








**UNIVERSITÉ DE NICE - SOPHIA ANTIPOLIS - UFR Sciences**

École Doctorale en Sciences Fondamentales et Appliquées

**SAPIENZA UNIVERSITÀ DI ROMA**

Dipartimento di Ingegneria Informatica Automatica e Gestionale
Ingegneria dei Sistemi

PhD THESIS
to obtain the title of

**Docteur en Sciences**
de l'Université de Nice-Sophia Antipolis
Spécialité: Sciences de l'Univers
and
**Dottore in Ingegneria dei Sistemi**
Sapienza Università di Roma

Defended by

**Silvia Paris**

# Sparsity-based detection strategies for faint signals in noise. Application to astrophysical hyperspectral data.

**Defended on October 4, 2013**

**Jury :**

| | | | | |
|---|---|---|---|---|
| Reviewers : | Fulvio Gini | Professor | - | Università di Pisa |
| | Eric Moreau | Professor | - | Université du Sud - Toulon - Var |
| Examinators : | Florent Chatelain | MdC | - | GIPSA-LAB, INPG |
| | Antonio Napolitano | Professor | - | Università di Napoli Parthenope |
| | Eric Slézak | Astronomer | - | Observatoire de la Côte d'Azur |
| Advisors : | André Ferrari | Professor | - | UNS |
| | David Mary | MdC | - | UNS |
| | Salvatore Monaco | Professor | - | Sapienza - Università di Roma |

**Sparsity-based detection strategies for faint signals in noise. Application to astrophysical hyperspectral data.**

**Abstract:** This thesis deals with the problem of detecting unknown signals at low Signal-to-Noise Ratio. This work focuses on the definition, study and implementation of efficient methods able to discern only-noise observations from those that presumably carry the information of interest in a sparse way. The relevance of these methods is assessed on hyperspectral data as an applicative part.

In the first part of this work, the basic principles of statistical hypothesis testing together with a general overview on sparse representations, estimation and detection are introduced.

The second part of the manuscript is divided into two sub-parts. In the first sub-part, two statistical hypotheses tests are proposed and studied. Both are adapted to the detection of sparse signals thanks to sparse estimates that can be interpreted in the Maximum A Posteriori or in the Penalized Least Squares frameworks. The first approach is the Posterior Density Ratio test, which computes the ratio of the a posteriori distribution under each hypothesis of the data model. The second is a Likelihood Ratio test in which the Maximum A Posteriori estimate replaces the Maximum Likelihood estimate. The behaviors and the relative differences between these constrained likelihood ratios are theoretically investigated through a detailed study of their analytical and structural characteristics. The tests' detection performances are compared with those of classical frequentist and Bayesian methods such as the unconstrained Generalized Likelihood Ratio test and the Bayes Factor.

According to the three-dimensional data sets considered in the applicative part, and to be closer to realistic scenarios involving data acquisition systems, the second sub-part presents detection methods seeking: i) to account for spectrally variable noise; ii) to exploit the spectral similarities of neighbors pixels in the spatial domain and iii) to exploit the greater accuracy brought by dictionary-based models, which take into account the spatio-spectral blur of information caused by instrumental Point Spread Functions.

The tests are finally applied to massive astrophysical hyperspectral data in the context of the European Southern Observatory's Multi Unit Spectroscopic Explorer (whose first light is foreseen for end 2013). The proposed methods have shown to be efficient for the detection of general sources with this instrument. A particular attention is finally paid to the detection of very faint and sparse sources known as Ly$-\alpha$ emitters, characterized by a spectrum that contains essentially a sole visible emission line.

**Keywords:** Statistical Signal Processing, Sparsity, Detection, Estimation, Hyperspectral Imaging.

**Méthodes de détection parcimonieuses pour signaux faibles dans du bruit. Application à des données hyperspectrales de type astrophysique.**

**Abstract:** Cette thèse contribue à la recherche de méthodes de détection de signaux inconnus à très faible Rapport Signal-à-Bruit. Ce travail se concentre sur la définition, l'étude et la mise en œuvre de méthodes efficaces capables de discerner entre observations caracterisées seulement par du bruit de celles qui au contraire contiennent l'information d'intérêt supposée parcimonieuse. Dans la partie applicative, la pertinence de ces méthodes est évaluée sur des données hyperspectrales.

Dans la première partie de ce travail, les principes à la base des tests statistiques d'hypothèses et un aperçu général sur les représentations parcimonieuses, l'estimation et la détection sont introduits.

