2. \tag{5.37}$$

Here, the minimax optimization problem depends on the plane $\Pi_{\boldsymbol{D}}$. If we write problem (5.37) in an epigraph form, we have

$$\begin{aligned} \boldsymbol{D}^{*\mathrm{Alt}} = \quad & \text{minimize} \quad -t \\ & \text{subject to} \quad t - \|\Pi_{\boldsymbol{D}}\ \boldsymbol{s}_i\|_2 \leq 0,\ i = \{1, \ldots, L\}\ , \\ & \qquad\qquad \|\boldsymbol{d}_j\|_2 \leq 1,\ j = \{1, \ldots, K\} \end{aligned} \tag{5.38}$$

which is a non convex function (for $K > 1$) because the constraint $t - \|\Pi_{\boldsymbol{D}}\ \boldsymbol{s}_i\|_2 \leq 0$ is non linear.

Under $\mathcal{H}_0$ of (4.1), $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Following the computation of the probability of detection above, the distributions of the test statistics $\boldsymbol{x}^T \Pi_{\boldsymbol{D}} \boldsymbol{x}$ in (5.31) under $\mathcal{H}_0$ leads to

$$T_{\text{Alt}}(\boldsymbol{x}; \mathcal{H}_0) = \boldsymbol{n}^\top \Pi_{\boldsymbol{D}} \boldsymbol{n} \sim \chi^2_K. \tag{5.39}$$

Then, the probability of false alarm can be written as function of the corresponding CDF, denoted by $\Phi_{\chi^2_K}$

$$\text{P}_{\text{FA}} = \mathbb{P}(T_{\text{Alt}} > \xi \,; \mathcal{H}_0) = 1 - \Phi_{\chi^2_K}(\xi). \tag{5.40}$$

For this model, the probability of false alarm is independent of $\boldsymbol{D}$ ($\forall \boldsymbol{D}$: $\text{rank}(\boldsymbol{D}) = K$).

In the case where $\boldsymbol{D}^{*\text{Alt}}$ has single atom ($K = 1$), the orthogonal projection matrix on the subspace of $\boldsymbol{D}$ becomes $\Pi_{(\boldsymbol{D}=\boldsymbol{d})} = \boldsymbol{d}(\boldsymbol{d}^\top \boldsymbol{d})^{-1}\boldsymbol{d}^\top = \boldsymbol{d}\boldsymbol{d}^\top$ (as $\boldsymbol{d}$ has normalized columns). $\boldsymbol{d}^{*\text{Alt}}$ is equivalent to the solution (5.13) found for the previous reduced model (5.1).

For the first RD model (5.1), the 1-sparse constraint ($\|\boldsymbol{\beta}\|_0 = 1$) under the alternative hypothesis means that we are working on the axes, one dimension at each instance. This model is close to the exact model (4.1) which uses $\boldsymbol{S}$. For the second RD model (5.26), we work on the hyperplanes where we want to maximize the energy.

It is not clear which RD model should be the best, but imposing $\|\boldsymbol{\beta}\|_0 = 1$ under $\mathcal{H}_1$ allowed to find greedy solution easily (which is not the case for the second RD model).

5.4 Discussion

The main (theoretical) contributions of this dissertation were presented in this Chapter. Following the study of detection problem investigated in Chapter 4, we identified the necessity to devise robust detection tests that operate in subspaces of reduced dimension while still maintaining good performances in worst-case scenarios.

Studies done in the present Chapter produced an exact solution to one-dimensional minimax (worst-case) optimization problem (5.13). We have seen in Example 5.2.1 that this approach improved the worst-case performance w.r.t. the classical low-rank approximation (SVD). Such an extreme reduction in dimension might however be exaggerated w.r.t. the intrinsic diversity of the reference library $\boldsymbol{S}$ (see Figure 7.4(c)). With this in mind, we investigated an approximation to the exact minimax optimization problem for an arbritary number of learned columns K, for the first reduced model. As a follow-up of these analyses, we propose in the next Chapter (6) two learning algorithms aimed at building $\boldsymbol{D}^*$ using the minimax correlation function (5.22), or the approximation criterion (5.23).

An alternative RD model (5.26) without sparsity constraint under $\mathcal{H}_1$ was also investigated in this Chapter which yields the same optimization problem as the previous one for $K = 1$. Between the two minimax RD models, it is not clear which one should be the best. However imposing $\|\boldsymbol{\beta}\|_0 = 1$ under $\mathcal{H}_1$ in the first RD model (5.1) allowed to find greedy solution easily (which is not the case for the alternative RD model). We have not pursued our investigations on the alternative RD model because the algorithms based on the first model proposed in the next Chapter give satisfactory results. Furthermore, the first RD model is closer to the true model (4.1).

In Chapter 6 the one-dimensional minimax atom $\boldsymbol{d}^*$ will be the basis of three minimax learning algorithms. The first algorithm functions in a greedy approach based on the analysis of Section 5.2.4. As a second algorithm, a variant arises from the general strategy of injecting minimax objectives in standard dictionary learning algorithms. The third algorithm combines clustering techniques found in the literature with the exact solution of one-dimensional minimax optimization problem.

Chapter 6

Minimax learning techniques of an arbitrary size dictionary

Contents

Some analyses and results presented in this Chapter were published in [Suleiman *et al.* 2013b, Suleiman *et al.* 2014a, Suleiman *et al.* 2014b].

6.1 Introduction

This Chapter proposes several minimax (or worst-case) learning algorithms for an arbitrary size dictionary according to problem (5.6). The first algorithm (Section 6.2) is a greedy type minimax learning algorithm based on the analysis of Section 5.2.4. The second algorithm involves a general strategy of incorporating a minimax criterion in standard dictionary learning algorithms, specifically in the dictionary update stage using (5.13)-(5.14). In this regard, we propose a variant of K-SVD algorithm, or more precisely in our setting, a variant of the gain-shape Vector Quantization [Gersho & Gray 1991]. We call the resulting algorithm K-minimax (Section 6.3). As a third algorithm, we propose a possible variant for worst-case learning approach, where we combine clustering techniques existing in literature (such as the Spherical K-means [Dhillon & Modha 2001]) with one-dimensional minimax solution (5.13)-(5.14).

Aside from these worst-case learning techniques, we also study a possible gradient descent type method to solve the one-dimensional minimax optimization problem (without using quadratic programming). We investigate this gradient descent method hoping to generalize it to the case $K > 1$. This approach is however not fully worked out. A description and related issues are provided in Appendix B.3.

Note that we use the term *clusters* or *classes* interchangeably in this Chapter. Both terms refer to the *groups* to which the alternatives belong, regardless of the methods defined in Chapter 3.

6.2 Greedy minimax: a heuristic approach

Following the analytical analysis in Section 5.2.4, we propose a heuristic optimization to learn a minimax dictionary. This algorithm (see the pseudo-code in Algorithm 1 below) samples the distribution in a greedy manner to open new cluster, and the dictionary update stage for each cluster is done through the one-dimensional minimax solution (5.13)-(5.14).

The approach is described below, and illustrated in Figure 6.1 for $K = 3$:

i). Compute through (5.13) the global minimax atom $\boldsymbol{d}^*$ (red star) representing the whole set of alternatives in $\boldsymbol{S}$ (black dots).

ii). Identify the alternative $\boldsymbol{s}_{i^*}$ that is the most poorly represented by the dictionary (i.e., of minimum correlation) (white dot). If there are several, pick one at random. 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 $\mathcal{S}$ is then classified into $j = 2$ classes $(\mathbf{C}_1^*, \mathbf{C}_2^*)$ by nearest neighbor rule, and one atom $\boldsymbol{d}^*$ is generated through (5.13) for each cluster, representing the updated learned dictionary columns (red stars).

iii). A new class is opened using one of the farthest alternatives to the current columns. Nearest neighbor rule results in three new classes whose minimax centers constitute the final dictionary $\boldsymbol{D}_3^*$.

In regard to clustering techniques in literature, greedy minimax learning algorithm can be viewed as a special case of Divisive clustering (top to bottom approach), where all data are initially grouped in a cluster an then split into many clusters. In our case however, the number of clusters (equivalent to the number of atoms K) is fixed at initialization.

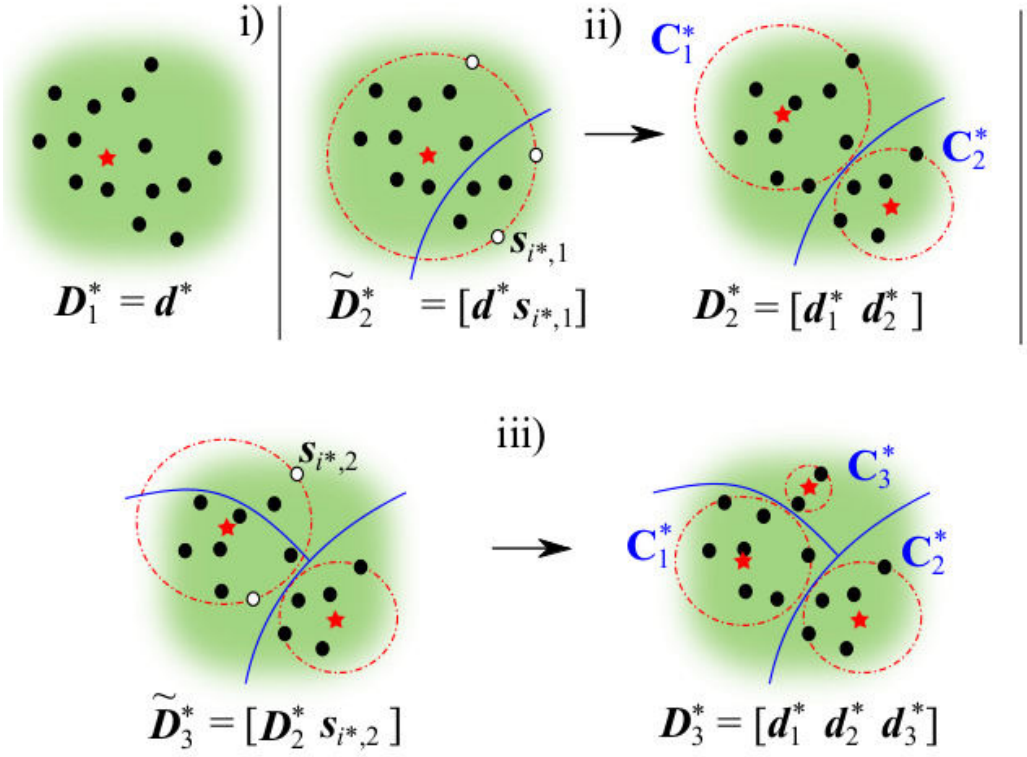


Figure 6.1: Illustration of the greedy minimax algorithm for $K = 3$. Black dots: alternatives $\boldsymbol{s}_i$ on the unit sphere, red stars: minimax atoms, white dots: the least correlated alternative w.r.t. $\boldsymbol{D}^*$. The blue lines delimitate the classes. After initialization i), the farthest alternative $\boldsymbol{s}_{i^*}$ w.r.t. $\boldsymbol{D}_1^*$ is identified and the alternatives are divided into two clusters. Minimax atoms for each cluster are computed, its concatenation forms $\boldsymbol{D}_2^*$. These processes are continued in sequences, until obtaining the desired K atoms.

Algorithm 1 Greedy minimax

Inputs: Data set $\boldsymbol{S} \in \mathbb{R}^{N\times L} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_L]$, number of atoms K.

Initialization: $j = 1$, $\boldsymbol{D}_j^* = \boldsymbol{d}^* \in \mathbb{R}^N$ as obtained in (5.13).
Set: $j = 2$,
$\boldsymbol{s}_{i^*,j-1} = \arg \min_{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_L} |\boldsymbol{d}^{*\top} \boldsymbol{s}_i|$,
$\widetilde{\boldsymbol{D}}_j^* = [\boldsymbol{d}^* \ \boldsymbol{s}_{i^*,j-1}]$,

while $j \leq K$ **do**

- *Clustering stage*:

  for $i = 1, \ldots, L$
  Assign $\boldsymbol{s}_i$ to the class of atom $\widetilde{\boldsymbol{d}}_l^*$ of $\widetilde{\boldsymbol{D}}_j^*$ if
  $|\boldsymbol{s}_i^\top \widetilde{\boldsymbol{d}}_l^*| > |\boldsymbol{s}_i^\top \widetilde{\boldsymbol{d}}_n^*|, \forall n \neq l$.
  end

This yields j clusters $\mathbf{C}_l^*$, $l = 1, \ldots, j$.

- *Dictionary update stage*:

for $l = 1, \ldots, j$
$\boldsymbol{d}_l^* = \arg \max_{\boldsymbol{d}:\|\boldsymbol{d}\|_2=1} \min_{\boldsymbol{s}_i \in \mathbf{C}_l^*} |\boldsymbol{d}^\top \boldsymbol{s}_i|$.
end
$\boldsymbol{D}_j^* = [\boldsymbol{d}_1^*, \ldots, \boldsymbol{d}_j^*]$,
$\boldsymbol{s}_{i^*,j} = \arg\min_{\boldsymbol{s}_i} \|\boldsymbol{D}_j^{*\top} \boldsymbol{s}_i\|_\infty$,
$\widetilde{\boldsymbol{D}}_{j+1}^* = [\boldsymbol{D}_j^* \ \boldsymbol{s}_{i^*,j}]$,
$j = j + 1$.
end while

Output: Greedy minimax dictionary of K atoms: $\boldsymbol{D}_K^* = \boldsymbol{D}_j^*$.

6.3 K-minimax: a variant of K-SVD approach

Here, we focus on dictionary learning for dimension reduction of some known library $\boldsymbol{S}$. As an example of an algorithm that can be useful for such a task, we focus on the classical K-SVD algorithm [Aharon *et al.* 2006] (the pseudo-code is given in Appendix A.5.3). K-SVD would in our notations optimize the dictionary $\boldsymbol{D}$ by finding an approximate solution of

$$\min_{\boldsymbol{D},\boldsymbol{A}} \|\boldsymbol{S} - \boldsymbol{D}\boldsymbol{A}\|_F^2 \quad \text{subject to } \forall i \in [1, L],\ \|\boldsymbol{\alpha}_i\|_0 \leq k,$$

where $\boldsymbol{A}$ is an unknown sparse representation matrix. In our framework, we set the K-SVD sparsity parameter t_0 to 1, in agreement with the unit ℓ_0 pseudonorm considered in the test (5.1)-(5.3). This encourages each $\boldsymbol{s}_i$, $\{i = 1, \ldots, L\}$, to be well represented by at least one column of $\boldsymbol{D}$. The K-SVD algorithm in this setting is equivalent to gain-shape VQ [Gersho & Gray 1991].

In Chapter 3, we have discussed some fundamental aspects on sparse dictionary learning. Most often, they rely on a MSE criterion, which may not be suitable for a minimax approach (e.g., see Examples 5.2.1 and 6.4.1). Hence, we propose to replace the SVD dictionary update step of each class by the minimax optimization (5.13) applied to the alternatives of the considered class. Figure 6.2 depicts this approach:

i). Start with an initial dictionary of K atoms.

ii). Sparse coding stage: Identify the most correlated atom of the dictionary to each $\boldsymbol{s}_i$. Divide accordingly the set $\mathcal{S}$ into K clusters $\mathbf{C}_j^{\mathrm{K}*}$, $\{j = 1, \ldots, K\}$, by nearest neighbor rule (largest correlation).

iii). Minimax dictionary update : $\boldsymbol{d}^*$ is computed for each class $\mathbf{C}_j^{\mathrm{K}*}$ by (5.13), resulting in a minimax centroid $\boldsymbol{D}_j^{\mathrm{K}*}$.

The steps ii) and iii) are repeated until convergence or a stopping rule, as in K-SVD. The final dictionary is noted $\boldsymbol{D}_K^{\mathrm{K}*}$. The pseudo-code of K-minimax is given in Algorithm 2.

This algorithm can for instance be initialized by K samples randomly chosen among $\mathcal{S}$, or by first computing the global ($K = 1$) minimax atom $\boldsymbol{d}^*$, and then selecting $K - 1$ atoms among the atoms that are less correlated to $\boldsymbol{d}^*$. We have found that the latter initialization samples better the marginal alternatives. This will be the initialization used for the applications in Chapter 7.

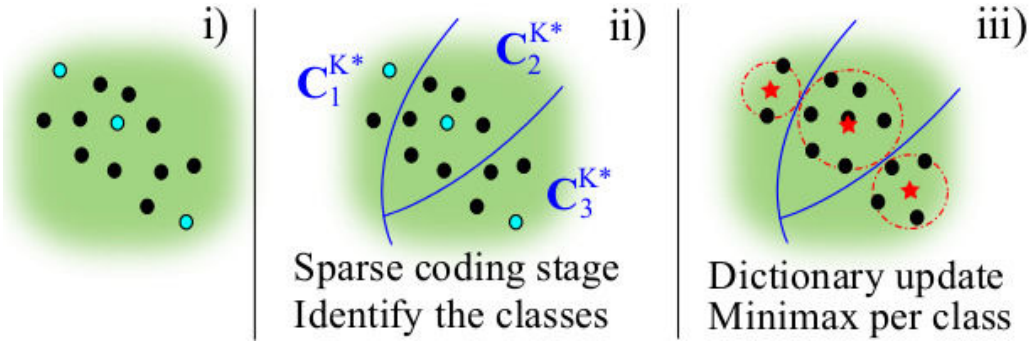


Figure 6.2: Illustration of the K-minimax algorithm for $K = 3$. Black dots : alternatives $\boldsymbol{s}_i$ on the unit sphere, cyan dots: initial dictionary atoms. The blue lines delimitate the classes and the red stars are the minimax atoms for each class. After initialization i), the algorithm iterates between steps ii) and iii). These are done until a stopping rule. We obtain a "K-minimax dictionary" of K atoms.

Algorithm 2 K-minimax

Inputs: Data set $\boldsymbol{S} \in \mathbb{R}^{N \times L} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_L]$, number of atoms K.

Initialization:

Choose an initial dictionary $\boldsymbol{D}_K^{\mathrm{K}*} \in \mathbb{R}^{N \times K}$.
Set: $k = 1$, $\boldsymbol{D}_K^{\mathrm{K}*(1)} = \boldsymbol{D}_K^{\mathrm{K}*}$.

Repeat until convergence (stopping rule):

- *Sparse Coding (Clustering) stage*:
  for $i = 1, \ldots, L$
  Assign $\boldsymbol{s}_i$ to the class of atom $\boldsymbol{d}_j^{\mathrm{K}*(k)}$ of $\boldsymbol{D}_K^{\mathrm{K}*(k)}$ if
  $|\boldsymbol{s}_i^\top \boldsymbol{d}_j^{\mathrm{K}*(k)}| > |\boldsymbol{s}_i^\top \boldsymbol{d}_n^{\mathrm{K}*(k)}|, \forall n \neq j$.
  end
  This yields K clusters $\mathbf{C}_j^{\mathrm{K}*(k)}$, $j = 1, \ldots, K$.

- *Dictionary update stage*:
  Update dictionary column $\boldsymbol{d}_j^{\mathrm{K}*(k)}$ for each class $\mathbf{C}_j^{\mathrm{K}*(k)}$:
  for $j = 1, \ldots, K$
  $$\boldsymbol{d}_j^{\mathrm{K}*(k)} = \arg \max_{\boldsymbol{d}:\|\boldsymbol{d}\|_2=1} \min_{\boldsymbol{s}_i \in \mathbf{C}_j^{\mathrm{K}*(k)}} |\boldsymbol{d}^\top \boldsymbol{s}_i|.$$
  end

$\boldsymbol{D}_K^{\mathrm{K}*(k+1)} = [\boldsymbol{d}_1^{\mathrm{K}*(k)}, \ldots, \boldsymbol{d}_K^{\mathrm{K}*(k)}]$,
$k = k + 1$.

Output: K-minimax dictionary of K atoms: $\boldsymbol{D}_K^{\mathrm{K}*} = \boldsymbol{D}_K^{\mathrm{K}*(k-1)}$.

6.4 Clustering technique combined with 1D minimax

6.4.1 SKM-minimax

Apart from the learning algorithms proposed above, we present here another possible approach to learn minimax dictionary. We combine clustering technique on the unit hypersphere from the literature with the one-dimensional minimax solution. We exemplify this approach here with the Spherical K-Means (SKM) clustering [Dhillon & Modha 2001]. The dictionary obtained from this "SKM clustering-1D minimax" combination is denoted as $\boldsymbol{D}^{\mathrm{SKM}*}$.

This approach results in a simple two-step method without any iteration:

i). Partition the data samples $\boldsymbol{S}$ into K clusters using the SKM algorithm (presented below in Section 6.4.2).

ii). For each cluster j, $j = 1, \ldots, K$ compute the corresponding minimax atom $\boldsymbol{d}_j^{\mathrm{SKM}*}$. The concatenation of these K minimax atoms forms the final dictionary $\boldsymbol{D}_K^{\mathrm{SKM}*} = [\boldsymbol{d}_1^{\mathrm{SKM}*}, \ldots, \boldsymbol{d}_K^{\mathrm{SKM}*}]$.

Example 6.4.1 illustrates the atoms learned using this approach, and compares them to those learned using the above minimax proposed approaches and to K-SVD dictionary. In Section 7.3.1, we will show numerical comparison of RD detection tests using $\boldsymbol{D}^{\mathrm{SKM}*}$.

6.4.2 Spherical K-Means

Spherical K-Means is a variant of the K-Means method, which uses *cosine similarity* instead of minimizing the (squared) Euclidean distance. Assume that we have two ℓ_2 normalized vectors $\boldsymbol{s}$ and $\boldsymbol{c} \in \mathbb{R}_+^N$, then the angle θ between these two vectors is defined $0 \leq \theta(\boldsymbol{s}, \boldsymbol{c}) \leq \pi/2$. The cosine similarity is the inner product between these two vectors, that is

$$\boldsymbol{s}^\top \boldsymbol{c} = \|\boldsymbol{s}\| \|\boldsymbol{c}\| \cos(\theta(\boldsymbol{s}, \boldsymbol{c})) = \cos(\theta(\boldsymbol{s}, \boldsymbol{c})). \tag{6.1}$$

The centroid of each cluster is the nearest in cosine similarity with all the members belonging to its cluster. SKM is a heuristic approach that approximates the solution to the following optimization problem [Dhillon & Modha 2001]

$$\{C_j^{\mathrm{SKM}}\}_{j=1}^K = \arg\max_{\{C_j\}_{j=1}^K} \sum_{j=1}^K \sum_{\boldsymbol{S} \in C_j} \boldsymbol{s}^\top \boldsymbol{c}_j, \tag{6.2}$$

where $\boldsymbol{s} \in \boldsymbol{S}$ are ℓ_2 normalized data samples to be partitioned, $\{C_j\}_{j=1}^K$ are K disjoint clusters ($C_j \cap C_k = \emptyset$ if $j \neq k$ and $\bigcup_{j=1}^K C_j = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_L\}$) and $\boldsymbol{c}_j$ is the normalized mean centroids (also called the *concept vectors*) associated to each cluster.

SKM may sometimes yield empty cluster(s) for some values of K. This happens when no data sample is assigned to the corresponding cluster(s). As a result, we could not build $\boldsymbol{D}_K^{\mathrm{SKM}*}$ for some values of K. In such cases, SKM-minimax will not be evaluated in the main applicative part (Chapter 7) of this dissertation.