La deuxième partie du manuscrit est divisée en deux sous-parties. Dans la première sous-partie, deux tests d'hypothèses statistiques sont proposés et étudiés. Les deux sont adaptés à la détection de signaux parcimonieux grâce à des estimations parcimonieuses qui peuvent être interprétées au sens du Maximum A Posteriori ou dans le cadre des Moindres Carrés Pénalisés. La première approche est le test du Rapport des Distributions a Posteriori (PDR, *Posterior Density Ratio*), qui calcule le rapport des distributions a posteriori sous chaque hypothèse du modèle. La seconde approche est le test du Rapport de Vraisemblance où l'estimation par Maximum A Posteriori remplace l'estimation par Maximum de Vraisemblance (LRMAP, *Likelihood Ratio with Maximum A Posteriori estimate*). Les comportements et les différences relatives entre ces rapports de vraisemblance contraints sont étudiés à travers une étude détaillée de leurs caractéristiques analytiques et structurelles. Les performances de détection des tests sont comparés à celles de méthodes fréquentistes et Bayésiennes classiques telles que le test du Rapport de Vraisemblance Généralisé non contraint et le facteur de Bayes.

Conformément aux données tridimensionnelles considérées dans la partie applicative, et pour se rapprocher de scénarios plus réalistes impliquant des systèmes d'acquisition de données, la deuxième sous-partie présente des méthodes de détection qui cherchent à: i) prendre en compte le bruit fortement variable spectralement, ii) exploiter les similitudes spectrales de pixels spatialement voisins, et iii) exploiter un modèle plus précis basé sur des dictionnaires qui prennent en compte l'effet d'étalement spatio-spectral de l'information causée par les fonctions d'étalement du point de l'instrument.

Les tests sont finalement appliqués à des données astrophysiques massives de type hyperspectral dans le contexte du *Multi Unit Spectroscopic Explorer* de l'Observatoire Européen Austral (première lumière prévue fin 2013). Les méthodes proposées s'avèrent efficaces pour la détection de sources générales avec cet instrument. Une attention particulière est finalement accordée à la détection des sources très faibles et rares appelées émetteurs Ly$-\alpha$, caractérisés par un spectre qui contient essentiellement une seule raie en émission.

**Keywords:** Traitement Statistique du Signal, Parcimonie, Détection, Estimation, Imagerie Hyperspectrale.

**Metodi di sparsi di detezione per segnali deboli in presenza di rumore. Applicazione a dati iperspettrali di tipo astrofisico.**

**Abstract:** Questa tesi affronta il problema della detezione di segnali non noti caratterizzati da un basso Rapporto Segnale/Rumore. In particolare, questo lavoro si concentra sulla definizione, lo studio e l'implementazione di metodi efficienti in grado di discernere le osservazioni composte totalmente da rumore da quelle che contengono presumibilmente l'informazione di interesse. La rilevanza di questi metodi é stata verificata valutando l'applicazione al contesto di dati iperspettrali.

Nella prima parte di questo lavoro, si introducono i principi dei test di verifica di ipotesi, delle rappresentazioni sparse, e della stima e detezione del segnale.

La seconda parte del manoscritto è divisa in due sotto-parti. Nella prima, vengono proposti ed analizzati due test statistici di verifica di ipotesi. Entrambi sono adattati alla rilevazione di segnali sparsi grazie all'utilizzo di stime sparse che possono essere interpretate sia nel contesto del Massimo A Posteriori che in quello dei Minimi Quadrati Pesati. Il primo metodo proposto è il test del Rapporto delle Densità a Posteriori (PDR, *Posterior Density Ratio*), il quale calcola il rapporto tra le distribuzioni a posteriori sotto ciascuna delle ipotesi del modello considerato. Il secondo metodo è un test del Rapporto di Verosimiglianza, nel quale la stima del Massimo A Posteriori sostituisce la stima di Massima Verosimiglianza (LRMAP, *Likelihood Ratio with Maximum A Posteriori estimate*). I comportamenti e le differenze relative tra questi due rapporti di verosimiglianza vincolati sono considerati attraverso uno studio dettagliato delle loro caratteristiche analitiche e strutturali. Le prestazioni in termini di detezione dei test proposti sono confrontate con quelle di metodi classici frequentisti e Bayesiani, quali ad esempio il test del Rapporto di Verosimiglianza Generalizzato ed il Fattore di Bayes.

Conformemente ai dati tridimensionali considerati nella parte applicativa, la seconda sotto-parte presenta metodi di rilevazione che cercano di: i) tener conto della forte variabilità spettrale del rumore, ii) sfruttare le somiglianze spettrali tra pixel spazialmente vicini e iii) trarre vantaggio dalla maggiore precisione ottenuta considerando modelli di dati basati sull'utilizzo di dizionari, tenenti conto della dispersione spazio-spettrale della informazione causata dalla Funzione di Dispersione di un Punto dello strumento.