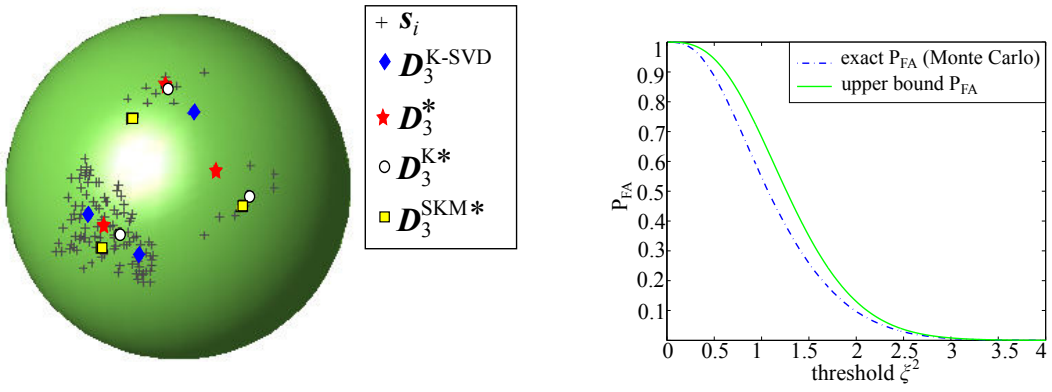
Example 6.4.1. Comparison of four dictionary learning methods for $K = 3$.

This is an extension of the previous Example 5.2.1 concerning a simple numerical experiments using random positives alternatives forming a reference library $\boldsymbol{S} \in \mathbb{R}^{3\times 115}$. Here, we will learn more than one atoms ($K = 3$.) We compare the corresponding atoms learned using four learning algorithms. The first, denoted by $\boldsymbol{D}_3^{\text{K-SVD}}$, is obtained by the standard K-SVD approach (in which the sparsity $k = 1$), and the other three ($\boldsymbol{D}_3^{*}$, $\boldsymbol{D}_3^{\text{K}*}$ and $\boldsymbol{D}_3^{\text{SKM}*}$) are respectively the greedy minimax, the K-minimax and the SKM-minimax approaches described above.

Figure 6.3(a) shows the alternatives (black crosses, which exhibit here three subpopulations or clusters) and the learning results of each method. Two of the K-SVD atoms (blue diamonds) lie within the main subpopulation of the alternatives. In contrast, all three of the minimax approaches: greedy minimax (red stars), K-minimax (white circles) and SKM-minimax (yellow squares) approaches have one atom in or close to each of the three clusters. The minimax approaches yield nearly similar results and on this example one of the K-minimax atoms coincides with one of the greedy minimax atoms.

Turning to the minimax correlations (5.22), which are the figures of merit retained for the worst-case optimization, we obtain $\rho^{(3)}(\boldsymbol{D}_3^{\text{K-SVD}}) = 0.797$, $\rho^{(3)}(\boldsymbol{D}_3^{*}) = 0.913$, $\rho^{(3)}(\boldsymbol{D}_3^{\text{K}*}) = 0.881$ and $\rho^{(3)}(\boldsymbol{D}_3^{\text{SKM}*}) = 0.873$. W.r.t. the case $K = 1$ (see Example 5.2.1), where $\rho^{(1)}(\boldsymbol{u}) = 0.475$ and $\rho^{(1)}(\boldsymbol{d}^*) = 0.714$, the minimax correlations are improved for $K > 1$, with indeed better scores for the minimax algorithms.

Figure 6.3(b) also compares the $\text{P}_{\text{FA}}(\boldsymbol{D}_3^{*})$ as estimated by Monte Carlo simulations with the bound (5.25) derived in Chapter 5 as a function of the threshold ξ^2. This numerical experiment shows that the exact P_{FA} is indeed upper bounded by (5.25). For small and large P_{FA}, these values are nearly the same and the bound is tight.



(a) (b)

Figure 6.3: (a) An example of the atoms learned by the proposed approaches (greedy minimax: red stars, K-minimax: white circles and SKM-minimax: yellow squares) and by K-SVD: blue diamonds on the unit sphere for $K = 3$. (b) Comparison of the $\text{P}_{\text{FA}}(\boldsymbol{D}_3^{*})$ by Monte Carlo simulation to the upper bound (5.25).

■

Example 6.4.2. Marginal alternatives as dictionary atoms for worst-case detection testing

In this example, we show the results of an intuitive approach in forming a worst-case dictionary. This approach uses the marginal alternatives directly as the dictionary atoms. This type of approach can be viewed as a direct Feature Selection method to dimension reduction. Feature Selection is a method that selects few data samples from the library w.r.t. a criterion and uses them as a reduced representatives set without applying any transformation (e.g., see [Guyon & Elisseeff 2003]).

One might think that since there is only a couple of alternatives in $\boldsymbol{S}$ that dominate the worst-case detection performances (see Figure 5.3), including these marginal alternatives in the dictionary should do a good job in improving the minimax performance. The present example investigates this strategy and shows that it is infact inefficient and yields very poor performance.

From a reference library of 100 spectral lines, $\boldsymbol{S} \in \mathbb{R}^{100\times 100}$, we learn the optimized minimax atom $\boldsymbol{d}^*$ (5.13). Then, we identify K alternatives that are the least correlated to $\boldsymbol{d}^*$. In this example, we choose $K = 3$. The concatenation of these 3 marginal alternatives forms a dictionary, named $\boldsymbol{D}_3^{\text{marginals}}$. Then, we train a greedy minimax dictionary of 3 atoms, denoted by $\boldsymbol{D}_3^*$.

We perform next a RD test of the form presented in Section 5.2.1, that is

$$T_{\boldsymbol{D}}(\boldsymbol{x}) = \max_{j=1,\ldots,K} (\boldsymbol{d}_j^\top \boldsymbol{x})^2 \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \xi^2,$$

where the $\{\boldsymbol{d}_j\}$ are K columns of the considered dictionaries ($\boldsymbol{D}_3^*$ and $\boldsymbol{D}_3^{\text{marginals}}$). The results are obtained by Monte Carlo simulations, where under $\mathcal{H}_0 : \boldsymbol{x} = \boldsymbol{n}$, under $\mathcal{H}_1 : \boldsymbol{x} = \boldsymbol{s}_i\alpha_i + \boldsymbol{n}$, with $\boldsymbol{n} \sim \mathcal{N}(\boldsymbol{0}, \mathbf{I})$, and $\alpha_i = 2.2, \forall i \in [1, 100]$. To study the worst-case detection performances, the tests are executed for all possible alternatives activated one by one under $\mathcal{H}_1$, i.e., $i = 1, \ldots, 100$.

Figure 6.4 depicts the numerical results of this test (evaluated by AUCs of the resulting ROC curves). We can see that testing using the greedy minimax dictionary $\boldsymbol{D}_3^*$ yields (far) better worst-case (and also average) detection performances than using the marginal dictionary $\boldsymbol{D}_3^{\text{marginals}}$.

The reason of this behavior is that in order to be efficient in worst-case scenarios, the algorithm should be able to separate the distribution of the alternatives into distinct classes and place representative (minimax) centroids for each class. Simply including the marginal alternatives in the dictionary is a too extreme way of accounting for marginal alternatives as it results in letting aside many alternatives of the core of the distribution.

Note finally that without the proposed method for computing the "center" $\boldsymbol{d}^*$, it may be difficult (apart from running exhaustive Monte Carlo simulations) to know a priori which alternatives are "marginals" when all $\boldsymbol{s}_i$ are correlated.

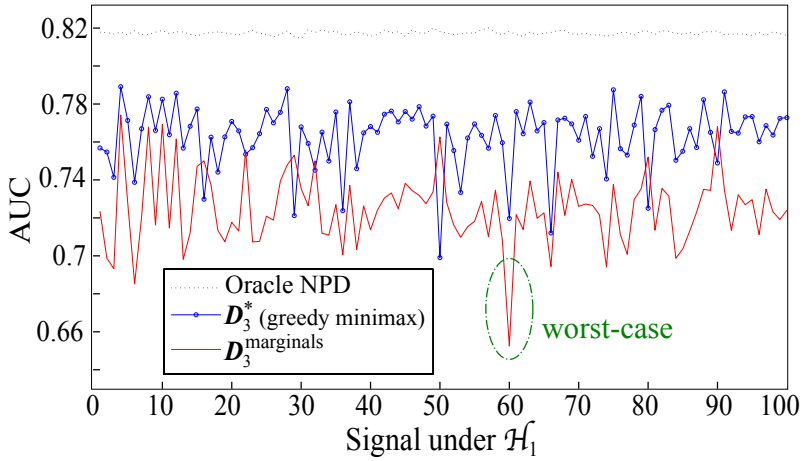


Figure 6.4: AUC of the ROCs over 100 alternatives activated one by one, under $\mathcal{H}_1$.

■

6.5 Discussion

Several minimax dictionary learning techniques were presented in this Chapter. Among them, we will be focusing on greedy minimax and K-minimax algorithms for the application part in the next Chapter.

Apart from the evaluation of detection performances of different tests, we discuss in Chapter 7 the problem of choosing the optimal number of K atoms w.r.t. the worst-case criterion in these learning process. Although there is no specific formulation derived, we will however outline a possible way to choose K.

In Chapter 8, we present further possible machine learning applications of algorithms presented in this Chapter, where we will learn minimax dictionaries of faces and apply the proposed algorithms for worst-case handwritten digits recognition.

Part III

Applications: Astrophysics and Machine Learning

The third Part of this manuscript highlights some possible applications and presents numerical results regarding the sparsity constrained GLR detection tests and minimax learning algorithms proposed in this dissertation. This final Part consists in two Chapters, each Chapter considering a different application framework.

In **Chapter 7**, we evaluate detection performances of a testing scheme using a specific reference library $\boldsymbol{S}$, and we compare the results with approaches using different learned dictionaries of reduced dimension. The considered reference library contains approximately 10000 spectral lines, provided by the MUSE (Multi Unit Spectroscopic Explorer) consortium. These spectral lines are obtained from highly realistic astrophysical simulations. The worst-case detection performances for each test are compared for tests using spectral profiles and spatio-spectral profiles. Furthermore, we will suggest an approach to determine the best number of K atoms for dictionary learning in our worst-case setting.

Chapter 8 investigates the interest of the approach for pattern recognition tasks. This Chapter shows two possible applications of the proposed minimax learning algorithms, for learning faces and for worst-case recognition of handwritten digits.

Chapter 7

An application in Astrophysics

Contents

Some results presented in this Chapter were published in [Suleiman *et al.* 2013b,Suleiman *et al.* 2014a, Mary *et al.* 2014].

7.1 Introduction

Numerical results presented in this Section regard the detection of spectral lines in HSI data that are acquired by the integral field spectrograph MUSE (Multi Unit Spectroscopic Explorer). This instrument delivers data images of 300×300 pixels at approximately 3600 spectral channels in the visible spectrum (see Figure 7.1).

One of the main objectives of the MUSE instrument is the detection and characterization of very distant and faint galaxies known as Lyman-α emitters. Their spectra contain essentially one single emission line, whose profile varies with the considered object (see Figures 7.4(a) and 7.4(c)). These signatures can be simulated using astrophysical models leading to a large library of alternatives. We consider that this spectral library represents all possible alternatives under $\mathcal{H}_1$.

In our case, we obtained from astrophysical simulations performed by the MUSE consortium a library of 9745 specific profiles ($\boldsymbol{S} \in \mathbb{R}^{3600\times 9745}$). In the data, the emission lines can essentially be centered at *any* of the 3600 wavelength channels (because the astronomical redshift is related to the distance of the galaxy to Earth), thus the effective dimension of $\boldsymbol{S}$ is $L = 9745 \times 3600 \approx 3.5 \times 10^7$. Using the constrained GLR (4.9) on the exact model, the L

alternatives should be tested over each of the 90000 spectra of the data cube, an approach whose complexity is prohibitive [Paris *et al.* 2013b].

The problem of detecting a spectral profile in a data spectrum is typically the "one among many" detection problem stated by (4.1). We face the two associated difficulties, computational complexity and worst behavior of standard RD tests. Minimax RD tests are of particular interest here because the most "exotic" emission profiles (the typical ones) may also be the most interesting ones from an astrophysical viewpoint (precisely because they are atypical), so such profiles should not be left aside from the detection process.

Before evaluating the proposed approaches for detection tests, we introduce in Section 7.2 an overview of the MUSE instrument and we show the spectral line possibly presents in its data cubes. Most of the works in this Chapter focus on spectral model (i.e., pixel-vectors along the wavelength channels without taking into account the spatial leakage in neighboring pixel-vectors) but we also consider a case of "spatio-spectral" profiles in Section 7.4. A spatio-spectral profile is a three-dimensional signal (or, subcube) at spatial position (x, y) along the wavelength channels (third dimension, see Figure 7.7 for an illustration).

This Chapter ends with an approach allowing to evaluate the best number of atoms K that should be used in the minimax RD tests.

7.2 The MUSE spectrograph and the Lyman-α emitters

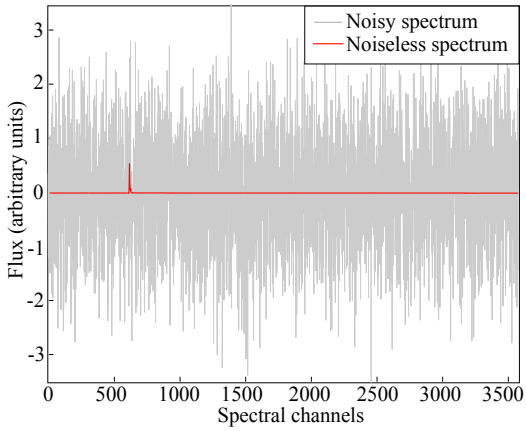
MUSE is a wide-field optical integral-field spectrograph [Caillier *et al.* 2012] built for the Very Large Telescope (VLT) in Chile, of the European Southern Observatory (ESO). An integral field spectrograph provides spectral information at every pixel of a two-dimensional scene (spatial sky positions x, y) forming three-dimensional image cubes (with the wavelength information as third dimension). The spectral range that this instrument covers is from 0.465μm to 0.93μm visible spectrum, with a spectral resolution of 0.13μm. MUSE is considered as a wide field spectrograph because it has 24 integral field unit (IFU), allowing to capture a large field of view. Figure 7.1(b) depicts the architecture of this instrument.

The main objective of MUSE is to study the formation of young galaxies (e.g., high redshift Lyman-α emitters), of nearby galaxies (e.g., supermassive back holes, interacting galaxies), of starts and resolved stellar populations (e.g., early stages of stellar evolution) and the study of solar system. In the framework of this Ph.D thesis (which started in December 2011), we use highly realistic simulated data set $\boldsymbol{S}$ provided by the MUSE consortium [Verhamme *et al.* 2012] (the MUSE instrument was successfully mounted on VLT on 19 January 2014, and saw first light on 31 January 2014).

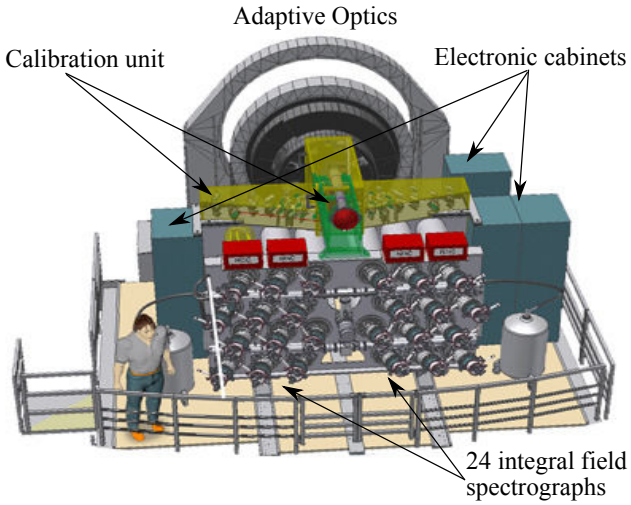
Figure 7.2 shows the data acquisition process of MUSE. In brief, once the light reaches the telescope, it is split two times and dispersed by the spectrograph, before being stored into a data cube. The first splitting divides the light into 24 subfields, each directed to the integral field unit (IFU). The second splitting is in each IFU, where each ray is split into 48 slices.

The study of formation of galaxies is a real challenge because it involves the observation of very far galaxies, more than ten billion light years from earth. The emission lines of young galaxies are associated to Hydrogen emission lines. They can be characterized by

redshifted Lyman-α emitters which are visible only through a faint and narrow emission line. Furthermore, the acquired data presents a very low SNR, affected by atmospheric perturbation and by Poisson and detectors' noise. Figure 7.1(a) compares a noisy Lyman-α spectral line against a noiseless Lyman-α spectral line. We can see that the target signature is buried in noise. These characteristics make the detection of Lyman-α sources in the data cube a difficult task (e.g., see [Paris *et al.* 2013a]).



(a) One noiseless and corresponding noisy spectra of a Lyman-α emitter in the MUSE data cube.



(b) View of the MUSE instrument.

Figure 7.1: (a) Example of a noiseless and corresponding noisy spectrum in MUSE data cube. (b) The structure of MUSE (Image credit to ESO). We can see different parts such as the calibration unit on top (in yellow and green), the electronic cabinets on each side, and all 24 of the integral field spectrographs (in gray).

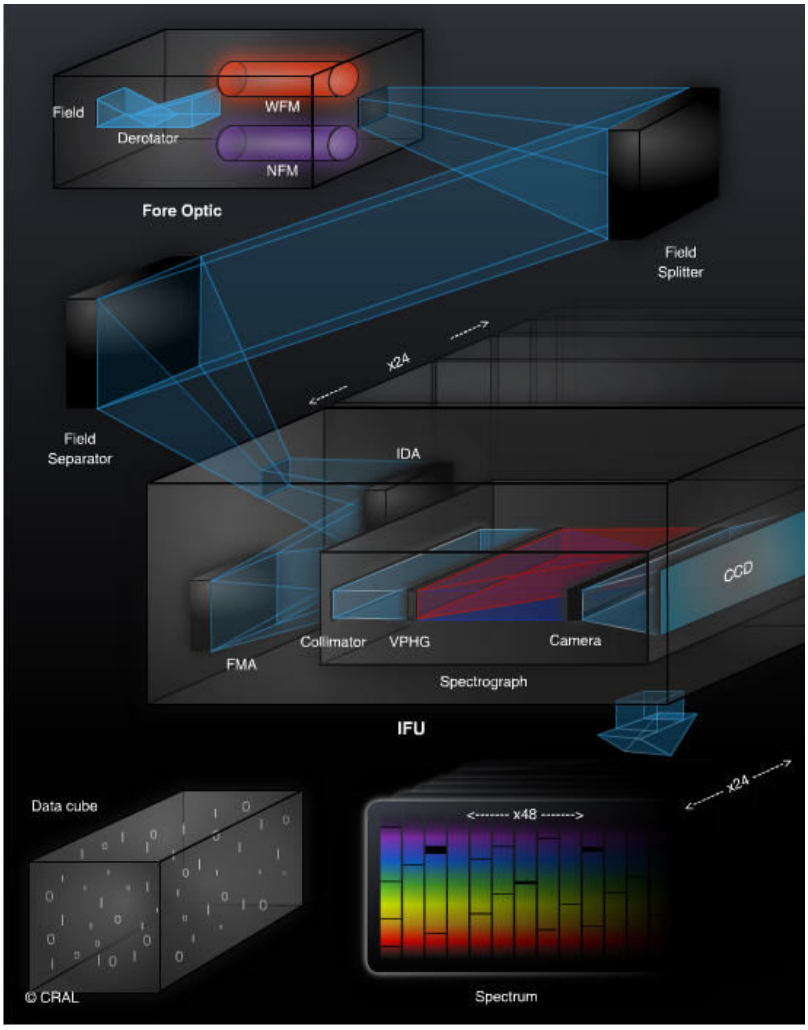


Figure 7.2: The acquisition process of the MUSE instrument. Once the light arrives, the optical rotator compensates the rotation of the field of view (of the telescope). Then, it passes through a set of optics. The light is then split into 24 subfields, each directed to one IFU. In each IFU, the light is split again into 48 slices. Then, a spectrograph disperses the light w.r.t. wavelength, which finally arrives at a detector that stores the signal, yielding a data cube. Image credit to ESO.

7.3 Worst-case detection of spectral profiles

For the purpose of computing sufficiently averaged ROC curves over the whole library for all tests, we make two simplifications in the simulations below. First, we consider alternatives of smaller dimension. To do so, we restrict the spectrum of each line to an interval of about $N = 100$ contiguous wavelength channels centered around the line maximum, see Figure 4.1(a). There is almost no energy at the rest of the wavelength channels anyway. Second, we do not consider possible translations under $\mathcal{H}_1$. These simplifications have no impact on the validity of the presented results but allow an exhaustive statistical analysis. Testing using a smaller sample of alternatives allow much more realizations of Monte Carlo simulations in shorter duration. Full size detection tests can be performed following the approach of [Paris *et al.* 2013b]. Here, we train the considered test dictionaries for a set of $L = 100$ (Section 7.3.1) and set of $L = 9745$ spectral profiles (Section 7.3.2).

In this Section, we compare the performances of several RD tests:

i. 1-dimensional reference test: Oracle NPD using the active alternative $\boldsymbol{s}_i$ under $\mathcal{H}_1$.

ii. 1-dimensional test using minimax atom $\boldsymbol{d}^*$.

iii. 1-dimensional test using the best rank-one approximation (SVD) of $\boldsymbol{S}$, $\boldsymbol{u}$.

iv. K-dimensional test using K-SVD dictionary $\boldsymbol{D}_K^{\text{K-SVD}}$.

v. K-dimensional test using greedy minimax dictionary $\boldsymbol{D}_K^*$.

vi. K-dimensional test using K-minimax dictionary $\boldsymbol{D}_K^{\text{K}*}$.

vii. K-dimensional test using SKM-minimax dictionary $\boldsymbol{D}_K^{\text{SKM}*}$, if applicable.

To recall, the one-dimensional tests above (i, ii, iii) are of the form[1]

$$|\boldsymbol{x}^\top \boldsymbol{a}| \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma, \tag{7.1}$$

where $\boldsymbol{a}$ is replaced by the corresponding atom (i.e., $\boldsymbol{s}_i$ for Oracle NPD i, $\boldsymbol{d}^*$ for ii, and $\boldsymbol{u}$ for iii). In contrast, the K-dimensional tests above (iv, v, vi, vii) have the form

$$T_{\boldsymbol{D}}(\boldsymbol{x}) = \max_{j=1,\ldots,K} |\boldsymbol{d}_j^\top \boldsymbol{x}| \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma, \tag{7.2}$$

where $\boldsymbol{D} = [\boldsymbol{d}_1, \ldots, \boldsymbol{d}_K]$ corresponds to the learned dictionary of different approaches (i.e., $\boldsymbol{D} = \boldsymbol{D}_K^{\text{K-SVD}}$ for iv, $\boldsymbol{D} = \boldsymbol{D}_K^*$ for v, $\boldsymbol{D} = \boldsymbol{D}_K^{\text{K}*}$ for vi and $\boldsymbol{D} = \boldsymbol{D}_K^{\text{SKM}*}$ for vii).