I test sono infine applicati a dati iperspettrali di tipo astrofisico nel contesto del *Multi Unit Spectroscopic Explorer* dell'Osservatorio Europeo Australe. I risultati ottenuti mostrano che i metodi proposti sono efficaci nel rilevamento di segnali informativi con questo strumento. Particolare attenzione è stata rivolta alla individuazione di fonti molto deboli e sparse note come emettitori Ly$-\alpha$, caratterizzate da uno spettro contenente essenzialmente una sola linea visibile in emissione.

**Keywords:** Trattamento Statistico del Segnale, Sparsità, Detezione, Stima, Imaging Iperspettrale.

# Contents

**Contents** ix

**Introduction and notations** 1

**I Statistical hypothesis testing and sparsity: a survey** 11

**1 Statistical hypothesis testing: a general overview** 13
1.1 Introduction . . . 13
1.2 Building a hypothesis test . . . 14
1.2.1 Definition of the hypotheses . . . 14
1.2.2 The test statistic . . . 15
1.2.3 Probability of error and significance level of a test . . . 16
1.2.4 The $p$-value . . . 19
1.2.5 Performance of a test: ROC curves and AUC . . . 20
1.3 Major approaches to hypothesis testing . . . 23
1.3.1 The Likelihood Ratio test . . . 23
1.3.2 The Generalized Likelihood Ratio test . . . 26
1.3.3 Bayesian approach to hypothesis testing and the Bayes Factor . . . 30
1.4 Conclusion . . . 32

**2 Sparse representations, estimation and detection** 33
2.1 Introduction . . . 33
2.2 Sparse representations of signals . . . 34
2.2.1 Sparsity of signals in a dictionary of atoms . . . 34
2.2.2 The choice of the dictionary . . . 35
2.3 Sparse estimation . . . 36
2.3.1 Penalty functions and thresholding estimators . . . 37
2.3.2 Bayesian framework for sparsity-promoting estimators . . . 41
2.3.3 MAP sparsity-promoting estimate of unknown parameters . . . 44
2.4 Sparse detection . . . 47
2.4.1 Previous works in sparse detection . . . 47
2.4.2 The Higher Criticism . . . 48
2.4.3 A glance at sparse detection in Hyperspectral Imaging . . . 51
2.5 Conclusion . . . 54

**II Sparsity-based Detection Tests: Theory and Methods** 55

**A — Description and study of the proposed approaches** 57

**3 Sparsity-based detection tests** **59**
3.1 Introduction . . . 59
3.2 A first approach: the LRMAP detection test . . . 60
3.3 A second approach: the PDR detection test . . . 61
3.4 Comparison of the tests performances in the scalar case . . . 63
3.4.1 Scalar model and test statistics . . . 63
3.4.2 Analytical properties of the tests . . . 64
3.4.3 Numerical simulations . . . 68
3.5 Comparison of the tests performances in the vector case . . . 69
3.5.1 $T_{\text{PDR}}$ and $T_{\text{LRMAP}}$ parameter dependency and asymptotic results . . . 69
3.5.2 On the relative power of the PDR and LRMAP tests . . . 70
3.5.3 Numerical simulations . . . 73
3.6 Conclusion . . . 74

**4 Geometrical study of the PDR and LRMAP tests** **75**
4.1 Introduction . . . 75
4.2 Study of the detection surfaces for $N = 2$ . . . 76
4.2.1 Shape of the detection surfaces . . . 76
4.2.2 Special cases of equivalence . . . 78
4.3 General N-dimensional case . . . 79
4.4 Comparison of the tests performances . . . 81
4.5 Conclusion . . . 81

**B — Refined sparsity-based detection strategies** **83**

**5 Spectral model based strategies** **85**
5.1 Introduction . . . 85
5.2 1D (spectral) model . . . 86
5.2.1 ML estimate of $\boldsymbol{\theta}$ and $T_{\text{GLR}}$ for a dictionary-based model . . . 87
5.2.2 MAP estimate of $\boldsymbol{\theta}$ for a dictionary-based model . . . 87
5.2.3 $T_{\text{PDR}}$ and $T_{\text{LRMAP}}$ test statistics for a dictionary-based model . . . 88
5.2.4 Computation of $\text{P}_{\text{FA}}$ and $\text{P}_{\text{DET}}$ at $\gamma = 0$: the PDR/LRMAP test . . . 89
5.3 The *Max* test as a constrained GLR . . . 90
5.4 Comparison between the $\text{GLR}_{\text{1s}}^{(\text{1D})}$ test and the HC . . . 91
5.5 Conclusion . . . 95

**6 Spatio-spectral model based strategies** **97**
6.1 Introduction . . . 97
6.2 First attempts to account for spatial dependencies . . . 98
6.2.1 A simple toy-case with equal spectra and same noise statistics . . . 98
6.2.2 Refined detection model: proportional neighbor spectra and spatially variable noise statistics. . . . 102
6.2.3 Multiple-round detection strategy . . . 106
6.3 A (spatio-spectral) 3D model-based strategy . . . 109
6.3.1 3D data Model and $\text{GLR}_{\text{1s}}^{(\text{3D})}$ detection test . . . 110
6.3.2 Design strategies for the choice of the dictionary . . . 111

6.4 Conclusion . . . 112

III Detection of Faint Sources in Astrophysical Hyperspectral Data 113

7 Detection results using spectral model based strategies 115
7.1 Introduction . . . 115
7.2 Hyperspectral Imaging and its application to Astrophysics . . . 116
7.2.1 Spectroscopy in Astronomy . . . 116
7.2.2 Hyperspectral Imaging . . . 118
7.2.3 An instance of Hyperspectral Imaging in Astrophysics: The MUSE instrument . . . 119
7.2.4 MUSE instrument's Point Spread Function . . . 120
7.3 Sparsity-constrained vs unconstrained GLR test . . . 121
7.3.1 The design of the dictionary $\mathbf{R}$ . . . 122
7.3.2 Comparison of the detection performances on a single test spectrum 123
7.3.3 Comparison of the detection performances on hyperspectral data cubes 126
7.3.4 Analysis of the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ test detection results in terms of $p$-value . . . 129
7.4 Conclusion . . . 131

8 Detection results using spatio-spectral model based strategies 133
8.1 Introduction . . . 133
8.2 Multiple-round detection strategy applied to MUSE data . . . 133
8.3 Detection results of the spatio-spectral model based approach on MUSE data 139
8.3.1 3D (spatio-spectral) dictionary . . . 139
8.3.2 Spatial $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ vs spatio-spectral $\mathrm{GLR}_{1\mathrm{s}}^{(3\mathrm{D})}$: Detection results on three faint Ly-$\alpha$ emitters . . . 141
8.4 Conclusion . . . 146

Conclusions and perspectives 147

A Appendix 155
A.1 Proof of NP Lemma . . . 155
A.2 Proof of BF as the minimum $\mathrm{P_E}$ detector . . . 155
A.3 Proof of Property 3.4.1 . . . 156
A.4 Proof of Proposition 4.3.1 . . . 158

B Appendix 161
B.1 Computation of the Hard Thresholding operator . . . 161
B.2 Computation of the BF test statistic for model (3.18) . . . 162
B.3 Learning techniques for dictionary selection: the K-SVD approach . . . 163
B.4 Asymptotic approximation of $T_{\mathrm{LRMAP}}$ . . . 165
B.5 A tentative of comparison of the tests by analysis of the $\mathrm{P_{FA}}$ and $\mathrm{P_{DET}}$ bounds 166

Bibliography 169

# List of Figures


1.1 Probabilities of right and wrong decisions evaluating a decision rule . . . . . 17
1.2 Test statistics distributions and decision regions: one-dimensional case. . . . 19
1.3 Illustration of the $P_{FA}$ and the $p$-value for a one-sided, right-tailed test. . . 20
1.4 Example of Receiver Operating Characteristic . . . . . 22
1.5 Test statistics distributions and decision regions: two-dimensional case. . . . 26
1.6 Scalar GLR test: data distributions and decision regions . . . . . 29
1.7 Vector GLR test: data distributions and decision regions . . . . . 29
1.8 Comparison of the LR and GLR ROC curves . . . . . 30

2.1 Synthesis of a $N-$dimensional vector $\mathbf{x}$ in a redundant dictionary $\mathbf{R}$ . . . . 35
2.2 Examples of $l_1$ penalty functions and corresponding ST operators . . . . . 38
2.3 Example of clipped $l_1$ penalty function and corresponding HT operator . . . 40
2.4 Example of SCAD penalty function and corresponding thresholding operator 40
2.5 Examples of Quadratic and Hit-or-Miss Loss functions . . . . . 42
2.6 Examples of Laplacian and GG Prior distributions . . . . . 45
2.7 Analytical and Empirical CDF of $p$-values and illustration of the HC statistic 50