Note that we also include in this study the *Max* test over all alternatives (see (4.9): $\max_{i=1,\ldots,L} |\boldsymbol{s}_i^\top \boldsymbol{x}| \gtrless_{\mathcal{H}_0}^{\mathcal{H}_1} \gamma$), but only for comparison purposes, since it may not be implementable in practice on a full size data set.

For all the tests above, under $\mathcal{H}_0$: $\boldsymbol{x} = \boldsymbol{n}$ and under $\mathcal{H}_1$: $\boldsymbol{x} = \boldsymbol{s}_i \alpha_i + \boldsymbol{n}$, where $\boldsymbol{n} \sim \mathcal{N}(\boldsymbol{0}, \mathbf{I})$, i and α_i are unknown. From the library $\boldsymbol{S}$, the alternatives $\boldsymbol{s}_i$ are activated one

[1] The absolute value correspond to a two sided test where the amplitude can be positive (emission) or can be negative (absorption). Considering only emission lines will remove the absolute values in (7.1)-(7.2).

by one from $i = 1$ until $i = L$, allowing to perform an exhaustive worst-case performances evaluation. The corresponding amplitude α_i is fixed for all i (so that the performances are comparable w.r.t. SNR). Thus, α_i is the parameter controlling the SNR of the simulation. The threshold γ is obtained by Monte Carlo simulations for the case where $\boldsymbol{D}$ has more than one atom ($K > 1$). In the case $K = 1$, the threshold can be computed analytically.

7.3.1 Results using a library of 100 spectral lines

For the considered reference library of $L = 100$ spectral lines (denoted by $\boldsymbol{S}_{100}$), it turns out that 5 alternatives fall on the smallest enclosing circle defined by (5.13), whose center is the minimax atom $\boldsymbol{d}^*$ (learned from the 100 spectral lines). We called these the most marginal alternatives. To train dictionaries of $K > 1$ atoms, we would naturally choose the number of atoms K superior to the number of the most marginal alternatives. By this, we expect that the learned dictionaries will sufficiently capture the intrinsic diversity of the library $\boldsymbol{S}_L$. Recall that in Chapter 5, the numerical simulations for $N = 3$ showed that in the case of $K = 1$ (Example 5.2.1), the optimized minimax atom was hold by three marginal alternatives. Thus, learning a dictionary of $K = 1$ might be insufficient to represent well the various profiles of $\boldsymbol{S}_L$. Then, in Example 6.4.1 where $K = 3$, we have seen that the minimax correlation function (5.22) increases w.r.t. the case where $K = 1$. With this in mind, we investigate the detection performances of tests using $K = 1$ ($\boldsymbol{d}^*$ and $\boldsymbol{u}$) and $K = 6$ atoms in the learned dictionaries ($\boldsymbol{D}_6^{\text{K-SVD}}$, $\boldsymbol{D}_6^*$, $\boldsymbol{D}_6^{\text{K}*}$ and $\boldsymbol{D}_6^{\text{SKM}*}$).

The results presented here (Figure 7.3 and Table 7.1) are obtained from 2×10^5 realizations (uncertainty: ± 0.001). The uncertainty is calculated by the difference between the minimum and the average AUC of the Oracle NPD. This value is close to 3 times the standard deviation of the (roughly) Gaussian estimation noise on the ROC.

Table 7.1 summarizes the AUCs of the ROC curves for different tests. The *best* worst-case (or minimax) performances are achieved by tests using greedy minimax (blue circles line) and K-minimax (orange crosses line) dictionaries, respectively $\boldsymbol{D}_6^*$ of Section 6.2 and $\boldsymbol{D}_6^{\text{K}*}$ of Section 6.3. This is followed by the test using the SKM-minimax (gray line) dictionary $\boldsymbol{D}_6^{\text{SKM}*}$ (presented in Section 6.4). The one-dimensional test using $\boldsymbol{d}^*$ (studied in Section 5.2.2) is the third best minimax performance (cyan circle line).

More classical RD methods based on minimizing the MSE in the dictionary update stage, (K)-SVD algorithm (cf. Example A.5.3) suffer from a large detection power loss for some alternatives, particularly in one-dimensional testing (using $\boldsymbol{u}$). These power losses can be seen in Figure 7.3 (pink, dash-dots) for alternatives such as $\boldsymbol{s}_1$, $\boldsymbol{s}_4$, etc., with a particularly large detection power loss at $\boldsymbol{s}_{90}$.

The third column of Table 7.1 shows the average detection performances of these tests. Testing using SVD atom ($\boldsymbol{u}$) yields the best average performance, as it is close to the Oracle NPD (black dots). The average performances of minimax RD tests ($K > 1$) are equivalent to that of *Max* test (using $\boldsymbol{S}_{100}$: red dashes, close to the blue and gray lines in Figure 7.3).

In summary, the proposed minimax RD tests perform indeed better in worst-case scenarios than the (K)-SVD RD tests. Increasing the number K from one atom to a few atoms improves the worst-case performance and the upshot of better sampling increases the average power.

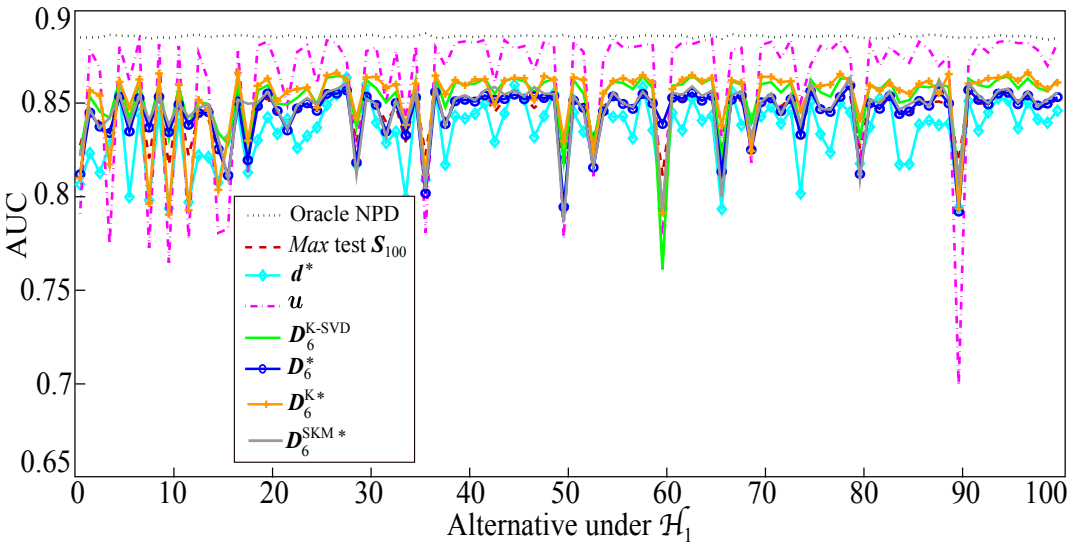


Figure 7.3: AUCs for all i, $\{i = 1, \ldots, 100\}$ instances under $\mathcal{H}_1$. The Figure compares the detection performances of RD models using 6 learned dictionaries. Two of them are one-dimensional atoms, and the rest of them are $K = 6$ atoms. The exact model $\boldsymbol{S}_{100}$ (red dashes, close to the blue and gray lines) and the reference AUC are also provided (Oracle NPD: black dots). We can see that RD tests using the classical approaches (SVD: pink dash-dots, and K-SVD: green solid line) suffer from large losses for certain alternatives, e.g., $\boldsymbol{s}_{60}$ and $\boldsymbol{s}_{90}$. On the contrary, minimax approaches maintain as much power as possible in these worst-case scenarios.

Dictionary	**Min AUC (worst-case)**	**Average AUC**
Oracle NPD	0.886	0.887
$\boldsymbol{S}_{100}$	0.813	0.847
$\boldsymbol{d}^*$	0.768	0.836
$\boldsymbol{u}$	0.700	0.863
$\boldsymbol{D}_6^{\text{K-SVD}}$	0.753	0.843
$\boldsymbol{D}_6^*$	0.794	0.846
$\boldsymbol{D}_6^{\text{K}*}$	0.792	0.854
$\boldsymbol{D}_6^{\text{SKM}*}$	0.788	0.848

Table 7.1: Results over 100 alternatives. (Uncertainty due to the estimation noise of the ROCs: ± 0.001). This table shows that RD detection test using SVD ($\boldsymbol{u}$) suffers from a large loss w.r.t. Oracle NPD, while the loss for the minimax atom ($\boldsymbol{d}^*$) is (maximally) minimized. By adding more atoms to the learned dictionary ($K = 6$), the worst-case performance is improved (compare for instance the test using greedy minimax $\boldsymbol{D}_6^*$ to the test using one-dimensional minimax $\boldsymbol{d}^*$). The average performances of RD tests for $K = 6$ are comparable to that of *Max* test using $\boldsymbol{S}_{100}$.

Numerical simulations in the next Section regard a library $\boldsymbol{S}_L$ with much more diversity (where L is approximately 10000).

7.3.2 Results using a library of approximately 10000 spectral lines

For the full size reference library, it turns out that 16 alternatives out of 9745 lie on the smallest enclosing circle defined by (5.13) for the minimax atom $\boldsymbol{d}^*$ (those are shown in Figure 7.4(c)). Thus, we set in the following K to a sufficiently large value ($K = 70$) for the proposed minimax approaches. As mentioned in the previous Section 7.3.1, a too low value of K would not allow to capture the diversity of the marginal subpopulations of $\mathcal{S}$, resulting in no or little improvement of the worst-case performance w.r.t. the case $K = 1$.

Before turning to the tests comparison, let us first consider Figure 7.4(d), which shows that the function $\rho^{(K)}$ (5.22) used as a proxy for greedy minimax learning (Section 6.2) is increasing rapidly in K. This fulfills the objective set in Section 5.2.4: increase the minimax correlation as a proxy for the worst detection power.

Table 7.2 and Figure 7.5 summarize the results in terms of AUC of the ROC. For these simulations, the uncertainty caused by estimation noise on the ROC is ± 0.003.

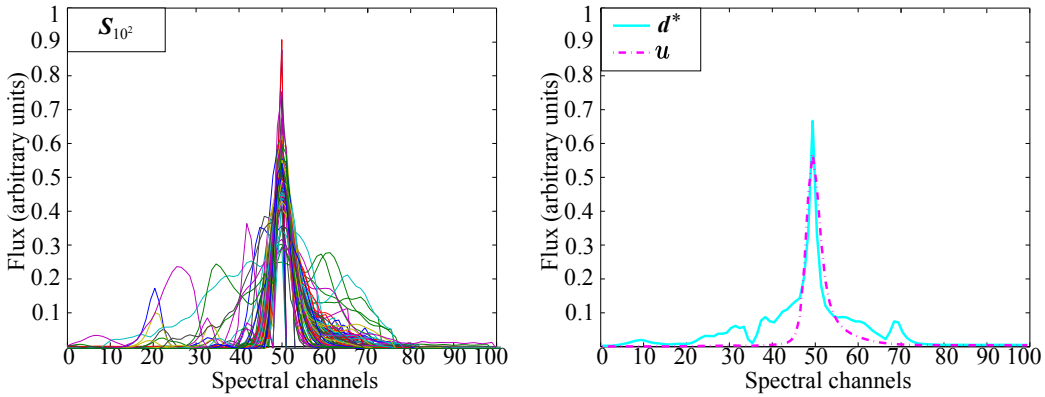
If we first consider AUC results *averaged over all alternatives* (i.e., the most usual criterion, but indeed not the one under focus), the second column of Table 7.2 shows that the best performances are obtained by the standard SVD: the first is $\boldsymbol{u}$ (pink, dash-dots), which is nearly as good as the reference (i.e., the Oracle NPD for each alternative, black dots), and the second is $\boldsymbol{D}_{70}^{\text{K-SVD}}$ (green solid). As argued before, $\boldsymbol{u}$ and K-SVD tend to capture sketchy features shared by most alternatives (compare, in Figure 7.4(b), $\boldsymbol{u}$ to the learned minimax spectral profile $\boldsymbol{d}^*$ for $K = 1$). In terms of detection power this translates into good results in average. As visible in Figure 7.5 however, this comes at the price of large power losses for some alternatives (e.g., $\boldsymbol{s}_{90}$, $\boldsymbol{s}_{108}$, $\boldsymbol{s}_{111}$ and others).

In contrast, RD testing with the proposed dictionaries $\boldsymbol{d}^*, \boldsymbol{D}_{70}^*$ and $\boldsymbol{D}_{70}^{\text{K}*}$ perform better than with $\boldsymbol{u}$ and K-SVD dictionary in terms of worst-case scenario performances. The overall performances is more stable (e.g., limited power loss for $\boldsymbol{s}_{90}$, $\boldsymbol{s}_{108}$, see Figure 7.5).

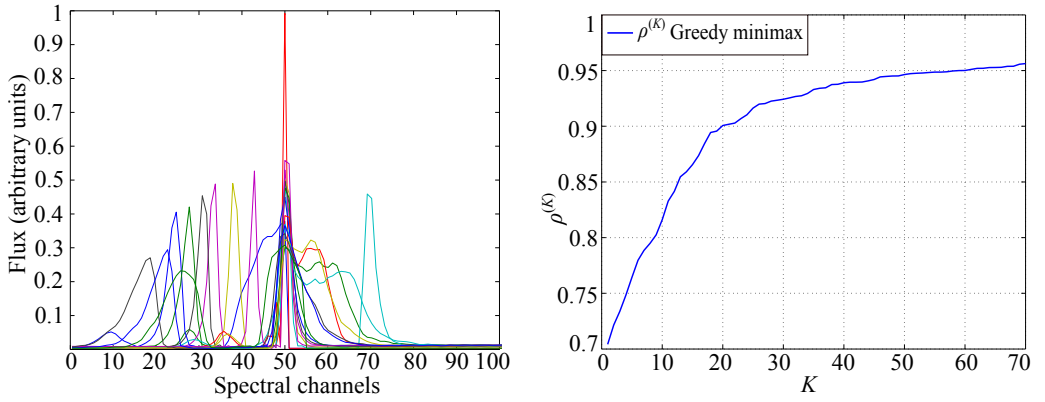
Comparing now the proposed optimization approaches for $K > 1$ (in Figure 7.5: greedy minimax $\boldsymbol{D}_{70}^*$, blue circles; K-minimax $\boldsymbol{D}_{70}^{\text{K}*}$, orange crosses) to the minimax dictionary for $K = 1$ ($\boldsymbol{d}^*$, cyan diamonds), we see that the worst-case performances are improved w.r.t. $K = 1$, which was the main objective of the study of Section 5.2.4. Note that as a side effect of better sampling the diversity of the alternatives, we also gain in average performances.

Table 7.2 shows that the worst-case performance of the *Max* test using $\boldsymbol{S}_{9745}$ is comparable to those of the RD tests with $\boldsymbol{d}^*, \boldsymbol{D}_{70}^*$ and $\boldsymbol{D}_{70}^{\text{K}*}$. The proposed approaches thus allow here to obtain essentially the same performances as the *Max* test (red dashes, close to the blue circles and orange crosses in Figure 7.5) with much lower complexity by limiting the power losses inherent to classical dimensionality reduction methods.

As a final remark, comparing the results obtained in Section 7.3.1 for $L = 100$ with the results obtained here for $L \approx 10000$, the relative behaviors of the RD tests are the same. However, for a very large library of alternatives, the number of K atoms has to be increased sufficiently, hence the interest of Section 5.2.4.



(a) 100 spectral lines in $\boldsymbol{S}_{9745}$ (b) Comparison of 1D learned atoms



(c) 16 marginal alternatives of $\boldsymbol{S}_{9745}$ (d) Minimax function of (5.22)

Figure 7.4: (a) 100 of the alternatives in library $\boldsymbol{S}_{9745}$. (b) $\boldsymbol{d}^*$ (minimax) and $\boldsymbol{u}$ (SVD) atoms, learned over 9745 alternatives. (c) The 16 alternatives of $\boldsymbol{S}_{9745}$ lying on the smallest enclosing circle $\mathcal{C}$ w.r.t. $\boldsymbol{d}^*$. (d) Minimax correlations $\rho^{(K)}$ for the greedy minimax, where $K = 1, \ldots, 70$.

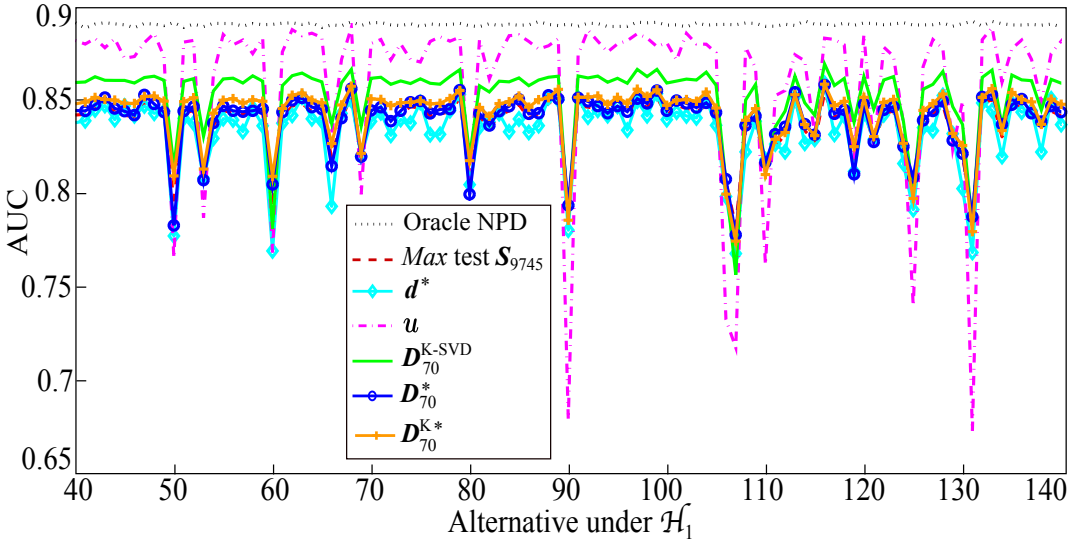


Figure 7.5: AUC shown for 100 alternatives under $\mathcal{H}_1$, $\{i = 41, \ldots, 140\}$. The simulations were done for $\{i = 1, \ldots, 9745\}$, activated one by one under $\mathcal{H}_1$ (given $\boldsymbol{S}_{9745}$). Results over the whole alternatives are summarized in Table 7.2. We compare here the detection performances of RD models using 5 learned dictionaries, and the exact model $\boldsymbol{S}_{9745}$ (red dashes, close to the blue circles and orange crosses). Two of the learned dictionaries are 1D ($\boldsymbol{d}^*$: cyan diamond line and $\boldsymbol{u}$: pink dash-dots), and the rest of them consist of $K = 70$ atoms. We also include the reference AUC (Oracle NPD: black dots). Minimax approaches are more robust w.r.t. some alternatives inducing maximum power losses (e.g., $\boldsymbol{s}_{90}, \boldsymbol{s}_{108}, \boldsymbol{s}_{111}, \boldsymbol{s}_{125}$ and $\boldsymbol{s}_{131}$).

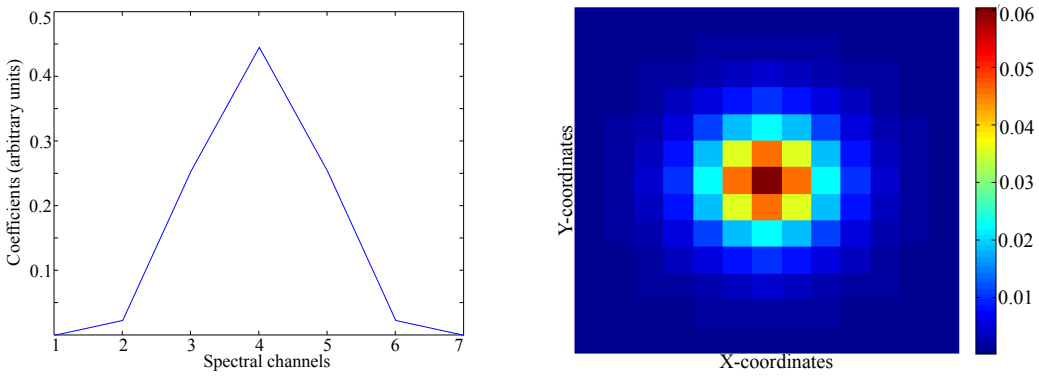
Dictionary	**Min AUC (worst-case)**	**Average AUC**
Oracle NPD	0.884	0.887
$\boldsymbol{S}_{9745}$	0.772	0.841
$\boldsymbol{d}^*$	0.763	0.838
$\boldsymbol{u}$	0.670	0.870
$\boldsymbol{D}_{70}^{\text{K-SVD}}$	0.737	0.856
$\boldsymbol{D}_{70}^{*}$	0.773	0.844
$\boldsymbol{D}_{70}^{\text{K}*}$	0.769	0.846

Table 7.2: Results over 9745 alternatives. Uncertainty due to the estimation noise of the ROCs: ± 0.003. SKM-minimax is not included here, because the clustering for $K = 70$ yields some empty clusters. Similar to Table 7.1, the maximum loss in a worst-case scenario for these simulations, occurs for RD test using SVD ($\boldsymbol{u}$) (for $K = 1$) and K-SVD ($\boldsymbol{D}_{70}^{\text{K-SVD}}$) (for $K > 1$). While for greedy minimax ($\boldsymbol{D}_{70}^{*}$), the worst-case performance is equivalent to those of the exact model ($\boldsymbol{S}_{9745}$).

7.4 Worst-case detection of spatio-spectral (3D) profiles

7.4.1 3D PSF and 3D atoms

Extending the spectral composite hypothesis models (4.1) and (5.1) presented in Chapter 4 and Chapter 5, we examine here a spatio-spectral model, taking into account the spatial Point Spread Function (PSF) of the MUSE instrument[2]. This PSF describes the instrument's spatial and spectral response to a point source. The instrument induces a spatio-spectral leakage of a point source found in the spectral and spatial domains. Figure 7.6 illustrate the PSF of MUSE in spectral form and in a spatial scene respectively.



(a) MUSE's PSF: spectral view (b) MUSE's PSF: spatial view

Figure 7.6: The PSF of MUSE's instrument. Figure (a) shows a spectral view and Figure (b) shows a spatial view.

The motivation for accounting for the spatio-spectral PSF is to gain in power because the test will then account for the corresponding spatio-spectral information leakage under $\mathcal{H}_1$. With regard to the spectral model of (4.1), we obtain a corresponding "spatio-spectral" model by convolving each $\boldsymbol{s}_i \in \mathbb{R}^N$ with the spatio-spectral PSF (represented by a cube of $13 \times 13 \times 7$) yielding 3D sub-cubes representing the spatio-spectral signatures of the $\{\boldsymbol{s}_i\}$.

Similarly, for the first reduced composite hypothesis model (5.1), the atoms of the dictionary $\boldsymbol{D} \in \mathbb{R}^{N \times K}$ are convolved with the PSF yielding a dictionary of 3D atoms. The numerical simulations presented in Section 7.4.2 studies the case where $\boldsymbol{D} = \boldsymbol{d}$ (single atom, $K = 1$). Figures 7.7 illustrate respectively the resulting 3D minimax atom $\boldsymbol{d}^*$, and the resulting 3D SVD atom $\boldsymbol{u}$ in three dimensions.

[2]This is a simplified model for the purpose of illustration. In reality, this PSF is both variable spectrally and spatially.

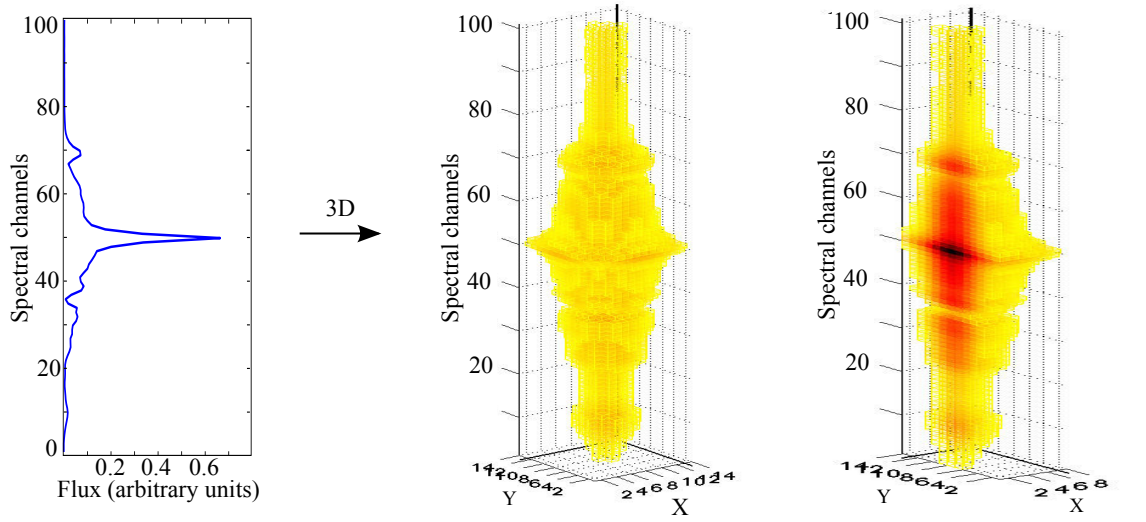
7.4.2 Detection results on simulations

The numerical simulations presented below deal with a simulated 3D data cube of dimension $50 \times 50 \times 100$. This data cube contains 9 spatio-spectral profiles, as shown in Figure 7.8. For simplicity we do not simulate the fact that the spectral lines of the 9 objects may actually be centered at any of the 3600 wavelengths. All are centered at spectral channel 50 and the data cube is thus shorter in wavelength than actual MUSE data cubes.

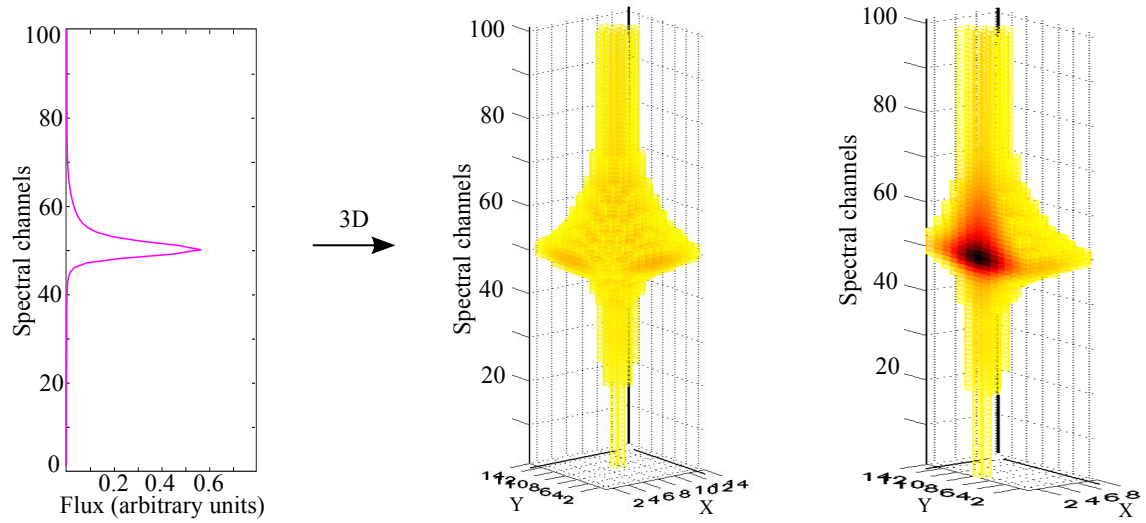
Figure 7.8(a) depicts the mean (in wavelength, $N = 100$) of the noiseless data cube. Figures 7.8(b) - 7.8(f) illustrate the corresponding 9 Lyman-α profiles under $\mathcal{H}_1$. Five of them have similar shapes ($\boldsymbol{s}_1$ to $\boldsymbol{s}_5$) and four of them were chosen among marginal alternatives ($\boldsymbol{s}_6$ to $\boldsymbol{s}_9$). In the presence of high noise (SNR $= -17$dB), all of the 9 profiles are totally buried in noise as shown in Figure 7.9(b).

The numerical results (at fixed $\mathrm{P_{FA}} = 0.01$ and SNR $= -17$dB) show that testing using minimax 3D atom (Figure 7.9(c)) yields better detection power for the marginal alternatives (in circles, particularly profiles number 7 to 9, while the powers are almost equals for profile number 6) than using SVD 3D atom (Figure 7.9(d)). The comparative evaluation of the performances is more clear in Figure 7.9(e) where we plot the difference of detection powers (i.e., $\Delta\mathrm{P_{Det}} = \mathrm{P_{Det}}(\boldsymbol{d}^*) - \mathrm{P_{Det}}(\boldsymbol{u})$). We can see (in Figure 7.9(e)) that $\Delta\mathrm{P_{Det}}$ are positives for profiles number 7 to 9 (it is near to zero for profile number 6), meaning that the detection powers of testing using minimax 3D atom are larger than those using SVD 3D atom for these profiles. In contrast, $\Delta\mathrm{P_{Det}}$ are negatives for profiles 1 to 5, meaning that the detection powers of testing using SVD 3D atom are larger than those using minimax 3D atom for these profiles.

The worst detection rate for the test using SVD atom is 0.3649 (profile number 9), while it is 0.4770 for the test using minimax atom (profile number 8), which illustrate the better behavior of the proposed test in worst-case scenario.

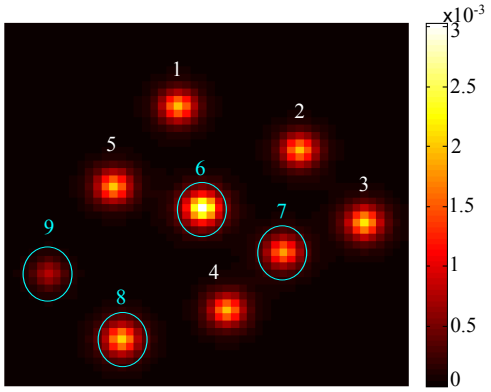


(a) Minimax atom, $\boldsymbol{d}^*$

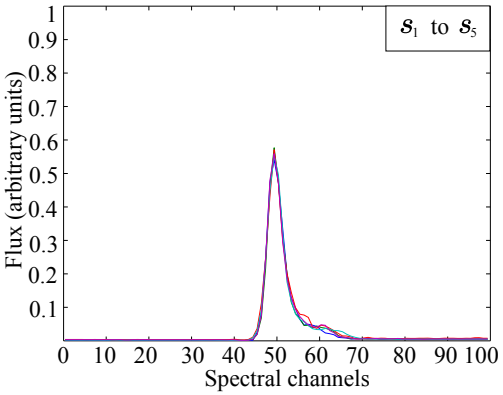


(b) SVD atom, $\boldsymbol{u}$

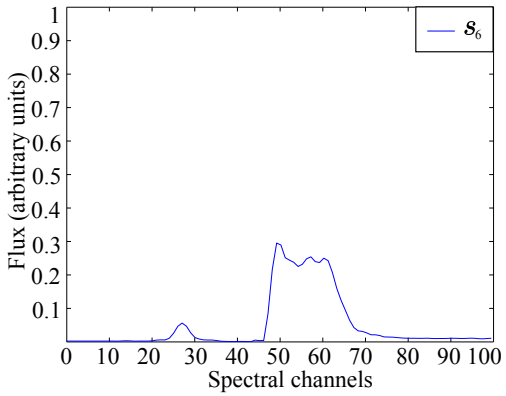
Figure 7.7: Example of 3D (spatio-spectral) learned atoms. (a) Minimax atom, and (b) SVD atom. For both subfigures, the left panels show the spectral profiles, the middle panels show the corresponding 3D learned atoms and the right panels show a cut of the 3D atoms.



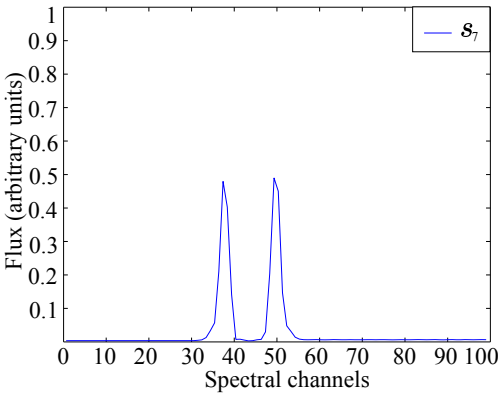
(a) Mean over the wavelengths of the noiseless data cube



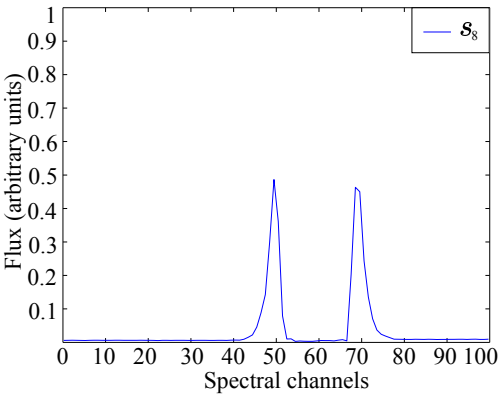
(b) $\boldsymbol{s}_1$ to $\boldsymbol{s}_5$ (all have similar shapes)



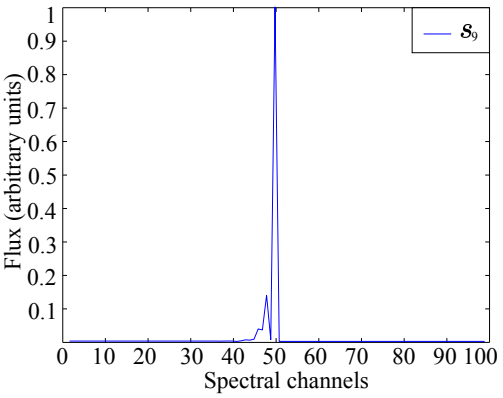
(c) $\boldsymbol{s}_6$



(d) $\boldsymbol{s}_7$

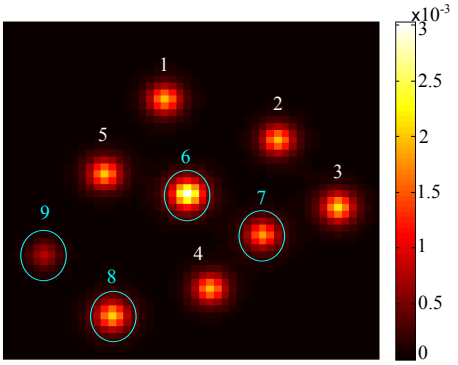


(e) $\boldsymbol{s}_8$

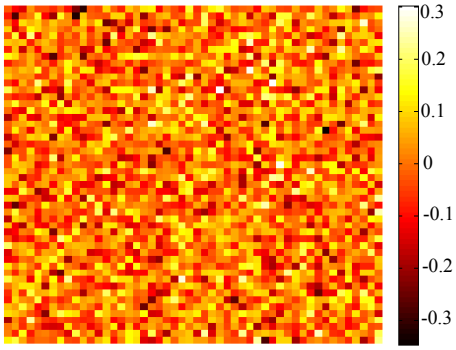


(f) $\boldsymbol{s}_9$

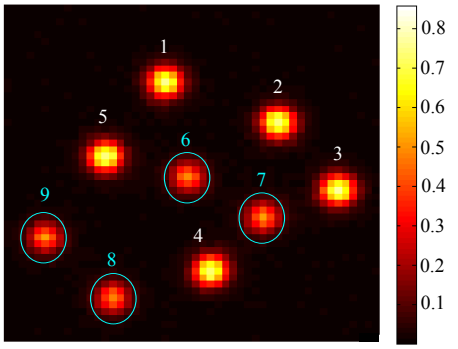
Figure 7.8: (a) Noiseless data cube (with spectral profiles convolved by the 3D PSF), averaged over spectral channels. This cube contains 9 Lyman-α profiles. Profiles 1 to 5 have a similar shape, while the rest are marginal profiles (number 5 to number 9).



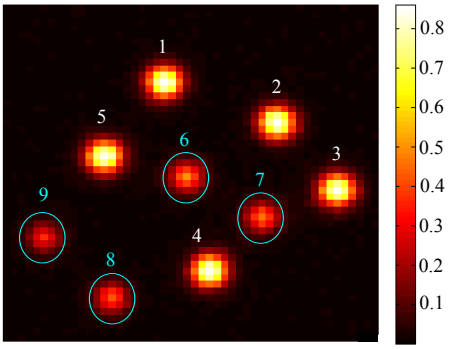
(a) Mean over the wavelengths of the noiseless data cube



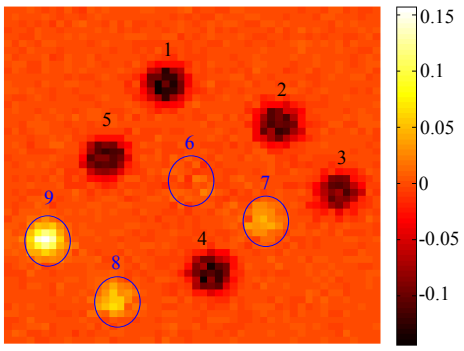
(b) Mean over the wavelengths of the noisy data cube (SNR=−17dB)



(c) Detection map for tests using 3D version of $\boldsymbol{d}^*$



(d) Detection map for tests using 3D version of $\boldsymbol{u}$



(e) $\mathrm{P}_{\mathrm{Det}}(\boldsymbol{d}^*) - \mathrm{P}_{\mathrm{Det}}(\boldsymbol{u})$

Figure 7.9: (a) and (b) show respectively the mean (over the wavelengths) of the noiseless and noisy data cube. (c) and (d): Detection maps for spatio-spectral hypothesis testing at fixed $\mathrm{P}_{\mathrm{FA}} = 0.01$, for SNR $= -17$dB. The results show that tests using minimax 3D atom (subfigure (b)) yield better detection power for the marginal alternatives (in circles, particularly for profiles 7 to 9) than using SVD 3D atom (subfigure (b)). Subfigure (e) depicts the difference of performances of test using 3D minimax vs. using 3D SVD (see text).

7.5 Strategies for determining best number of K atoms w.r.t. minimax criterion

Generally, when performing classification or clustering tasks, we have to set the number of desired classes, K. Finding an optimal procedure to choose K is still an open problem. A common practice is to execute a large number of numerical simulations and pick the best value of K w.r.t. a defined criterion.

For the considered worst-case criterion, a similar approach to evaluate the best number of K atoms is to train several dictionaries (train greedy minimax dictionary of $K = 1$ atom and several dictionaries of $K = 2, \ldots, 100$ atoms). Then we evaluate the AUCs over L alternatives for each dictionary, noting the worst-case AUC value for each "K-trained" dictionary.

We plot below the values of these worst-case performances w.r.t. the number of atoms K for SNR = 8dB: we obtain the curve shown in Figure 7.10. From this example, we can see that increasing the number of learned atoms K does not necessarily improve the worst-case performance (for low values of K the minimum AUC drops). However, the minimax performance globally increases with K until we reach a roughly constant level of minimum AUC. We can then pick the value of K corresponding to the left side of the plateau (about $K = 36$ here).

Of course, this method is time consuming especially if the number of alternatives L is very large. For instance, if $L \approx 10000$, it is not possible to learn dictionaries until say, $K = L/2 = 5000$ (with $L \approx 10000$ this means, in order to obtain the worst-case scenario in each case, performing $K(K+1)L = 2.5 \times 10^{11}$ Monte Carlo simulations of at least 1000 realizations each, which is completely out of reach). So much so, we could not go beyond $K \approx 100$.

Another simple approach is to try a restricted number of values for K, with values sufficiently larger than the number of marginal alternatives (w.r.t. $\boldsymbol{d}^*$), and evaluate the minimax detection performances only for those values. For example, there are 16 marginal alternatives of $\boldsymbol{S}_{9745}$ on the circle that defines $\boldsymbol{d}^*$. We can try testing dictionaries of for instance $K = 17, 20, 35, 70, 100$ atoms (red circles in Figure 7.10). For $\boldsymbol{S}_{9745}$, the computation time[3] to train a greedy minimax dictionary of $K = 10$ atoms ($\boldsymbol{D}_{10}^*$) is about 1.94 minutes. Performing Oracle NPD and greedy minimax RD tests (using $\boldsymbol{D}_{10}^*$) over 9745 alternatives (for 10^5 Monte Carlo realizations) takes about 12 hours.

[3]Using a standard machine of 2.7GHz processor and 4Go of DDR3 RAM.

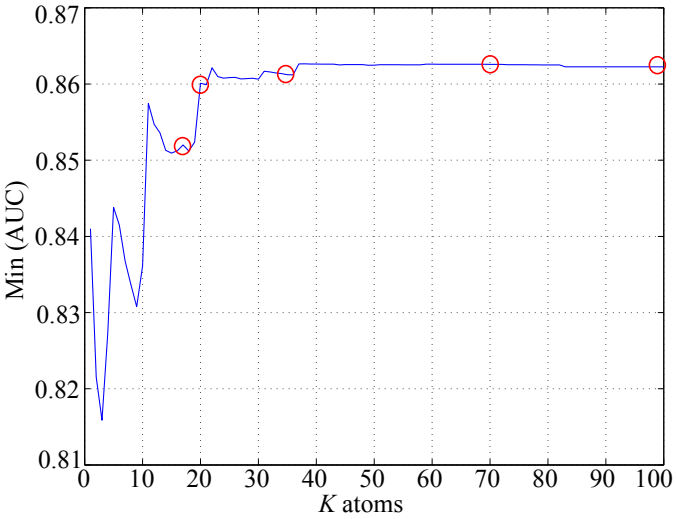


Figure 7.10: The blue line represents worst-case performances (minimum AUC of the ROC curves) of $\boldsymbol{D}^*_{(K)}$, for $K = 1, \ldots, 100$ over 100 alternatives ($\boldsymbol{S}_{100}$), for SNR = 8dB ($\alpha = 2.5$). In this example, there are no improvements of the detection power for $K \geq 36$. One can then choose $K = 36$. The red circles mark some picked values of K and the corresponding worst-case performances.

7.6 Discussion

This Chapter illustrates on an astrophysical application our studies of Part II, where we had proposed minimax RD tests based on GLR. The interest of the minimax strategy resides in this application in the implementation of tests that are robust for marginal spectral profiles, allowing astrophysicists to better detect such unusual sources. Choosing the best number of columns for the dictionary (w.r.t. minimax objectives) remains an issue, but we have suggested two approaches that can be used in practice.

In the next Chapter, two other possible applications are presented for worst-case pattern recognition. By using the proposed minimax learning algorithms, we will first learn dictionaries of faces from a known database. Then we will turn to the worst-case recognition rates when identifying handwritten digits.

Chapter 8

Machine learning applications

Contents

Some results presented in this Chapter were published in [Suleiman *et al.* 2014a,Suleiman *et al.* 2014b].

8.1 Introduction

In this Chapter, we turn to a more "visual" application, which is about image (or pattern) recognition. Pattern recognition methods generally seek to assign a new input to an appropriate cluster based on a similarity criterion. The performance of a recognition algorithm is usually evaluated in term of average recognition rate. The approach of this thesis instead focusing on the *worst-case recognition rate*.

First, we compare the proposed minimax dictionaries with (K)-SVD dictionaries learned from a known database of 40 faces. Next, we apply the proposed algorithm in a one-dimensional setting to train minimax dictionaries of handwritten digits. We also use for comparison a trained dictionary corresponding to the best rank-one approximation (SVD) of handwritten digits. We test then the recognition rate on another set of data using each of the learned dictionaries described above in the worst-case setting.

8.2 Minimax learning of faces

We illustrate in this Section the results of the algorithms in terms of learned atoms in the case where the library $\boldsymbol{S}$ is a database of faces. In the next Section 8.3, we will evaluate the worst-case recognition rates. $L = 40$ subjects were selected from the ORL Database of Faces by AT&T Laboratories Cambridge [Samaria & Harter 1994], representing the set of possible alternatives under $\mathcal{H}_1$ (Fig. 8.1(a) shows 20 faces).

For $K = 1$, the one-dimensional learned atoms exhibit the behaviors discussed in previous Chapters. The minimax atom $\boldsymbol{d}^*$ (Fig. 8.1(b)) captures marginal features (glasses, different eyes, noses, and mouths positions), while the SVD atom $\boldsymbol{u}$ (Fig. 8.1(c)) represents an "average face" with shared and smoothed characteristics.

Setting $K > 1$ allows for more atoms to be learned, forming worst-case detection dictionaries whose axes dissociate to focus on specific sets of "atypical" signatures. For $K = 3$ for instance (see Fig. 8.1(d)), the second minimax atom is identified to a woman face alternative whose features are very dissimilar (in a correlation sense) from the others. The first atom of $\boldsymbol{D}_3^*$ is similar to $\boldsymbol{d}^*$, with some features removed that are transferred to the third atom. In contrast, the K-SVD method (Fig. 8.1(e)) yields in this case three "smooth faces".

(a) Some alternatives in library $\boldsymbol{S}$

(b) $\boldsymbol{d}^*$

(c) $\boldsymbol{u}$

(d) $\boldsymbol{D}_3^* = [\boldsymbol{d}_1^*, \boldsymbol{d}_2^*, \boldsymbol{d}_3^*]$

(e) $\boldsymbol{D}_3^{\text{K-SVD}} = [\boldsymbol{d}_1^{\text{K-SVD}}, \boldsymbol{d}_2^{\text{K-SVD}}, \boldsymbol{d}_3^{\text{K-SVD}}]$

Figure 8.1: (a) 20 faces in the database of 40 faces, front-facing. (b) One-dimensional minimax face, and (c) SVD face. (d) Greedy minimax faces, $K = 3$, and (e) K-SVD faces, $K = 3$. K-SVD represents average features while worst-case algorithms capture marginal features.

8.3 Worst-case recognition rate of handwritten digits

Another possibly interesting application of this minimax approach is for the recognition of handwritten digits. We use the MNIST database [LeCun *et al.* 1998] which contains 60000 examples of handwritten digits (from 0 to 9), collected from approximately 250 writers. For our demonstration, we use 4121 training examples and 1300 test examples for each digit.

In this case, classification methods aim at maximizing the probability that the digits are correctly classified (e.g., a handwritten 1 should be correctly classified in class 1). In our framework, we wish to learn dictionaries aimed at maximizing the probability of correct classification in the worst-case of the training database (for instance, a dictionary will be learnt for class 1 in order to be robust for all instances of handwritten 1 digits).

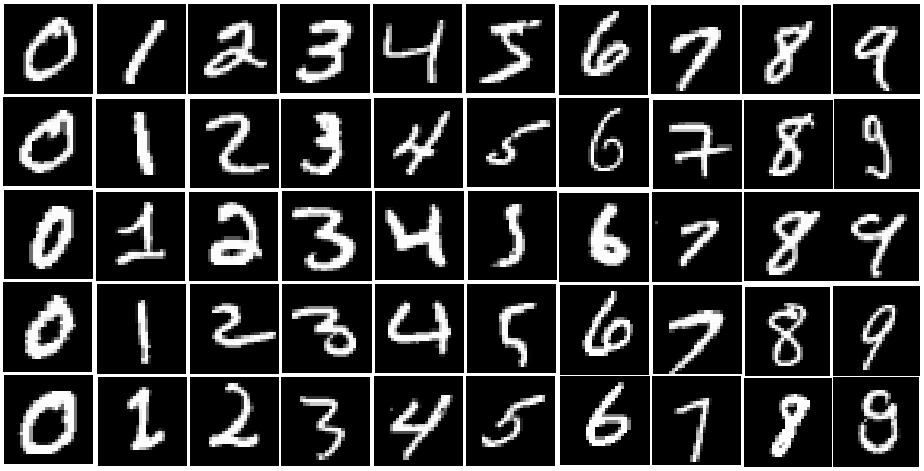
For each database (library) $\boldsymbol{S}^c, c = 0, \ldots, 9$ (each composed of 4121 alternatives), we learned the corresponding minimax atom $\boldsymbol{d}^{*\,c}$. We concatenated these ten atoms into a matrix $\boldsymbol{B}^* = [\boldsymbol{d}^{*\,0}, \ldots, \boldsymbol{d}^{*\,9}]$. The same was done for the SVD method resulting in a matrix $\boldsymbol{B}^{\mathrm{SVD}}$. Figure 8.2(b) and 8.2(c) depicts the atoms of the learned dictionaries for the two methods.

We evaluated the probability of correct classification in the worst-case scenario for each method and for each class. To do so, in the case of $\boldsymbol{B}^*$ for instance, we identified for each class c which of the 1300 test alternatives had the minimum correlation with $\boldsymbol{B}^* = [\boldsymbol{d}^{*\,0}, \ldots, \boldsymbol{d}^{*\,9}]$. Let $\boldsymbol{s}^c_{i^*}$ denote this alternative. Correct classification is achieved for this worst-case alternative if $\arg \max_{j=0,\ldots,9} \boldsymbol{d}^{*\,j\,\top} \boldsymbol{s}^c_{i^*} = c$. The same procedure was used to evaluate the correct classification of the SVD approach with $\boldsymbol{B}^{\mathrm{SVD}}$.

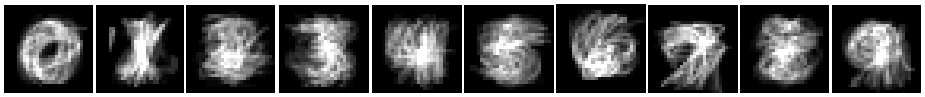
We executed this experiment 1000 times (allowing random permutation between training and test examples). Table 8.1 shows the probability of correct recognition in the worst-case scenario for this experiment. These results show that the "minimax approach" always classifies better the worst alternative than SVD.

Note that SVD is always mistaken (null recognition rate) by such alternatives in seven cases out of ten. The minimax approach is indeed more robust.

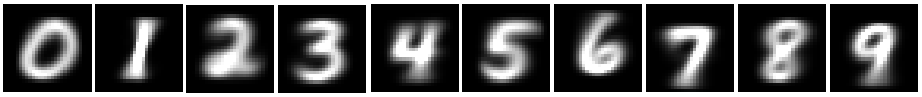
Of course, a much better classification results could be obtained in this context by injecting discriminative principles between classes in the learning process and by increasing K. The purpose of these simulations is only to illustrate another application where the proposed worst-case scenario learning methods could be used and elaborated.



(a) Some samples of the handwritten digits in the library $\boldsymbol{S}^c$



(b) $\boldsymbol{B}^* = [\boldsymbol{d}^{*\,0}, \boldsymbol{d}^{*\,1}, \ldots, \boldsymbol{d}^{*\,9}]$



(c) $\boldsymbol{B}^{\mathrm{SVD}} = [\boldsymbol{u}^0, \boldsymbol{u}^{,1}, \ldots, \boldsymbol{u}^9]$

Figure 8.2: (a) Some samples in the database of handwritten digits. Figures (b) and (c) learned atoms ($K = 1$) for each digit by different approaches: minimax and SVD, respectively.

Digit	**Worst-case recognition** (%)	
	Minimax	SVD
0	45.6	27.7
1	31.9	0
2	0.8	0
3	5.8	0
4	21.3	13.7
5	18.8	0
6	14.3	0
7	15.5	0
8	20.9	0.9
9	2.4	0

Table 8.1: Worst-case recognition rates for handwritten digits.

8.4 Discussion

This Chapter glanced at other possibly interesting applications for minimax learning, beside detection tests. The proposed approach may be advantageous in certain applications where we want to maintain the recognition rate in worst-case scenarios.

The purpose of this Chapter was to propose other illustrations of the algorithms of Chapter 6. Of course, a comprehensive evaluation w.r.t. the existing methods in the considered field (i.e., machine learning) would require more elaborated studies.

Conclusions and Future Works

Conclusions

Before diving into the core of this research work, we found it important to present several technical topics connected to this dissertation. Those are fundamentals aspects which are required to understand the proposed contribution.

Both of the initial chapters were related to detection theory. In Chapter 1, basic principles of statistical hypothesis testing for a detection problem were reviewed, and the concept of minimax strategy was outlined. Conforming with the main applicative part of this dissertation, namely application to astrophysical hyperspectral data, Chapter 2 reviewed approaches for target detection in HSI. With respect to the two categories defined in HSI (Spectral Matching detector and Anomaly detector), our method can be viewed as a particular case of Spectral Matching, where the learned subspaces are robust in the worst-case scenario.

Moving away from detection topics, Chapter 3 was dedicated to dimension reduction techniques. We discussed several families of RD approaches: classification, clustering, and sparse learning. A branch of this approach, namely sparsity-based learning techniques, were particularly emphasized. We have recalled some classical sparse dictionary learning algorithms such as the MOD and K-SVD. In our work, (K)-SVD was the approach that acted as a benchmark to our proposed learning methods.

The preliminary observations reported in Chapter 4 showed that RD tests using low rank matrix approximation (SVD) yield good detection performances on average. However, the behavior of such approach (whose learning process is based on the MSE) tends to capture common forms of the elements in the reference library. The intrinsic diversity of the very large library $\boldsymbol{S}$ is not totally covered and this results in low detection power for some alternatives.

With these observations in mind, we tried in Chapter 5 to devise a RD test which is robust against all shapes of target signatures. This translated into a minimax (or worst-case) criterion. Designing this robust RD test entailed learning subspaces that maximize the worst probability of detection. Chapter 5 explained the corresponding optimization problems and investigated the theoretical framework in learning minimax dictionaries. In the first stage, we examined the simplest case of $K = 1$ atom. The exact solution of the one-dimensional minimax problem was found in (5.13), which can be resolved by a QP solver. Next, we examined the optimization problem for an arbitrary value of K. The resulting problem (5.6) is very intricate to be solved because it involves correlated variables. As a remedy, we proposed to study the corresponding bounds for the probability of detection and the probability of false alarm, resulting in a proxy minimax function. This function was used to design strategies to build greedy minimax algorithm as elaborated in Chapter 6.

Following the theoretical studies of Chapter 5, Chapter 6 elaborated several minimax learning algorithms. Three variants were proposed. The first was called greedy minimax, based on the approximation to the optimization problem (5.6). It can be viewed as a special case of Divisive clustering (top to down approach), where all data are initially set

in a cluster and then split into many clusters, but the number of clusters K in our case is fixed at initialization. The second algorithm (named K-minimax) relied on injecting one-dimensional minimax solution in the dictionary update stage of K-SVD algorithm (or more precisely gain-shape VQ). The third kind of algorithm (SKM-minimax) combined the SKM clustering technique on the unit hypersphere with the one-dimensional minimax solution.

Some applications of the proposed approaches were presented in Part III of this dissertation. The principal application was presented in Chapter 7, where we focused on the target detection of Lyman-α emission lines in astrophysical data cubes. The results showed that RD tests using minimax learned dictionaries attained the main objectives of this dissertation. The proposed testing approach is far less complex than testing the full library (via the *Max* test over $\boldsymbol{S}$), and in the same time robust against all alternatives under $\mathcal{H}_1$. Their worst-case performances are in fact as high as those of *Max* test (which would test all alternatives). Moreover, quite good average performances were also obtained for a number of atoms K sufficiently large.

Apart from this detection application, we also outlined in Chapter 8 other possible applications that may be interesting in a worst-case recognition framework. Minimax dictionaries were learned from known databases of faces and handwritten digits. The one-dimensional minimax atom showed a much better worst-case recognition rate w.r.t. a (naive) approach based on SVD.

In summary, the originality of this dissertation lies in the proposition of detection tests for composite hypothesis testing that are optimized w.r.t. a minimax criterion. Recognizing the modern problem of computation complexity when testing a very large known library $\boldsymbol{S}$, we pointed out the importance to devise RD tests. The proposed minimax RD tests appear original w.r.t. the literature and also relevant for many applications.

This type of strategy can be very advantageous in all domains where detection concerns known (up to an amplitude factor) alternatives (which are possibly in large number), and where it is important that the detection is robust w.r.t. all possible alternatives. Other instances of such application include the determination of cancerous cells (where the known alternatives are cells taken from blood samples or bone marrow samples) or for the detection of gas leakage in pipelines (where the known alternatives are chemical spectra collected from gas flow measurements).

Future works

We list below several future works that can be conducted following the studies of this dissertation.

1. Real data from MUSE instrument

Firstly, the minimax subspaces learning for detection tests proposed in this dissertation were mainly evaluated on simulated astrophysical data (presented in Chapter 7). It will be interesting to see the outcome of these approaches for the application of real data cubes from the MUSE instrument which will be available in the future months.

2. One target Vs. the rest of alternatives

Problem (4.1) can be viewed as *one-class* detection problem, where the model is pure noise under the null hypothesis against a "one among many" alternative model. An extension branch that one can think of is a *multi-class* detection problem, where the model is one target against the rest of the alternatives:

$$\begin{cases} \mathcal{H}_0 & : \boldsymbol{y} = \boldsymbol{s}_\ell + \boldsymbol{n}, \quad \boldsymbol{n} \sim \mathcal{N}(\boldsymbol{0}, \mathbf{I}) \\ \mathcal{H}_1 & : \boldsymbol{y} = \boldsymbol{S}'\boldsymbol{\alpha} + \boldsymbol{n}, \quad \|\boldsymbol{\alpha}\|_0 = 1, \ \boldsymbol{\alpha} \text{ unknown} \end{cases} \tag{8.1}$$

where $\boldsymbol{y}$, $\boldsymbol{n}$ and $\boldsymbol{s}_\ell \in \mathbb{R}^N$, $\boldsymbol{s}_\ell \notin \boldsymbol{S}'$, $\boldsymbol{s}_\ell$ and $\boldsymbol{S}' \in \mathbb{R}^{N \times L}$ are column-normalized. Assuming that $\boldsymbol{s}_\ell$ is known under $\mathcal{H}_0$, the GLRT for model (8.1) is:

$$T_{\text{GLR}}(\boldsymbol{y}) = \max_{\boldsymbol{\alpha}:\|\boldsymbol{\alpha}\|_0=1} \frac{p(\boldsymbol{y}|\boldsymbol{S}'\boldsymbol{\alpha})}{p(\boldsymbol{y}|\boldsymbol{s}_\ell)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma' \tag{8.2}$$

Maximizing the numerator implies: $\max_{\boldsymbol{\alpha}:\|\boldsymbol{\alpha}\|_0=1} \ p(\boldsymbol{y}|\boldsymbol{S}'\boldsymbol{\alpha}) = \min_{\boldsymbol{\alpha}:\|\boldsymbol{\alpha}\|_0=1} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{S}'\boldsymbol{\alpha}\|_2^2$. For index i fixed ($\forall i \neq \ell$), we obtained $\alpha_i^{\text{ML}} = \boldsymbol{s}_i^\top \boldsymbol{y}$, then $\hat{i} = \arg\max_{i=1,\dots,L} |\boldsymbol{s}_i^\top \boldsymbol{y}|$. Taking the logarithm of the GLRT and injecting $\alpha_{\hat{i}}^{\text{ML}}$ in (8.2) yields:

$$T_{\text{GLR}}(\boldsymbol{y}, \boldsymbol{s}_\ell, \boldsymbol{S}') = \max_{\substack{i=1,\dots,L \\ i \neq \ell}} (\boldsymbol{s}_i^\top \boldsymbol{y})^2 - 2\boldsymbol{s}_\ell^\top \boldsymbol{y} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma, \quad \text{where } \gamma = 2\ln\gamma' - \|\boldsymbol{s}_\ell\|_2^2. \tag{8.3}$$

This is an interesting problem for future works (along with generalization to unknown amplitudes under $\mathcal{H}_0$. This goes in the direction of extending MSD [Scharf & Friedlander 1994] to minimax principle).

3. RD Model with $k > 1$ sparse constraint under $\mathcal{H}_1$

The RD model (5.1) imposes 1-sparse constraint under $\mathcal{H}_1$. We have seen that the Lyman-α spectral library $\boldsymbol{S}_{9745}$ possesses intrinsic diversity. Some of the spectral lines have two peaks (see Figures 7.8(d) and 7.8(e)). Thus, it may be interesting to study the case where the sparse constraint under $\mathcal{H}_1$ is set to superior to one:

$$\begin{cases} \mathcal{H}_0 & : \boldsymbol{x} = \boldsymbol{n}, & \boldsymbol{n} \sim \mathcal{N}(\boldsymbol{0}, \mathbf{I}) \\ \mathcal{H}_1 & : \boldsymbol{x} = \boldsymbol{D}\boldsymbol{\beta} + \boldsymbol{n}, & \|\boldsymbol{\beta}\|_0 = k,\ k > 1 \end{cases} . \tag{8.4}$$

4. Best number of K atoms

In clustering and classification tasks, there is no definitive procedure allowing to obtain the number of optimal clusters (or classes) K. We have presented in Section 7.5 two possible ways to determine K. This value depends on the intrinsic diversity of the samples in the reference library $\boldsymbol{S}$ but also on the SNR. We have not investigated the dependencies of the best value of K w.r.t. SNR. This could easily be done through simulations from which theoretical insight would be foreseen.

5. Concatenation of greedy minimax and K-SVD dictionaries

We have seen throughout this dissertation that testing using K-SVD dictionaries yielded a very good average detection performance, while testing using minimax dictionaries optimized the worst-case detection performances. In practice, an efficient strategy would be to concatenate, say, greedy minimax dictionaries $\boldsymbol{D}^*$ with K-SVD dictionary $\boldsymbol{D}^{\text{K-SVD}}$. GLR testing using $\boldsymbol{D}' = [\boldsymbol{D}^* \, \boldsymbol{D}^{\text{K-SVD}}]$ should give both a good worst-case and average detection performances. Note however that the issue of optimizing the number of columns remains.

Appendix A

Appendix of Part I

A.1 Proof of Neyman-Pearson Lemma

Proof. We seek the rejection region of $\mathcal{H}_0$, in which we decide $\mathcal{H}_1$: $R_1 = \{\boldsymbol{x} : \mathcal{H}_1\}$ so that $\mathrm{P_{Det}}$ is maximum for a fixed $\mathrm{P_{FA}} = \alpha$, as described below

$$\max \mathrm{P_{Det}} = \max \int_{R_1} p(\boldsymbol{x}; \mathcal{H}_1)\, \mathrm{d}\boldsymbol{x} \quad \text{subject to } \mathrm{P_{FA}} = \int_{R_1} p(\boldsymbol{x}; \mathcal{H}_0)\, \mathrm{d}\boldsymbol{x} = \alpha. \tag{A.1}$$

In the problem above, $p(\boldsymbol{x}; \mathcal{H}_i)$ is the probability distribution function of $\boldsymbol{x}$ under hypothesis $\mathcal{H}_i$, $i = 0, 1$. Let us consider the following Lagrangian $\mathcal{L}$

$$\mathcal{L} = \mathrm{P_{Det}} + \lambda(\mathrm{P_{FA}} - \alpha), \tag{A.2}$$

where $\lambda \in \mathbb{R}$ is a Lagrange multiplier. Injecting the corresponding expression of $\mathrm{P_{Det}}$ and $\mathrm{P_{FA}}$ of (A.1) in $\mathcal{L}$ yields

$$\begin{aligned} \mathcal{L} &= \int_{R_1} p(\boldsymbol{x}; \mathcal{H}_1)\, \mathrm{d}\boldsymbol{x} + \lambda \left(\int_{R_1} p(\boldsymbol{x}; \mathcal{H}_0)\, \mathrm{d}\boldsymbol{x} - \alpha \right) \\ &= \int_{R_1} \left(p(\boldsymbol{x}; \mathcal{H}_1) + \lambda\, p(\boldsymbol{x}; \mathcal{H}_0) \right) \mathrm{d}\boldsymbol{x} - \lambda\alpha. \end{aligned} \tag{A.3}$$

To maximize $\mathcal{L}$, $\boldsymbol{x}$ should be included in R_1 if

$$p(\boldsymbol{x}; \mathcal{H}_1) + \lambda\, p(\boldsymbol{x}; \mathcal{H}_0) > 0 \;\Rightarrow\; \frac{p(\boldsymbol{x}; \mathcal{H}_1)}{p(\boldsymbol{x}; \mathcal{H}_0)} > -\lambda. \tag{A.4}$$

LR is nonnegative, hence if we denote by $\gamma = -\lambda$, a positive threshold ($\gamma > 0$) obtained from $\mathrm{P_{FA}} = \alpha$, we decide $\mathcal{H}_1$ when

$$\frac{p(\boldsymbol{x}; \mathcal{H}_1)}{p(\boldsymbol{x}; \mathcal{H}_0)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{>}} \gamma, \tag{A.5}$$

which proves the Neyman-Pearson lemma (as described in lemma 1). □

A.2 Proof of the Bayes detector that minimizes the probability of error

Proof. Let us denote the Bayes risk as $\mathcal{R}$, and A_{ij} are constants correspond to the cost if we decide $\mathcal{H}_i$ but $\mathcal{H}_j$ is true. The Bayes risk is defined as

$$\begin{aligned}\mathcal{R} =& \sum_{i=0}^{1}\sum_{j=0}^{1} A_{ij}\mathbb{P}(\mathcal{H}_i|\mathcal{H}_j)\mathbb{P}(\mathcal{H}_j)\\ \mathcal{R} =& A_{00}\mathbb{P}(\mathcal{H}_0|\mathcal{H}_0)\mathbb{P}(\mathcal{H}_0) + A_{01}\mathbb{P}(\mathcal{H}_0|\mathcal{H}_1)\mathbb{P}(\mathcal{H}_1) + A_{10}\mathbb{P}(\mathcal{H}_1|\mathcal{H}_0)\mathbb{P}(\mathcal{H}_0)\\ &+ A_{11}\mathbb{P}(\mathcal{H}_1|\mathcal{H}_1)\mathbb{P}(\mathcal{H}_1). \end{aligned} \tag{A.6}$$

Recall that the probability error (1.28) defined in Section 1.5 is

$$\mathrm{P_E} = \mathbb{P}(\mathcal{H}_0|\mathcal{H}_1)\,\mathbb{P}(\mathcal{H}_1) + \mathbb{P}(\mathcal{H}_1|\mathcal{H}_0)\,\mathbb{P}(\mathcal{H}_0),$$

where $\mathcal{R} = \mathrm{P_E}$ if $A_{01} = A_{10} = 1$ and $A_{00} = A_{11} = 0$. Thus, to proof that relation (1.29) allows to minimize $\mathrm{P_E}$, we can use the definition of $\mathcal{R}$. Denote by $R_1 = \{\boldsymbol{x} : \mathcal{H}_1\}$, the rejection region of $\mathcal{H}_0$, and $R_0 = \{\boldsymbol{x} : \mathcal{H}_0\}$ the acceptance region of $\mathcal{H}_0$. We can write from (A.6)

$$\begin{aligned}\mathcal{R} =& A_{00}\mathbb{P}(\mathcal{H}_0)\int_{R_1} p(\boldsymbol{x}|\mathcal{H}_0)\mathrm{d}\boldsymbol{x} + A_{01}\mathbb{P}(\mathcal{H}_1)\int_{R_0} p(\boldsymbol{x}|\mathcal{H}_1)\mathrm{d}\boldsymbol{x} + A_{10}\mathbb{P}(\mathcal{H}_0)\int_{R_1} p(\boldsymbol{x}|\mathcal{H}_0)\mathrm{d}\boldsymbol{x}\\ &+ A_{11}\mathbb{P}(\mathcal{H}_1)\int_{R_1} p(\boldsymbol{x}|\mathcal{H}_1)\mathrm{d}\boldsymbol{x}. \end{aligned} \tag{A.7}$$

Using

$$\int_{R_0} p(\boldsymbol{x}|\mathcal{H}_i)\mathrm{d}\boldsymbol{x} = 1 - \int_{R_1} p(\boldsymbol{x}|\mathcal{H}_i)\mathrm{d}\boldsymbol{x},$$

equation (A.7) becomes

$$\begin{aligned}\mathcal{R} =& A_{00}\mathbb{P}(\mathcal{H}_0) + A_{01}\mathbb{P}(\mathcal{H}_1)\\ &+ \int_{R_1} \left\{ \left[A_{10}\mathbb{P}(\mathcal{H}_0) - A_{00}\mathbb{P}(\mathcal{H}_0)\right] p(\boldsymbol{x}|\mathcal{H}_0) \;+\; \left[A_{11}\mathbb{P}(\mathcal{H}_1) - A_{01}\mathbb{P}(\mathcal{H}_1)\right] p(\boldsymbol{x}|\mathcal{H}_1) \right\} \mathrm{d}\boldsymbol{x}. \end{aligned} \tag{A.8}$$

From the integral in (A.8), we decide $\mathcal{H}_1$ if

$$(A_{10} - A_{00})\,\mathbb{P}(\mathcal{H}_0)\,p(\boldsymbol{x}|\mathcal{H}_0) < (A_{01} - A_{11})\,\mathbb{P}(\mathcal{H}_1)\,p(\boldsymbol{x}|\mathcal{H}_1).$$

If we assume that $A_{10} > A_{00}$ and $A_{01} > A_{11}$, we have

$$\frac{p(\boldsymbol{x}|\mathcal{H}_1)}{p(\boldsymbol{x}|\mathcal{H}_0)} > \frac{(A_{10} - A_{00})\,p(\mathcal{H}_1)}{(A_{01} - A_{11})\,p(\mathcal{H}_0)} = \gamma, \tag{A.9}$$

which is equivalent to the relation (1.29), and corresponds to the minimum $\mathrm{P_E}$ detector. □

A.3 Examples of several detection methods in literature

Example A.3.1. LR test statistic for testing a known mean vector.

Consider the following hypotheses model for normally distributed data $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma})$

$$\begin{cases} \mathcal{H}_0 & : \boldsymbol{\theta} = \mathbf{0} \\ \mathcal{H}_1 & : \boldsymbol{\theta} = \boldsymbol{\theta}_1 \end{cases}, \tag{A.10}$$

where $\boldsymbol{x} \in \mathbb{R}^N$, $\boldsymbol{\theta}_1$ is a known vector parameter under $\mathcal{H}_1$ and $\boldsymbol{\Sigma}$ is a known covariance matrix. Applying the definition of LR test (1.9)-(1.10) to the above hypothesis model implies

$$\text{LR} : \Lambda(\boldsymbol{x}) := \frac{p(\boldsymbol{x}; \boldsymbol{\theta}_1)}{p(\boldsymbol{x}; \mathbf{0})} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma', \tag{A.11}$$

where under $\mathcal{H}_1$

$$p(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{\det(\boldsymbol{\Sigma}^{-1/2})}{(2\pi)^{N/2}} \exp\left(\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\theta}_1)^\top \Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\theta}_1)\right). \tag{A.12}$$

The likelihood ratio is

$$\begin{aligned} \Lambda(\boldsymbol{x}) &= \frac{\frac{\det(\boldsymbol{\Sigma}^{-1/2})}{(2\pi)^{N/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\theta}_1)^\top \Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\theta}_1)\right)}{\frac{\det(\boldsymbol{\Sigma}^{-1/2})}{(2\pi)^{N/2}} \exp\left(-\frac{1}{2}\boldsymbol{x}^\top \Sigma^{-1}\boldsymbol{x}\right)} \\ &= \exp\left(\frac{1}{2}\left(2\boldsymbol{\theta}_1^\top \Sigma^{-1}\boldsymbol{x} - \boldsymbol{\theta}_1^\top \Sigma^{-1}\boldsymbol{\theta}_1\right)\right) \\ &= \exp\left(\boldsymbol{\theta}_1^\top \Sigma^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\theta}_1^\top \Sigma^{-1}\boldsymbol{\theta}_1\right). \end{aligned} \tag{A.13}$$

Taking the logarithm of (A.13) yields

$$\text{LR} : \boldsymbol{\theta}_1^\top \Sigma^{-1}\boldsymbol{x} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma, \tag{A.14}$$

where $\gamma = \ln \gamma' + \frac{1}{2}\boldsymbol{\theta}_1^\top \Sigma^{-1}\boldsymbol{\theta}_1$. ■

Example A.3.2. GLR test statistic for testing an unknown mean vector

Consider the following hypotheses model for normally distributed data $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma})$

$$\begin{cases} \mathcal{H}_0 & : \boldsymbol{\theta} = \mathbf{0} \\ \mathcal{H}_1 & : \boldsymbol{\theta} \neq \mathbf{0} \end{cases}, \tag{A.15}$$

where $\boldsymbol{x} \in \mathbb{R}^N$, $\boldsymbol{\theta}$ is unknown and $\boldsymbol{\Sigma}$ is a known covariance matrix. Applying the definition

of GLRT (1.13) to the model (A.15) implies

$$T_{\mathrm{GLR}}(\boldsymbol{x}) := \frac{\max\limits_{\boldsymbol{\theta}\neq 0} p(\boldsymbol{x};\boldsymbol{\theta})}{p(\boldsymbol{x};\mathbf{0})} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma', \tag{A.16}$$

where under $\mathcal{H}_1$

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \frac{\det(\boldsymbol{\Sigma}^{-1/2})}{(2\pi)^{N/2}} \exp\left(\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\theta})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\theta})\right). \tag{A.17}$$

We shall first calculate the ML estimate of the unknown parameter, by maximizing the numerator of (A.16)

$$\begin{aligned}\widehat{\boldsymbol{\theta}}^{\mathrm{ML}} &= \arg\max_{\boldsymbol{\theta}\neq 0} p(\boldsymbol{x};\boldsymbol{\theta}) \\ &= \arg\min_{\boldsymbol{\theta}\neq 0} \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\theta})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\theta}).\end{aligned} \tag{A.18}$$

Computing the derivative of (A.18) w.r.t. $\boldsymbol{\theta}$ and equating to 0 yields

$$\Sigma^{-1}(\boldsymbol{\theta}-\boldsymbol{x})\big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{\mathrm{ML}}} = 0 \tag{A.19}$$

$$\widehat{\boldsymbol{\theta}}^{\mathrm{ML}} = \boldsymbol{x}. \tag{A.20}$$

Injecting $\widehat{\boldsymbol{\theta}}^{\mathrm{ML}}$ in the test statistic (A.16) leads to

$$\begin{aligned}T_{\mathrm{GLR}}(\boldsymbol{x}) &= \frac{\frac{\det(\boldsymbol{\Sigma}^{-1/2})}{(2\pi)^{N/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\theta}^{\mathrm{ML}})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\theta}^{\mathrm{ML}})\right)}{\frac{\det(\boldsymbol{\Sigma}^{-1/2})}{(2\pi)^{N/2}} \exp\left(-\frac{1}{2}\boldsymbol{x}^\top \Sigma^{-1}\boldsymbol{x}\right)} \\ &= \exp\left(\frac{1}{2}\left(\boldsymbol{x}^\top \Sigma^{-1}\boldsymbol{x}\right)\right).\end{aligned} \tag{A.21}$$

Taking the logarithm of (A.21) implies the following test

$$T_{\mathrm{GLR}}(\boldsymbol{v}) = \|\boldsymbol{v}\|_2^2 \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma, \tag{A.22}$$

where $\boldsymbol{v} = \Sigma^{-1/2}\boldsymbol{x}$ and $\gamma = 2\ln\gamma'$. This kind of test statistic (A.22) is an Energy Detector, where the energy of the "whitened" signal $\boldsymbol{v}$ is compared to a threshold. The SNR of each data component in this example may be different, because the test statistic depends on the elements of Σ. ■

Example A.3.3. Bayesian test statistic.

Let us consider $\boldsymbol{x} = \boldsymbol{\theta} + \boldsymbol{n}$ with the following hypotheses

$$\begin{cases} \mathcal{H}_0 & : \boldsymbol{\theta} = \mathbf{0} \\ \mathcal{H}_1 & : \boldsymbol{\theta} \sim p_1(\boldsymbol{\theta}) \end{cases} \tag{A.23}$$

where $\boldsymbol{x}$, $\boldsymbol{n}$, and $\boldsymbol{\theta} \in \mathbb{R}^N$, $\boldsymbol{n}$ is an independent and identically distributed (i.i.d.) noise, $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_1]^\top$ and $p_1(\boldsymbol{\theta})$ is the prior probability of the random parameter under $\mathcal{H}_1$. Applying the definition (1.30) to the current model conducts to

$$T_{\text{Bayes}}(\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\mathcal{H}_1)}{p(\boldsymbol{x}|\mathcal{H}_0)} = \frac{\int p(\boldsymbol{x}|\boldsymbol{\theta})\, p_1(\boldsymbol{\theta})\, d\boldsymbol{\theta}}{p(\boldsymbol{x}|\mathbf{0})} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma, \tag{A.24}$$

and for a separable prior under $\mathcal{H}_1$, $p_1(\boldsymbol{\theta}) = \prod_{n=1}^{N} p_{1_n}(\theta_n)$, we have

$$T_{\text{Bayes}}(\boldsymbol{x}) = \prod_{n=1}^{N} \frac{\int_{\mathbb{R}} p(x_n|\theta_n)\, p_{1_n}(\theta_n)\, d\theta_n}{p(x_n|\mathbf{0})} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma. \tag{A.25}$$

■

A.4 Analysis sparse model

Elad, Milanfar and Rubinstein were among the first who highlighted two different models in sparse regression problems: synthesis and analysis [Elad *et al.* 2007]. The analysis sparse model is described below.

Assume that we have a signal $\boldsymbol{s}$. The analysis model (as illustrated in Figure A.1) imposes sparsity constraint on $\boldsymbol{\Omega s}$, where $\boldsymbol{\Omega} \in \mathbb{R}^{W \times N}$ is the analysis dictionary (typically, $W > N$). Contrary to the synthesis model, this model accentuates on the locations and number of *zeros* in $\boldsymbol{\Omega s}$ (which is called *co-sparsity* denoted by l). The sparsity of $\boldsymbol{\Omega s}$ is related to l by $\|\boldsymbol{\Omega s}\|_0 = W - l$. The optimization problem for an analysis sparse approximation problem can be written as

$$\widehat{\boldsymbol{s}} = \arg\min_{\boldsymbol{x}} \|\boldsymbol{\Omega x}\|_0 \quad \text{subject to } \|\widehat{\boldsymbol{s}} - \boldsymbol{x}\|_2^2 \leq \varepsilon, \tag{A.26}$$

or

$$\widehat{\boldsymbol{s}} = \arg\min_{\boldsymbol{s}} \|\widehat{\boldsymbol{s}} - \boldsymbol{x}\|_2^2 \quad \text{subject to } \|\boldsymbol{\Omega x}\|_0 \leq W - l. \tag{A.27}$$

The rows in $\boldsymbol{\Omega}$ that are orthogonal to $\boldsymbol{s}$ (shown in orange in Figure A.1) define the *co-support* of $\boldsymbol{s}$.

In a dictionary learning context, (A.27) in matrix form leads to the optimization problem

$$\min_{\boldsymbol{\Omega},\boldsymbol{X}} \|\boldsymbol{S}-\boldsymbol{X}\|_F^2 \quad \text{subject to } \|\boldsymbol{\Omega}\boldsymbol{x}_i\|_0 \leq W-l,\ \forall i=1,\ldots,L, \qquad \|\boldsymbol{\omega}_j\|_2=1,\ \forall j=1,\ldots,W, \tag{A.28}$$

where $\boldsymbol{\omega}_j$ is the rows of $\boldsymbol{\Omega}$ and $\boldsymbol{x}_i$ the columns of $\boldsymbol{X}$.

Following (A.27) and (A.28), quite a number of algorithms were proposed for analysis learning. For instance, the counterpart of OMP in analysis is called Greedy Analysis Pursuit (GAP) [Nam *et al.* 2011]. At initialization, GAP algorithm fixes the co-sparsity l and sets all rows $\boldsymbol{\omega}_j \in \boldsymbol{\Omega}^{(0)}$, $j=1,\ldots,W$ to be the co-support of $\boldsymbol{s}$ (i.e., all rows in $\boldsymbol{\Omega}^{(0)}$ are orthogonal to $\boldsymbol{s}$). Then, the non-zero elements in $\boldsymbol{\Omega}\boldsymbol{s}$ are identified, reducing the size of the co-support from W to l.

In [Rubinstein *et al.* 2012], the co-support is initialized to zero and the zero elements in $\boldsymbol{\Omega}\boldsymbol{s}$ are iteratively identified. So the co-support increases in dimension until a stopping rule. This analysis sparse coding step is named the Backward Greedy algorithm. This is used together with a dictionary update step via SVD, in an algorithm called Analysis K-SVD algorithm, which parallels with the Synthesis K-SVD algorithm described in Example A.5.3.

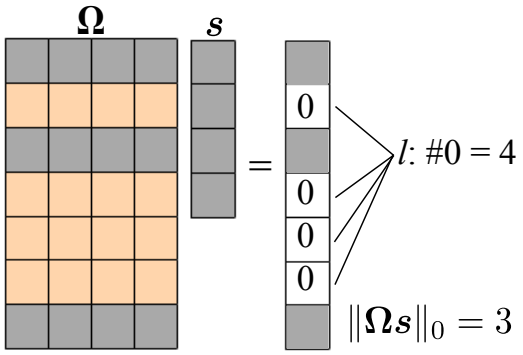


Figure A.1: The concept of analysis sparse modeling. $\boldsymbol{\Omega} \in \mathbb{R}^{W\times N}$ is the analysis dictionary. Sparsity constraint is imposed on $\boldsymbol{\Omega}\boldsymbol{s}$, by the *co-sparsity* l (which is the number of zeros in $\boldsymbol{\Omega}\boldsymbol{s}$). The rows in $\boldsymbol{\Omega}$ that are orthogonal to $\boldsymbol{s}$ define the *co-support* of $\boldsymbol{s}$ (shown in orange). Here, the size of the co-support is 4.

A.5 The algorithms of several RD learning techniques

A.5.1 K-Means algorithm

Algorithm 3 K-Means

Inputs: Data set $\boldsymbol{S} \in \mathbb{R}^{N \times L} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_L]$, number of clusters K.

Initialization: Set $t = 0$, choose initial centroids $\boldsymbol{M}^{(0)} \in \mathbb{R}^{N \times K} = [\boldsymbol{\mu}_1^{(0)}, \ldots, \boldsymbol{\mu}_K^{(0)}]$ (e.g., random K data vectors from $\boldsymbol{S}$).

Repeat until convergence (stopping rule):

- *Clustering stage*:

  for $i = 1, \ldots, L$

  Assign $\boldsymbol{s}_i$ to the cluster $C_j^{(t)}$ of centroid $\boldsymbol{\mu}_j^{(t)}$ of $\boldsymbol{M}^{(t)}$

  $C_j^{(t)} = \{\|\boldsymbol{s}_i - \boldsymbol{\mu}_j^{(t)}\|_2^2 < \|\boldsymbol{s}_i - \boldsymbol{\mu}_l^{(t)}\|_2^2,\ \forall j \neq l,\ j = 1, \ldots, K\}$,

  end

This yields K clusters $C_1^{(t)}, \ldots, C_K^{(t)}$.

- *Centroids update stage*:

  for $j = 1, \ldots, K$

  $v =$ size cluster $C_j^{(t)}$,

  $\boldsymbol{\mu}_j^{(t+1)} = \frac{1}{v} \sum_{c=1}^{v} \boldsymbol{s}_c \in C_j^{(t)}$,

  end

Set $t = t + 1$,
$\boldsymbol{M}^{(t)} = [\boldsymbol{\mu}_1^{(t)}, \ldots, \boldsymbol{\mu}_K^{(t)}]$,

Output: The centroids $\boldsymbol{M} \in \mathbb{R}^{N \times K} = \boldsymbol{M}^{(t)}$ and the K clusters $C_1^{(t)}, \ldots, C_K^{(t)}$.

A.5.2 MOD algorithm

Algorithm 4 MOD

Inputs: Data set $\boldsymbol{S} \in \mathbb{R}^{N\times L} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_L]$, number of atoms K, sparsity k, error threshold ε.

Initialization: Set $t = 0$, choose an initial dictionary $\boldsymbol{D}^{(0)} \in \mathbb{R}^{N\times K} = [\boldsymbol{d}_1, \ldots, \boldsymbol{d}_K]$ (e.g., random vectors). Perform column-normalization on $\boldsymbol{D}$.

Set $t = 1$, set error $e^{(t)} = 1$,
Repeat until $e^{(t)} \leq \varepsilon$**:**

- *Sparse coding stage*:
  for $i = 1, \ldots, L$
  Compute the sparse vectors $\widehat{\boldsymbol{\alpha}_i}$ via any pursuit algorithm (e.g., MP, OMP)

$$\widehat{\boldsymbol{\alpha}_i}^{(t)} \approx \arg\min_{\boldsymbol{\alpha}_i} \|\boldsymbol{s}_i - \boldsymbol{D}^{(t-1)}\boldsymbol{\alpha}_i\|_2^2 \quad \text{subject to } \|\boldsymbol{\alpha}_i\|_0 \leq k,$$

  end
  This yields sparse matrix $\boldsymbol{A}^{(t)} = [\widehat{\boldsymbol{\alpha}_1}^{(t)}, \ldots, \widehat{\boldsymbol{\alpha}_L}^{(t)}]$.

- *Dictionary update stage*:
  MOD dictionary is obtained by the general solution to a least squares problem

$$\boldsymbol{D}^{(t)} = \arg\min_{\boldsymbol{D}} \|\boldsymbol{S} - \boldsymbol{D}\boldsymbol{A}^{(t)}\|_F^2 = \boldsymbol{S}\boldsymbol{A}^{\top(t)}(\boldsymbol{A}^{(t)}\boldsymbol{A}^{\top(t)})^{-1} = \boldsymbol{S}\boldsymbol{A}^{+(t)},$$

  where $\boldsymbol{A}^+$ is the Moore-Penrose pseudoinverse of $\boldsymbol{A}$.

Set $t = t + 1$,
Compute error $e^{(t)} = \|\boldsymbol{S} - \boldsymbol{D}^{(t-1)}\boldsymbol{A}^{(t-1)}\|_F^2$,

Output: MOD optimized dictionary of K atoms: $\boldsymbol{D}_K^{\text{MOD}} = \boldsymbol{D}^{(t-1)}$.

A.5.3 K-SVD algorithm.

Algorithm 5 K-SVD

Inputs: Data set $\boldsymbol{S} \in \mathbb{R}^{N\times L} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_L]$, number of atoms K, sparsity k.

Initialization: Set $t = 0$, choose an initial dictionary $\boldsymbol{D}^{(0)} \in \mathbb{R}^{N\times K} = [\boldsymbol{d}_1, \ldots, \boldsymbol{d}_K]$ (e.g., random vectors). Perform column-normalization on $\boldsymbol{D}$.

Set $t = 1$,

Repeat until convergence (stopping rule):

• *Sparse coding stage*:

for $i = 1, \ldots, L$

Compute the sparse vectors $\widehat{\boldsymbol{\alpha}_i}$ via any pursuit algorithm

$$\widehat{\boldsymbol{\alpha}_i}^{(t-1)} \approx \arg\min_{\boldsymbol{\alpha}_i} \|\boldsymbol{s}_i - \boldsymbol{D}^{(t-1)}\boldsymbol{\alpha}_i\|_2^2 \quad \text{subject to } \|\boldsymbol{\alpha}_i\|_0 \le k,$$

end

This yields sparse matrix $\boldsymbol{A}^{(t-1)} = [\widehat{\boldsymbol{\alpha}_1}^{(t-1)}, \ldots, \widehat{\boldsymbol{\alpha}_L}^{(t-1)}]$.

• *Dictionary update stage*:

for $j = 1, \ldots, K$

Compute the overall representation error matrix $\boldsymbol{E}_j$, where all the dictionary atoms are froze, except one ($\boldsymbol{d}_j$):

$$\boldsymbol{E}_j = \boldsymbol{S} - \sum_{l\neq j} \boldsymbol{d}_l\boldsymbol{\alpha}_l.$$

For row j in $\boldsymbol{A}^{(t-1)}$, locate the $\alpha_{j,i}$ that use(s) $\boldsymbol{d}_j$:

$$w_j = \{i | i \le i \le L, |\alpha_{j,i} \neq 0\}.$$

Restrict the columns in $\boldsymbol{E}_j$ according to the indexes in w_j, yielding $\boldsymbol{E}_j^R$. Then decompose $\boldsymbol{E}_j^R$ according to SVD (3.2), and update the atom $\boldsymbol{d}_j$ and the sparse coefficients $\alpha_{j,R}$ by the best rank-one approximation of $\boldsymbol{E}_j^R$ in Frobenius norm:

$$\boldsymbol{E}_j^R = \mathbf{U}\boldsymbol{\Sigma}_r\mathbf{V}^\top, \; r = \text{rank of } \mathrm{E}_j^R, \quad \mathbf{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N], \quad \mathbf{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_R],$$

$$\boldsymbol{d}_j^{\mathrm{SVD}} = \boldsymbol{u}_1,$$

$$\alpha_{j,R}^{(t)} = \boldsymbol{v}_1\, \boldsymbol{\Sigma}_r(1,1).$$

end

The updated matrices are $\boldsymbol{D}^{(t)} = [\boldsymbol{d}_1^{\mathrm{SVD}}, \ldots, \boldsymbol{d}_K^{\mathrm{SVD}}]$ and $\boldsymbol{A}^{(t)} = [\boldsymbol{\alpha}_1^{(t)}, \ldots, \boldsymbol{\alpha}_L^{(t)}]$.

Set $t = t + 1$,

Output: K-SVD optimized dictionary of K atoms: $\boldsymbol{D}_K^{\text{K-SVD}} = \boldsymbol{D}^{(t-1)}$.

Appendix B

Appendix of Part II

B.1 Proof of Proposition 1

Proof. The following well-known result will be used :

Proposition 4. *Projections matrices (e.g., p.291 of [Sen & Srivastava 1990]).*
Let $\boldsymbol{A}$ be a symmetric ($\boldsymbol{A}^\top = \boldsymbol{A}$) and idempotent ($\boldsymbol{A}\boldsymbol{A} = \boldsymbol{A}$) matrix of rank k, and let $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{I})$. Then $\boldsymbol{z}^\top \boldsymbol{A}\boldsymbol{z}$ is distributed as a noncentral chi-squared distribution noted $\chi^2_{k,\lambda}$ with k degrees of freedom and noncentrality parameter $\lambda = \boldsymbol{\mu}^\top \boldsymbol{A}\boldsymbol{\mu}$.

Under $\mathcal{H}_0$ of (4.1), evaluating the $\mathrm{P_{FA}}(\boldsymbol{S})$ for (4.9) yields

$$\begin{aligned} \mathrm{P_{FA}}(\boldsymbol{S}, L) &= \mathbb{P}(\max_{i=1,\dots,L} (\boldsymbol{s}_i^\top \boldsymbol{n})^2 > \gamma^2; \mathcal{H}_0) \\ &= 1 - \mathbb{P}((\boldsymbol{s}_1^\top \boldsymbol{n})^2 < \gamma^2, \dots, (\boldsymbol{s}_L^\top \boldsymbol{n})^2 < \gamma^2) \end{aligned} \tag{B.1}$$

The $\boldsymbol{s}_i^\top \boldsymbol{n}$ being normally distributed and decorrelated for $i \neq j$ (since $\boldsymbol{S}$ is assumed orthonormal) they are independent. Noting that $(\boldsymbol{s}_i^\top \boldsymbol{n})^2 = \boldsymbol{n}^\top (\boldsymbol{s}_i \boldsymbol{s}_i^\top) \boldsymbol{n}$ and applying Proposition 4 with $\lambda = 0$ leads to

$$\begin{aligned} \mathrm{P_{FA}}(\boldsymbol{S}, L) &= 1 - \prod_{i=1}^{L} \mathbb{P}((\boldsymbol{s}_i^\top \boldsymbol{n})^2 < \gamma^2) \\ \mathrm{P_{FA}}(L) &= 1 - (\Phi_{\chi^2_1}(\gamma^2))^L, \end{aligned} \tag{B.2}$$

where $\Phi_{\chi^2_1}$ is the CDF of χ^2_1. Inverting Eq. (B.2) yields $\gamma^2 = \Phi^{-1}_{\chi^2_1}\left((1 - \mathrm{P_{FA}})^{\frac{1}{L}}\right)$ where the dependence of $\mathrm{P_{FA}}$ w.r.t. $\boldsymbol{S}$ has been dropped.

Under $\mathcal{H}_1$ of (4.1), using a similar reasoning and the orthogonality of $\boldsymbol{S}$ conducts to

$$\begin{aligned} \mathrm{P_{Det}}(\boldsymbol{s}_\ell, \boldsymbol{S}, L) &= \mathbb{P}(\max_{i=1,\dots,L} (\boldsymbol{s}_i^\top (\boldsymbol{s}_\ell + \boldsymbol{n}))^2 > \gamma^2; \mathcal{H}_1, \boldsymbol{s}_\ell) \\ &= 1 - \mathbb{P}((\boldsymbol{s}_1^\top (\boldsymbol{s}_\ell + \boldsymbol{n}))^2 < \gamma^2, \dots, (\boldsymbol{s}_L^\top (\boldsymbol{s}_\ell + \boldsymbol{n}))^2 < \gamma^2) \\ &= 1 - \Phi_{\chi^2_{1,\lambda}}(\gamma^2) \prod_{\substack{i=1 \\ i \neq \ell}}^{L-1} \mathbb{P}((\boldsymbol{s}_i^\top \boldsymbol{n})^2 < \gamma^2) \\ \mathrm{P_{Det}}(\boldsymbol{s}_\ell, L) &= 1 - \Phi_{\chi^2_{1,\lambda}}(\gamma^2) \, (\Phi_{\chi^2_1}(\gamma^2))^{L-1}, \end{aligned} \tag{B.3}$$

where $\Phi_{\chi^2_{1,\lambda}}$ is the CDF of $\chi^2_{1,\lambda}$ with the noncentrality parameter $\lambda = \boldsymbol{s}_\ell^\top (\boldsymbol{s}_\ell \boldsymbol{s}_\ell^\top) \boldsymbol{s}_\ell = 1$.

Using the expression of γ^2 as a function of $\mathrm{P_{FA}}$ in (B.3) yields

$$\mathrm{P_{Det}}(\boldsymbol{s}_\ell, L) = 1 - \Phi_{\chi^2_{1,1}}\left(\Phi^{-1}_{\chi^2_1}\left((1-\mathrm{P_{FA}})^{\frac{1}{L}}\right)\right)\ (1-\mathrm{P_{FA}})^{\frac{L-1}{L}}. \tag{B.4}$$

where the dependence of $\mathrm{P_{Det}}$ w.r.t $\boldsymbol{S}$ has been dropped. Using the limits of the CDF of $\chi^2_{1,1}$ and the inverse CDF of χ^2_1, it can easily be shown that for $0 < (1-\mathrm{P_{FA}}) < 1$:

$$\begin{aligned}&\lim_{L\to+\infty} \Phi_{\chi^2_{1,1}}\left(\Phi^{-1}_{\chi^2_1}\left((1-\mathrm{P_{FA}})^{\frac{1}{L}}\right)\right) = 1, \text{ hence}\\ &\lim_{L\to+\infty} \mathrm{P_{Det}}(\boldsymbol{s}_\ell, L) = 1-(1-\mathrm{P_{FA}}) = \mathrm{P_{FA}}.\end{aligned} \tag{B.5}$$

Using the monotonicity of the two CDFs in (B.4) and of $L \mapsto (1-\mathrm{P_{FA}})^{\frac{1}{L}}$ it can be verified that for a fixed value $0 < \mathrm{P_{FA}} < 1$, $\mathrm{P_{Det}}(\boldsymbol{s}_\ell)$ is a decreasing function of L. □

B.2 Proof of Proposition 3

We first remark that problem (5.13) is invariant by rotation: if $\boldsymbol{d}^*$ is a solution, for any rotation matrix $\boldsymbol{Q}$, $\boldsymbol{Q}\boldsymbol{d}^*$ is also solution of (5.13) where all the $\boldsymbol{s}_i$ are replaced by $\boldsymbol{Q}\boldsymbol{s}_i$. This can be verified by replacing in (5.13) $\boldsymbol{d}$ and $\boldsymbol{s}_i$ by $\boldsymbol{Q}\boldsymbol{d}^*$ and $\boldsymbol{Q}\boldsymbol{s}_i$ respectively, where $\boldsymbol{Q}^\top = \boldsymbol{Q}^{-1}$. We then assume that all $\boldsymbol{s}_i$, $\{i = 1, \ldots, L\}$ are in the interior of a cone. Since the one-dimensional minimax problem is invariant by rotation, without loss of generality, $\mathcal{S}$ is restricted to be in the interior of the positive orthant $\mathbb{R}^N_+$.

Lemma 3. *If $\boldsymbol{d}^*$ is a solution of* (5.13)*, $-\boldsymbol{d}^*$ is also a solution, and $\boldsymbol{d}^*$ or $-\boldsymbol{d}^* \in \mathbb{R}^N_+$.*

Proof. The proof is by contradiction. It is obvious from (5.13) that if $\boldsymbol{d}^*$ is the solution, $-\boldsymbol{d}^*$ is also the solution. Consider a decomposition of a solution of (5.13) as $\boldsymbol{d}^* = \boldsymbol{d}^+ + \boldsymbol{d}^-$ where $\boldsymbol{d}^+$ (resp. $\boldsymbol{d}^-$) contains the positive (respectively negative) coordinates of $\boldsymbol{d}^*$. Assume that $\boldsymbol{d}^+ \neq \boldsymbol{0}$ and $\boldsymbol{d}^- \neq \boldsymbol{0}$, i.e., $\boldsymbol{d}^*$ has positive and negative components. If we define $\boldsymbol{d}^\bullet = \boldsymbol{d}^+ - \boldsymbol{d}^-$, we have

$$(\boldsymbol{d}^{\bullet\top}\boldsymbol{s}_i)^2 - (\boldsymbol{d}^{*\top}\boldsymbol{s}_i)^2 = -4(\boldsymbol{d}^{+\top}\boldsymbol{s}_i)(\boldsymbol{s}_i^\top\boldsymbol{d}^-) > 0 \tag{B.6}$$

where the first term of the scalar product is strictly positive and the second term is strictly negative. Consequently we have $\min_i (\boldsymbol{d}^{*\top}\boldsymbol{s}_i)^2 < \min_i (\boldsymbol{d}^{\bullet\top}\boldsymbol{s}_i)^2$ which contradicts that $\boldsymbol{d}^*$ is solution of (5.13).

This implies that $\boldsymbol{d}^- = 0$ or $\boldsymbol{d}^+ = 0$. In the first case, $\boldsymbol{d}^*$ is in the interior of $\mathbb{R}^N_+$. Then, $-\boldsymbol{d}^*$ is also a solution which corresponds to the second case $\boldsymbol{d}^+ = 0$. □

It is then possible to add, without loss of generality, the constraint $\boldsymbol{d} \in \mathbb{R}^N_+$ to the optimization problem (5.13). In this case $\boldsymbol{d}^\top\boldsymbol{s}_i = \cos(\theta_i) \geq 0$ where θ_i is the angle between the two unit norm vectors $\boldsymbol{d}$ and $\boldsymbol{s}_i$. This shows that this new problem is equivalent to

$$\boldsymbol{d}^* = \underset{\substack{\boldsymbol{d}:\boldsymbol{d}\in\mathbb{R}^N_+,\\ \|\boldsymbol{d}\|_2=1}}{\arg\max}\ \min_i\ \boldsymbol{d}^\top\boldsymbol{s}_i. \tag{B.7}$$

The epigraph form of this optimization problem [Boyd & Vandenberghe 2004] is:

$$\begin{array}{ll} \text{minimize} & -t \\ \text{subject to} & t - \boldsymbol{d}^\top \boldsymbol{s}_i \leq 0,\ i = \{1, \ldots, L\} \\ & -\boldsymbol{d}_i \leq 0,\ i = \{1, \ldots, L\} \\ & \|\boldsymbol{d}\|_2 = 1. \end{array} \tag{B.8}$$

Optimization problem (B.8) is non convex because of the equality constraint $\|\boldsymbol{d}\|_2 = 1$. Lemma 4 proves that the solution of (B.8) is solution of the convex QP (5.14).

Lemma 4. *The solution of* (B.8) *is the solution of the convex QP optimization problem* (5.14)

Proof. The Lagrangian of (5.14) is:

$$\mathcal{L} = -t + \sum_{i=1}^{L} \lambda_i (t - \boldsymbol{d}^\top \boldsymbol{s}_i) + \mu(\|\boldsymbol{d}\|_2 - 1) \tag{B.9}$$

The KKT[1] conditions at the optimum are the primal constraints: $\forall i,\ t \leq \boldsymbol{d}^\top \boldsymbol{s}_i$ and $\|\boldsymbol{d}\|_2 \leq 1$, the dual constraints: $\forall i\ \lambda_i \geq 0$ and $\mu \geq 0$,

- the complementary slackness: $\forall i,\ \lambda_i(t - \boldsymbol{d}^\top \boldsymbol{s}_i) = 0$ and $\mu(\|\boldsymbol{d}\|_2 - 1) = 0$,
- the gradient of the Lagrangian w.r.t. primal variables equals zero:

$$\frac{\partial \mathcal{L}}{\partial t} = 0 : \sum_i^L \lambda_i = 1; \qquad \frac{\partial \mathcal{L}}{\partial \boldsymbol{d}} = \boldsymbol{0} : \sum_i^L \lambda_i \boldsymbol{s}_i = 2\mu \boldsymbol{d}. \tag{B.10}$$

Suppose $\mu = 0$. The second gradient condition in (B.10) implies that $\sum_{i=1}^{L} \lambda_i \boldsymbol{s}_i = 0$. This implies that $\forall i,\ \lambda_i = 0$, which contradicts the sum-to-one constraint imposed by the first gradient condition in (B.10). Hence, $\mu > 0$.

The complementary slackness implies that $\|\boldsymbol{d}\|_2 = 1$. The second condition of (B.10) implies that $\boldsymbol{d} = \frac{1}{2\mu} \sum_i^L \lambda_i \boldsymbol{s}_i$. Since $\mu > 0$, $\forall i,\ \lambda_i \geq 0$ and $\forall i,\ \boldsymbol{s}_i \in \mathbb{R}_+^N$, this implies that $\boldsymbol{d} \in \mathbb{R}_+^N$. □

[1] KKT: Karush-Kuhn-Tucker.

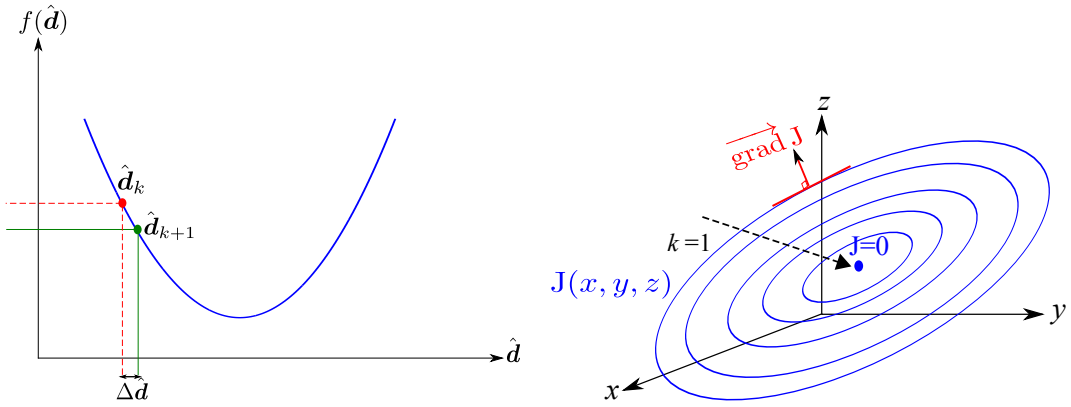
B.3 Gradient descent for 1D minimax problem

Based on the gradient descent approach, we seek the minimax atom for $N = 3$. We denote $\widehat{\boldsymbol{d}}$ the minimax atom by gradient descent. The objective of this study is to resolve the minimax optimization problem for $K = 1$, then contemplate to generalize it to the case $K > 1$.

Gradient descent takes the form

$$\widehat{\boldsymbol{d}}_{k+1} = \widehat{\boldsymbol{d}}_k - \Delta\widehat{\boldsymbol{d}}\,\overrightarrow{\text{grad}}\,\text{J}(\widehat{\boldsymbol{d}}_k), \tag{B.11}$$

where $\Delta\widehat{\boldsymbol{d}}$ is the step size between two iterations and $\text{J}(\widehat{\boldsymbol{d}}_k)$ is the objective function to be minimized in the direction of the negative gradient as shown in Figure B.1 when evaluating in term of distance.



(a) View of the step size between two iterations. (b) Minimizing the function J by iterations k.

Figure B.1: The concept of general gradient descent w.r.t. (B.11). (a) Step size between two iterations $\Delta\widehat{\boldsymbol{d}}$. (b) Function J is minimized in the direction of the negative gradient when evaluating in term of distance. Between two iterations: $\text{J}(\widehat{\boldsymbol{d}}_{k+1}) < \text{J}(\widehat{\boldsymbol{d}}_k)$.

By iterations k, we seek to minimize the maximum distance between a point $\widehat{\boldsymbol{d}}_k$ in the space to to the farthest alternative $\boldsymbol{s}_{i^*(k)}$ without increasing the distance of $\widehat{\boldsymbol{d}}_k$ to the other alternatives $\boldsymbol{s}_i(k)$. However, in term of correlation (which is considered in our setting), we want to maximize the minimum correlation between $\boldsymbol{S}$ and $\widehat{\boldsymbol{d}}$, that is the problem (5.13). In this regard, we define J as

$$\text{J}(\widehat{\boldsymbol{d}}) = \min_{i=1,\ldots,L} |\widehat{\boldsymbol{d}}^\top \boldsymbol{s}_i|. \tag{B.12}$$

Thus, by gradient descent, we want to "maximize" J at each iteration k: $\text{J}(\widehat{\boldsymbol{d}}_{k+1}) > \text{J}(\widehat{\boldsymbol{d}}_k)$. The optimization problem becomes

$$\widehat{\boldsymbol{d}}_{k+1} = \widehat{\boldsymbol{d}}_k + \Delta\widehat{\boldsymbol{d}}\,\overrightarrow{\text{grad}}\,\text{J}(\widehat{\boldsymbol{d}}_k). \tag{B.13}$$

Note that if $\widehat{\boldsymbol{d}}^\top \boldsymbol{s}_i = 0$, the function $\text{J}(\widehat{\boldsymbol{d}})$ is non-differentiable and this case has to be

managed apart.

In spherical coordinates, by defining the norm of $\widehat{\boldsymbol{d}}$ as r (where $r = 1$), the elevation angle $\phi \in [\frac{\pi}{2}, -\frac{\pi}{2}]$ and the azimuth angle $\theta \in [0, 2\pi]$, the total derivative of J writes

$$\mathrm{dJ} = \frac{\partial \mathrm{J}}{\partial r}\,\mathrm{d}r + \frac{\partial \mathrm{J}}{\partial \phi}\,\mathrm{d}\phi + \frac{\partial \mathrm{J}}{\partial \theta}\,\mathrm{d}\theta$$

$$\mathrm{dJ} = \overrightarrow{\mathrm{grad}}\,\mathrm{J} \cdot \overrightarrow{\mathrm{d}l} \tag{B.14}$$

$$\mathrm{dJ} = \overrightarrow{\mathrm{grad}}\,\mathrm{J} \cdot \begin{vmatrix} \mathrm{d}r = 0 \\ r\ \mathrm{d}\phi \\ r \sin\phi\ \mathrm{d}\theta \end{vmatrix} .$$

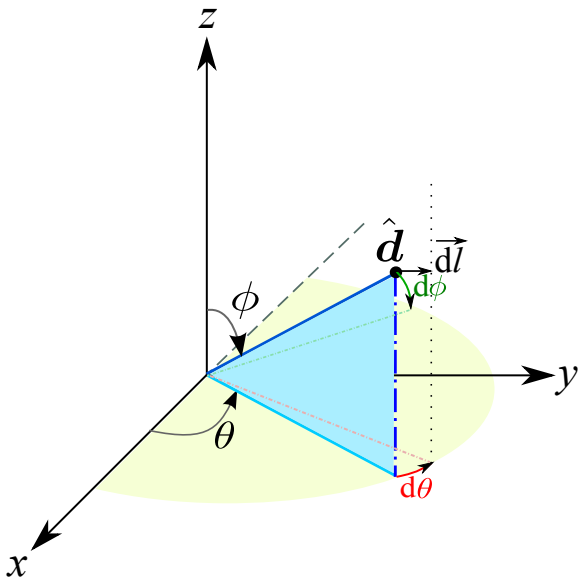
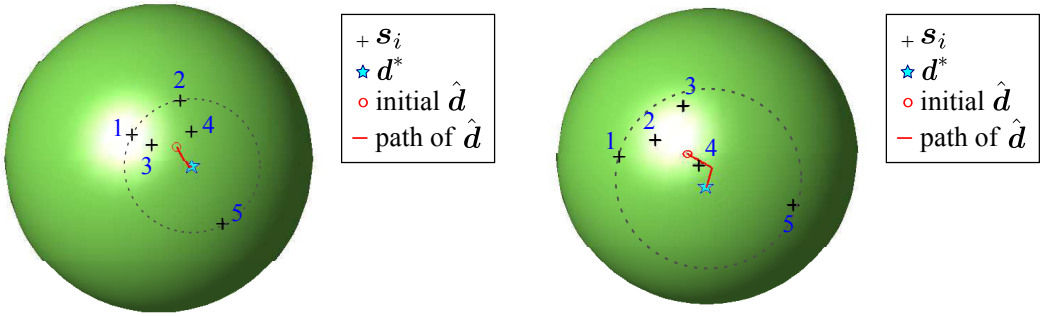


Figure B.2: Illustration of the elements in dJ.

Following the problem (B.13) (and thus (B.14)), when the point $\widehat{\boldsymbol{d}}_k$ arrives at an intersection of two planes, that is the distance or correlation of $\widehat{\boldsymbol{d}}_k$ and the two farthest alternatives in $\boldsymbol{S}$ (say, $\boldsymbol{s}_a$ and $\boldsymbol{s}_b$) is the same: $\mathrm{J}(\widehat{\boldsymbol{d}}_k) = |\widehat{\boldsymbol{d}}_k^{\top}\boldsymbol{s}_a| = |\widehat{\boldsymbol{d}}_k^{\top}\boldsymbol{s}_b|$, then the function J is non-differentiable. When this happen, we have to projects $\widehat{\boldsymbol{d}}_k$ to a new plane of mean($\boldsymbol{s}_a$,$\boldsymbol{s}_b$). To do so, we denote first $\boldsymbol{A} = [\widehat{\boldsymbol{d}}_k, \mathrm{mean}(\boldsymbol{s}_a, \boldsymbol{s}_b)]$. Then, we seek a new d$l$, defined as $\mathrm{d}l' = \boldsymbol{A}\boldsymbol{\beta}^*$, where

$$\boldsymbol{\beta}^* = \|\ \mathrm{d}l - \boldsymbol{A}\boldsymbol{\beta}\|_2^2 \ \Rightarrow \boldsymbol{\beta}^* = (\boldsymbol{A}^{\top}\boldsymbol{A})^{-1}\boldsymbol{A}^{\top}\ \mathrm{d}l. \tag{B.15}$$

By using $\mathrm{d}l' = \boldsymbol{A}\boldsymbol{\beta}^*$, we can then find $\widehat{\boldsymbol{d}}_{k+1}$.



(a) Example 1 (b) Example 2

Figure B.3: Two examples of gradient descent simulations to find minimax atom $\widehat{\boldsymbol{d}}$. The initialization point (red circle) is the mean of $\mathcal{S}$ (black crosses represent $\boldsymbol{s}_i \in \boldsymbol{S}$, where here, $\boldsymbol{S} \in \mathbb{R}^{3\times 5}$ is random normalized alternatives). The red line indicates the path of the gradient descent method to find $\widehat{\boldsymbol{d}}$. The minimax atom $\boldsymbol{d}^*$ (cyan star) is generated from a QP solver, based on (5.13). These simulations show that, $\widehat{\boldsymbol{d}} \approx \boldsymbol{d}^*$ (i.e., the exact solution) for $N = 3$. The gray dashes line indicates the smallest circle enclosing $\boldsymbol{S}$.

In Figure B.3, we show two examples of minimax gradient descent approach using different set of $\boldsymbol{S}$. For these numerical demonstrations, the initial point $\widehat{\boldsymbol{d}}_0$ is the normalized mean of the set $\boldsymbol{S}$ ($\widehat{\boldsymbol{d}}_0$ is represented by red circles in Figure B.3). The results obtained are compared to the minimax atom $\boldsymbol{d}^*$ (5.14), which is in the form of QP.

In Example 1 (Figure B.3(a)), at $k = 0$, $\widehat{\boldsymbol{d}}_0$ sees that $\boldsymbol{s}_5$ is the farthest from it. So at each iteration k, $\widehat{\boldsymbol{d}}_k$ will find the direction that yields $\mathrm{J}(\widehat{\boldsymbol{d}}_k) > \mathrm{J}(\widehat{\boldsymbol{d}}_{k-1})$, i.e., in this example, the direction that minimizes the distance between $\widehat{\boldsymbol{d}}_k$ and $\boldsymbol{s}_5$, while assuring that its distance to the other alternatives $(\boldsymbol{s}1, \ldots, \boldsymbol{s}_4)$ does not increase. At the final point, when there is no more significant improvement of the value J, $(\mathrm{J}(\widehat{\boldsymbol{d}}_k) - \mathrm{J}(\widehat{\boldsymbol{d}}_{k-1}) \approx 0)$, the minimax $\widehat{\boldsymbol{d}}_k$ is found and it conforms to $\boldsymbol{d}^*$. Here, three alternatives $(\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{s}_5)$ are at the minimax distance from $\widehat{\boldsymbol{d}}_k$.

Using a different set of $\boldsymbol{S}$, Example 2 (Figure B.3(b)) shows that at initialization, $\widehat{\boldsymbol{d}}_0$ sees that $\boldsymbol{s}_5$ is the farthest from it. Following towards this direction, $\widehat{\boldsymbol{d}}_k$ arrives at a point where its distance is equivalent to $\boldsymbol{s}_1$ and $\boldsymbol{s}_5$. $\widehat{\boldsymbol{d}}_k$ changes into the plane of mean$(\boldsymbol{s}_1, \boldsymbol{s}_5)$. Then, $\widehat{\boldsymbol{d}}_k$ continues to find direction that minimizes its distance w.r.t. the farthest alternative(s). In the end, we found $\widehat{\boldsymbol{d}} \approx \boldsymbol{d}^*$, where there are two alternatives $(\boldsymbol{s}_1, \boldsymbol{s}_5)$ lie on the border of the smallest circle of center $\widehat{\boldsymbol{d}}_k$.

We investigated this gradient descent approach for $K = 1$, hoping to generalize it to the case $K > 1$ and to obtain exact minimax solution for such cases. This generalization is however not achieved yet because it poses several important issues (complexity of coordinate system in $N > 3$-dimensional Euclidean space and the case where the function J (B.12) is non-differentiable).

Bibliography

[Abolghasemi *et al.* 2012] V. Abolghasemi, S. Ferdowsi and S. Sanei. *Blind Separation of Image Sources via Adaptive Dictionary Learning.* IEEE Trans. on Image Process., vol. 21, no. 6, pages 2921–2930, 2012. (Cited on page 43.)

[Aharon *et al.* 2006] 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), pages 4311–4322, November 2006. (Cited on pages 39, 43, and 83.)

[Ardekani *et al.* 1999] Babak A. Ardekani, Jeff Kershaw, Kenichi Kashikura and Iwao Kanno. *Activation detection in functional MRI using subspace modeling and maximum likelihood estimation.* IEEE Transactions on Medical Imaging, vol. 18, no. 2, pages 101–114, 1999. (Cited on page 35.)

[Arias-Castro *et al.* 2005] E. Arias-Castro, D.L. Donoho and X. Huo. *Near-optimal detection of geometric objects by fast multiscale methods.* IEEE Transactions on Information Theory, vol. 51, no. 7, pages 2402–2425, July 2005. (Cited on page 53.)

[Arias-Castro *et al.* 2010] 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), pages 2533–2556, 2010. (Cited on pages 24 and 53.)

[Arias-Castro *et al.* 2011] E. Arias-Castro, E.J. Candès and A. Durand. *Detection of an anomalous cluster in a network.* The Annals of Statistics, vol. 39, no. 1, pages 278–304, February 2011. (Cited on page 53.)

[Banerjee *et al.* 2006] A. Banerjee, P. Burlina and C. Diehl. *A support vector method for anomaly detection in hyperspectral imagery.* IEEE Trans. on Geoscience and Remote Sensing, vol. 44, no. 8, pages 2282–2291, 2006. (Cited on page 35.)

[Bellman 1957] Richard Bellman. Dynamic programming. Princeton Univ. Press, 1957. (Cited on page 60.)

[Berk *et al.* 2005] Alexander Berk, Gail P. Anderson, Prabhat K. Acharya, Lawrence S. Bernstein, Leonid Muratov, Jamine Lee, Marsha J. Fox, Steven M. Adler-Golden, James H. Chetwynd, Michael L. Hoke, Ronald B. Lockwood, Thomas W. Cooley and James A Gardner. *MODTRAN5: a reformulated atmospheric band model with auxiliary species and practical multiple scattering options.* In Proc. SPIE 5806, Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XI, pages 88–95. SPIE, 2005. (Cited on page 51.)

[Blumensath *et al.* 2007] T. Blumensath, M. Yaghoobi and M.E. Davies. *Iterative Hard Thresholding and ℓ_0 Regularisation.* In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), volume 3, pages III–877–III–880, April 2007. (Cited on page 42.)

[Boser *et al.* 1992] Bernhard E. Boser, Isabelle M. Guyon and Vladimir N. Vapnik. *A training algorithm for optimal margin classifiers.* In Proceedings of the fifth annual workshop on Computational learning theory, pages 144–152. ACM, 1992. (Cited on pages 23 and 66.)

[Boyd & Vandenberghe 2004] S. Boyd and L. Vandenberghe. Convex optimization. Cambridge University Press, 2004. (Cited on page 135.)

[Caillier *et al.* 2012] P. Caillier, M. Accardo, L. Adjali, H. Anwand, R. Bacon, D. Boudon, L. Brotons, L. Capoani, E. Daguisé, M. Dupieux, C. Dupuy, M. François, A. Glindemann, D. Gojak, G. Hansali, T. Hahn, A. Jarno, A. Kelz, C. Koehler, J. Kosmalski, F. Laurent, M. Le Floch, J.-L. Lizon, M. Loupias, A. Manescau, J. E. Migniau, C. Monstein, H. Nicklas, L. Parès, A. Pécontal-Rousset, L. Piqueras, R. Reiss, A. Remillieux, E. Renault, G. Rupprecht, O. Streicher, R. Stuik, H. Valentin, J. Vernet, P. Weilbacher and G. Zins. *The MUSE project face to face with reality.* In Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, volume 8449 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, September 2012. (Cited on pages ix, 28, 29, 31, and 94.)

[Chen *et al.* 1998] Scott Shaobing Chen, David L. Donoho and Michael A. Saunders. *Atomic decomposition by basis pursuit.* SIAM journal on scientific computing, vol. 20, no. 1, pages 33–61, 1998. (Cited on page 41.)

[Cortes & Vapnik 1995] C. Cortes and V.N. Vapnik. *Support-Vector Networks.* Machine Learning, vol. 20, pages 273–297, 1995. (Cited on page 66.)

[Dai & Milenkovic 2009] Wei Dai and Olgica Milenkovic. *Subspace pursuit for compressive sensing signal reconstruction.* IEEE Transactions on Information Theory, vol. 55, no. 5, pages 2230–2249, 2009. (Cited on page 42.)

[DeCarlo 1997] Lawrence T. DeCarlo. *On the meaning and use of kurtosis.* Psychological methods, vol. 2, no. 3, page 292, 1997. (Cited on page 35.)

[Dhillon & Modha 2001] I.S. Dhillon and D.S. Modha. *Concept decompositions for large sparse text data using clustering.* Machine learning, vol. 42, no. 1-2, pages 143–175, 2001. (Cited on pages 79 and 85.)

[Donoho & Jin 2004] D.L. Donoho and J. Jin. *Higher Criticism for Detecting Sparse Heterogeneous Mixtures.* Annals of Statistics, vol. 32, no. 3, pages 962–994, 2004. (Cited on pages 24, 53, and 55.)

[Donoho & Johnstone 1998] D.L. Donoho and I.M. Johnstone. *Minimax estimation via wavelet shrinkage.* Annals of Statistics, 1998. (Cited on page 23.)

[Donoho 2004a] David L. Donoho. *Compressed Sensing.* pages 1–34, 2004. (Cited on page 39.)

[Donoho 2004b] David L. Donoho. *For most large underdetermined systems of linear equations the minimal 1-norm solution is also the sparsest solution.* Communications on Pure and Applied Mathematics, vol. 59, no. 6, pages 797–829, 2004. (Cited on pages 42 and 43.)

[Eckart & Young 1936] Carl Eckart and Gale Young. *The approximation of one matrix by another of lower rank.* Psychometrika, vol. 1, no. 3, pages 211–218, 1936. (Cited on page 40.)

[Eismann *et al.* 2009] M.T. Eismann, A.D. Stocker and N.M. Nasrabadi. *Automated Hyperspectral Cueing for Civilian Search and Rescue.* Proceedings of the IEEE, vol. 97, no. 6, pages 1031–1055, 2009. (Cited on page 31.)

[Elad & Aharon 2006] M. Elad and M. Aharon. *Image Denoising Via Sparse and Redundant Representations Over Learned Dictionaries.* IEEE Trans. on Image Process., vol. 15, no. 12, pages 3736–3745, 2006. (Cited on page 43.)

[Elad *et al.* 2007] M. Elad, P. Milanfar and R. Rubinstein. *Analysis versus synthesis in signal priors.* Inverse problems, vol. 23, no. 3, page 947, 2007. (Cited on pages 39 and 127.)

[Emmanuel *et al.* 2004] Candès Emmanuel, Justin Romberg and Terence Tao. *Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information.* 2004. (Cited on page 39.)

[Engan *et al.* 1999] 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), volume 5, pages 2443–2446. IEEE, 1999. (Cited on pages 39 and 44.)

[Fernandes 2010] Leandro Augusto Frata Fernandes. *On the generalization of subspace detection in unordered multidimensional data.* PhD thesis, Federal University of Rio Grande do Sul (UFRGS), 2010. (Cited on page 35.)

[Fisher 1925] Ronald A. Fisher. Statistical methods for research workers. Oliver and Boyd, Edinburgh, 1925. (Cited on page 11.)

[Gersho & Gray 1991] A. Gersho and R.M. Gray. Vector quantization and signal compression. Kluwer Academic Publishers, Norwell, MA, USA, 1991. (Cited on pages 79 and 83.)

[Govender *et al.* 2007] M. Govender, K. Chetty and H. Bulcock. *A review of hyperspectral remote sensing and its application in vegetation and water resource studies.* Water Sa, vol. 33, no. 2, 2007. (Cited on page 27.)

[Gowda & Krishna 1978] K. C. Gowda and G. Krishna. *Agglomerative clustering using the concept of mutual nearest neighbourhood.* Pattern Recognition, vol. 10, no. 2, pages 105–112, 1978. (Cited on page 39.)

[Grant & Boyd 2014] Michael Grant and Stephen Boyd. *CVX: Matlab software for disciplined convex programming.* http://cvxr.com/cvx, March 2014. (Cited on page 65.)

[Green *et al.* 1998] R.O. Green, M.L. Eastwood, C.M. Sarture, T.G. Chrien, M. Aronsson, B.J. Chippendale, J.A. Faust, B.E. Pavri, C.J. Chovit, M. Solis, M.R. Olah and O. Williams. *Imaging Spectroscopy and the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS).* Remote Sensing of Environment, vol. 65, no. 3, pages 227–248, 1998. (Cited on page 28.)

[Guyon & Elisseeff 2003] I. Guyon and A. Elisseeff. *An Introduction to Variable and Feature Selection.* Journal of Machine Learning Research, vol. 3, pages 1157–1182, 2003. (Cited on page 87.)

[Hinnrichs *et al.* 2004] M. Hinnrichs, J.O. Jensen and G. McAnally. *Handheld hyperspectral imager for standoff detection of chemical and biological aerosols.* In Proc. SPIE 5416, Chemical and Biological Sensing V, 2004. (Cited on page 28.)

[Huang *et al.* 2014] H. Huang, L. Liu and M.O. Ngadi. *Recent Developments in Hyperspectral Imaging for Assessment of Food Quality and Safety.* Sensors, vol. 14, no. 4, pages 7248–7276, 2014. (Cited on page 28.)

[Ingster & Suslina 2003] Y. Ingster and I. Suslina. *Minimax Detection of Positive Signals.* Journal of Mathematical Sciences, vol. 118, no. 6, pages 5570–5585, 2003. (Cited on page 24.)

[Jiao *et al.* 2012] J. Jiao, Lin Zhang and R.D. Nowak. *Minimax-Optimal Bounds for Detectors Based on Estimated Prior Probabilities.* IEEE Transactions on Information Theory, vol. 58, no. 9, pages 6101–6109, Sept 2012. (Cited on page 25.)

[Kassam & Poor 1985] S.A. Kassam and H.V. Poor. *Robust techniques for signal processing: A survey.* In Proc. of IEEE, volume 73, pages 433–481. IEEE, 1985. (Cited on page 23.)

[Khan & Madden 2010] S.S. Khan and M.G. Madden. *A Survey of Recent Trends in One Class Classification.* In Artificial Intelligence and Cognitive Science, volume 6206, pages 188–197. Springer Berlin Heidelberg, 2010. (Cited on page 66.)

[Khatri 1968] C.G. Khatri. *Further contributions to some inequalities for normal distributions and their applications to simultaneous confidence bounds.* Annals of the Institute of Statistical Mathematics, vol. 22, no. 1, pages 451–458, 1968. (Cited on page 74.)

[Kokaly *et al.* 1998] R.F. Kokaly, R.N. Clark and K.E. Livo. *Mapping the Biology and Mineralogy of Yellowstone National Park using Imaging Spectroscopy.* In AVIRIS Workshop Proceedings (JPL publication 97-21), volume 1, pages 245–254, 1998. (Cited on page 31.)

[Kong & Wang 2012] S. Kong and D. Wang. *A Dictionary Learning Approach for Classification: Separating the Particularity and the Commonality.* In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato and Cordelia Schmid, editeurs, Computer Vision (ECCV), volume 7572 of *Lecture Notes in Computer Science*, pages 186–199. Springer Berlin Heidelberg, 2012. (Cited on page 43.)

[Kraut *et al.* 2001] S. Kraut, L.L. Scharf and L.T. McWhorter. *Adaptive Subspace Detectors.* IEEE Trans. Signal Process., vol. 49, no. 1, pages 1–16, January 2001. (Cited on pages 31 and 34.)

[Kwon & Nasrabadi 2005] H. Kwon and N.M. Nasrabadi. *Kernel RX-algorithm: a nonlinear anomaly detector for hyperspectral imagery.* IEEE Trans. on Geoscience and Remote Sensing, vol. 43, no. 2, pages 388–397, 2005. (Cited on page 35.)

[LeCun *et al.* 1998] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner. *Gradient-based learning applied to document recognition.* Proceedings of the IEEE, vol. 86, no. 11, pages 2278–2324, 1998. (Cited on page 113.)

[Lee & Cho 2006] S.H. Lee and H.K. Cho. *Detection of the pine trees damaged by pine wilt disease using high spatial remote sensing data.* In Proc. of the ISPRS Commision VII Symposium Remote Sensing: from pixels to processes, volume XXXVI, part 7, pages 8–11, 2006. (Cited on page 31.)

[Lehmann & Romano 2005] E.L. Lehmann and J.P. Romano. Testing statistical hypotheses. Springer, third edition, 2005. (Cited on pages 16, 23, and 24.)

[Lyold 1982] S.P. Lyold. *Least Squares Quantization in PCM.* IEEE Trans. on Information Theory, vol. 28, no. 2, pages 129–137, March 1982. (Cited on page 38.)

[MacQueen 1967] J.B. MacQueen. *Some Methods for Classification and Analysis of MultiVariate Observations.* In Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability, volume 1, pages 281–297. University of California Press, 1967. (Cited on page 38.)

[Mahalanobis 1936] P.C. Mahalanobis. *On the generalised distance in statistics.* In Proceedings of the National Institute of Sciences of India, volume 2, pages 49–55, 1936. (Cited on page 35.)

[Mallat & Zhang 1993] S.G. Mallat and Z. Zhang. *Matching Pursuits with Time-Frequency Dictionaries.* IEEE Trans. Signal Process., pages 3397–3415, December 1993. (Cited on pages 39 and 42.)

[Mallat *et al.* 1994] S. Mallat, G. Davis and Z. Zhang. *Adaptive time-frequency decompositions.* SPIE Journal of Optical Engineering, vol. 33, pages 2183 – 2191, 1994. (Cited on pages 39 and 42.)

[Mallat 2008] Stephane Mallat. A wavelet tour of signal processing: the sparse way. Academic Press, 2008. (Cited on page 43.)

[Manly *et al.* 2004] Kenneth F. Manly, Dan Nettleton and J.T. Gene Hwang. *Genomics, prior probability, and statistical tests of multiple hypotheses.* Genome Research, vol. 14, no. 6, pages 997–1001, 2004. (Cited on page 12.)

[Manolakis *et al.* 2009] D. Manolakis, R. Lockwood, T. Cooley and J. Jacobson. *Is there a best hyperspectral detection algorithm?* In Proc. SPIE 7334, Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XV, pages 733402.1– 733402.16, 2009. (Cited on pages 28 and 31.)

[Manolakis *et al.* 2014] D. Manolakis, E. Truslow, M. Pieper, T. Cooley and M. Brueggeman. *Detection Algorithms in Hyperspectral Imaging Systems: An Overview of Practical Algorithms.* Signal Processing Magazine, IEEE, vol. 31, no. 1, pages 24–33, 2014. (Cited on page 31.)

[Mary & Ferrari 2014] D. Mary and A. Ferrari. *An alternate standardization of binomial counts in Higher Criticism.* In International Symposium on Information Theory. IEEE, 2014. (Cited on page 24.)

[Mary *et al.* 2014] David Mary, André Ferrari and Raja Fazliza Raja Suleiman. *Détection de sources astrophysiques dans les données du spectographe intégral de champ MUSE.* Actes du 3e colloque scientifique SFPT-GH sur l'Imagerie Hyperspectrale, May 2014. (Cited on pages 5 and 93.)

[Matteoli *et al.* 2007] S. Matteoli, F. Carnesecchi, M. Diani, G. Corsini and L. Chiarantini. *Comparative analysis of hyperspectral anomaly detection strategies on a new high spatial and spectral resolution data set.* In Proc. SPIE 6748, Image and Signal Processing for Remote Sensing XIII, pages 67480E.1–67480E.11, 2007. (Cited on page 35.)

[Nam *et al.* 2011] Sangnam Nam, M.E. Davies, M. Elad and R. Gribonval. *Recovery of cosparse signals with Greedy Analysis Pursuit in the presence of noise.* In Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2011 4th IEEE International Workshop on, pages 361–364, Dec 2011. (Cited on page 128.)

[Nasrabadi 2014] N.M. Nasrabadi. *Hyperspectral Target Detection : An Overview of Current and Future Challenges.* Signal Processing Magazine, IEEE, vol. 31, no. 1, pages 34–44, 2014. (Cited on page 31.)

[Needell & Tropp 2009] Deanna Needell and Joel A. Tropp. *CoSaMP: Iterative signal recovery from incomplete and inaccurate samples.* Applied and Computational Harmonic Analysis, vol. 26, no. 3, pages 301–321, 2009. (Cited on page 42.)

[Nelsen 2006] R.B. Nelsen. An introduction to copulas. Springer, 2006. (Cited on page 72.)

[Neumann 1928] J. Von Neumann. *Zur Theorie der Gesellschaftsspiele: On the theory of parlor games.* Mathematische Annalenn, vol. 100, pages 295–300, 1928. (Cited on page 22.)

[Neyman & Pearson 1933] J. Neyman and E.S. Pearson. *On the Problem of the Most Efficient Tests of Statistical Hypotheses.* Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 231, pages 289–337, 1933. (Cited on pages 11 and 15.)

[Nutall 1974] A.H. Nutall. *Some Integrals Involving the (Q sub M) - Function.* Rapport technique AD-779846, Naval Underwater Systems Center New London, Connecticut, May 1974. (Cited on page 64.)

[Olshausen & Field 1996] B.A. Olshausen and D.J. Field. *Natural image statistics and efficient coding.* Network Computation in Neural Systems, vol. 7, no. 2, pages 333–339, 1996. (Cited on page 44.)

[Paris *et al.* 2013a] Silvia Paris, David Mary and André Ferrari. *Detection tests using sparse models, with application to hyperspectral data.* IEEE Trans. Signal Process., vol. 61, no. 6, pages 1481–1494, March 2013. (Cited on pages 28, 31, and 95.)

[Paris *et al.* 2013b] Silvia Paris, Raja Fazliza R. Suleiman, David Mary and André Ferrari. *Constrained Likelihood Ratios for detecting sparse signals in highly noisy 3D data.* In International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 3947–3951. IEEE, 2013. (Cited on pages 5, 60, 61, 94, and 97.)

[Pati *et al.* 1993] Yagyensh Chandra Pati, Ramin Rezaiifar and P.S. Krishnaprasad. *Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition.* In Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on, pages 40–44. IEEE, 1993. (Cited on pages 39 and 42.)

[Pearlman *et al.* 2011] J. Pearlman, S. Carman, C. Segal, P. Jarecke, P. Clancy and W. Browne. *Overview of the Hyperion Imaging Spectrometer for the NASA EO-1 mission.* In IEEE International Geoscience and Remote Sensing Symposium (IGARSS), volume 7, pages 3036–3038, 2011. (Cited on page 28.)

[Pearson 1901] Karl Pearson. *On lines and planes of closest fit to systems of points in space.* Philosophical Magazine, vol. 2, pages 559–572, 1901. (Cited on page 38.)

[Pearson 1905] Karl Pearson. *"Das Fehlergesetz und Seine Verallgemeinerungen Durch Fechner und Pearson." A Rejoinder.* Biometrika, vol. 4, no. 1/2, pages 169–212, 1905. (Cited on page 35.)

[Reed & Yu 1990] I.S. Reed and X. Yu. *Adaptive multiple-band CFAR detection of an optical pattern with unknown spectral distribution.* International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 38, pages 1760–1770, October 1990. (Cited on page 34.)

[Robey *et al.* 1992] F.C. Robey, D.R. Fuhrmann, E.J. Kelly and R. Nitzberg. *A CFAR adaptive matched filter detector.* IEEE Trans. on Aerospace and Electronic Systems, vol. 28, no. 1, pages 208–216, 1992. (Cited on page 31.)

[Roweis & Saul 2000] Sam T. Roweis and Lawrence K. Saul. *Nonlinear dimensionality reduction by locally linear embedding.* Science, vol. 290, no. 5500, pages 2323–2326, 2000. (Cited on page 40.)

[Rubinstein *et al.* 2012] R. Rubinstein, T. Faktor and M. Elad. *K-SVD dictionary-learning for the analysis sparse model.* In Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, pages 5405–5408, March 2012. (Cited on page 128.)

[Samaria & Harter 1994] 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, pages 138–142, 1994. (Cited on page 111.)

[Sammon 1969] John W. Sammon. *A nonlinear mapping for data structure analysis.* IEEE Transactions on computers, vol. 18, no. 5, pages 401–409, 1969. (Cited on page 40.)

[Scharf & Friedlander 1994] L.L. Scharf and B. Friedlander. *Matched Subspace Detectors.* IEEE Trans. Signal Process., vol. 42, no. 8, pages 2146–2457, 1994. (Cited on pages 16, 31, 33, and 121.)

[Schölkopf 1997] Bernhard Schölkopf. *Support vector learning.* PhD thesis, Technishen Universitat Berlin, 1997. (Cited on page 66.)

[Sen & Srivastava 1990] A. Sen and M. Srivastava. Regression analysis: Theory, methods and applications. Springer-Verlag New York Inc., 1990. (Cited on page 133.)

[Sion 1958] Maurice Sion. *On general minimax theorems.* Pacific Journal of Mathematics, vol. 8, no. 1, pages 171–176, 1958. (Cited on page 23.)

[Sklar 2011] B. Sklar. Digital communications: Fundamentals & applications. Prentice Hall, 2011. (Cited on page 51.)

[Student 1908] Student. *The probable error of a mean.* Biometrika, pages 1–25, 1908. (Cited on page 11.)

[Suleiman *et al.* 2013a] Raja Fazliza Raja Suleiman, David Mary and André Ferrari. *Minimax sparse detection based on one-class classifiers.* In International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5553–5557. IEEE, 2013. (Cited on pages 4, 51, and 61.)

[Suleiman *et al.* 2013b] Raja Fazliza Raja Suleiman, David Mary and André Ferrari. *Parcimonie, apprentissage et classification pour une approche minimax en détection.* In Actes du 24e Colloque GRETSI sur le Traitement du Signal et Images, 2013. (Cited on pages 5, 61, 79, and 93.)

[Suleiman *et al.* 2014a] Raja Fazliza Raja Suleiman, David Mary and André Ferrari. *Dimension reduction for hypothesis testing in worst-case scenarios.* IEEE Trans. Signal

Process., vol. 62, no. 22, pages 5973–5968, November 2014. (Cited on pages 4, 51, 61, 79, 93, and 111.)

[Suleiman *et al.* 2014b] Raja Fazliza Raja Suleiman, David Mary and André Ferrari. *Subspace learning in minimax detection.* In International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 3062–3066. IEEE, 2014. (Cited on pages 4, 61, 79, and 111.)

[Tax & Duin 2004] D.M.J. Tax and R.P.W. Duin. *Support Vector Data Description.* Machine Learning, vol. 54, pages 45–66, 2004. (Cited on pages 35 and 66.)

[Tibshirani 1994] Robert Tibshirani. *Regression Shrinkage and Selection Via the Lasso.* Journal of the Royal Statistical Society, Series B, vol. 58, pages 267–288, 1994. (Cited on pages 39 and 43.)

[Verhamme *et al.* 2012] A. Verhamme, Y. Dubois, J. Blaizot, T. Garel, R. Bacon, J. Devriendt, B. Guiderdoni and A. Slyz. *Lyman-α emission properties of simulated galaxies: interstellar medium structure and inclination effects.* Astronomy & Astrophysics, vol. 546, page A111, 2012. (Cited on pages 54 and 94.)

[Wald 1945] Abraham Wald. *Statistical decision functions which minimize the maximum risk.* Annals of Mathematics, pages 265–280, 1945. (Cited on page 23.)

[Wald 1950] Abraham Wald. Statistical decision functions. Wiley, 1950. (Cited on page 23.)

[Zhang & Li 2010] Q. Zhang and B. Li. *Discriminative K-SVD for dictionary learning in face recognition.* In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2691–2698, 2010. (Cited on page 43.)