3.1 Example of a thresholded test statistic considered in Property 3.4.1 . . . . . 65
3.2 NP, GLR, LRMAP, PDR and BF ROC curves in the case of a scalar parameter 68
3.3 Empirical distributions of the test statistics $T_{\text{LRMAP}}$ and $T_{\text{PDR}}$ . . . . . 70
3.4 PDR, LRMAP, BF and GLR ROC curves . . . . . 72
3.5 LRMAP and PDR ROC curves . . . . . 73

4.1 Illustration of the tests detection surfaces for $N = 2$ . . . . . 77
4.2 3D detection surfaces in the two different cases of equivalence of the tests . 78
4.3 3D $\mathcal{S}_{\text{PDR}}$ and $\mathcal{S}_{\text{LRMAP}}$ detection surfaces . . . . . 79
4.4 No intersection and superposition cases of the $\mathcal{S}_{\text{PDR}}$ and $\mathcal{S}_{\text{LRMAP}}$ 3D detection surfaces . . . . . 80
4.5 PDR and LRMAP AUCs difference as a function of $\eta$ threshold values . . . 82

5.1 Numerical comparisons of the $\text{GLR}_{\text{1s}}^{(\text{1D})}$ vs. the HC . . . . . 93
5.2 Numerical comparisons of the $\text{GLR}_{\text{1s}}^{(\text{1D})}$ test vs. the HC: case of less sparse signals . . . . . 94

6.1 Detection exploiting spatial dependencies: bright and faint noiseless simulated spectra . . . . . 102
6.2 ROC curves: "Spectra considered independently" vs. "Spatial model $\mathbf{x}_f = \mathbf{x}_b$" 103
6.3 Faint spectrum estimate using model (6.30) . . . . . 106
6.4 Detection performances of the LR-MP$\beta$ test . . . . . 107
6.5 An illustration of the multiple-round detection strategy . . . . . 108

7.1 Example of a star (in blue) and galaxy (in red) spectrum. . . . . 117
7.2 Example of hyperspectral image and spectrum associated to a pixel . . . . . 118

7.3 A 3D illustration of the MUSE instrument . . . 119
7.4 Example of noise variances affecting the components of a MUSE spectrum. . 120
7.5 MUSE's WSF . . . 121
7.6 MUSE's SSF . . . 121
7.7 Matrix form of the Wavelength Spread Function . . . 122
7.8 Examples of atoms in dictionary R . . . 123
7.9 Comparison of the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ and unconst. GLR tests on a MUSE spectrum . 125
7.10 Comparison of the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ and unconst. GLR tests on a MUSE data cube 127
7.11 Reconstructed white light image of the noiseless dry-run data cube . . . 128
7.12 Compared detection performances of the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ and unconst. GLR tests on the *dry-run* . . . 128
7.13 Comparison of the $p$-value maps computed for all spectra in a MUSE sub-cube 130

8.1 Detection results using a two-step detection strategy on a MUSE sub-cube . 137
8.2 Detection results using a two-step detection strategy on MUSE dry-run . . . 138
8.3 Typical spectra under $\mathcal{H}_1$ and corresponding data for $(16, 7)$. Coordinates $(x, y)$ correspond to spatial position in Figure 8.8(a). . . . 139
8.4 Examples of spectral lines and trained atoms. . . . 140
8.5 Example of 3D minimax atom . . . 141
8.6 Empirical correspondence between $\mathrm{P}_{\mathrm{FA}_0}$ and $\xi$ for the SVD 3D atom . . . 142
8.7 Examples of 3D SVD atoms . . . 144
8.8 Compared detection performances of the unconst. GLR and the 1D and 3D $\mathrm{GLR}_{1\mathrm{s}}$ . . . 145

A.1 Intersections between the $\mathcal{S}_{PDR}^{(n)}$ and $\mathcal{S}_{LRMAP}^{(n)}$ domains when $n \geq 2$ . . . 158

B.1 Maximizing-minimizing surfaces for the LRMAP and PDR tests . . . 167

# Introduction and notations

# Introduction

In everyday life, decision making takes place every time we have to make a choice in face of uncertainty. Imagine a doctor examining a patient and trying to diagnose whether the symptoms he shows are due to the presence of a particular disease or not. Even if the decision reduces to a simple yes-or-no response, the presence of ambiguity (e.g. misleading symptoms) can make the decision problem very difficult to solve. In this and countless others kinds of decision problems, precise scientific techniques are needed to help in returning the right verdict and assessing the risk of being misled by the observations.

A useful means to accurately describe, study and quantify in probabilistic terms the majority of decision problems is given by the Signal Detection Theory (SDT). The beginnings of SDT date back to the 1950s, when researchers in the field of radar systems needed reliable tools to determine the presence or absence of an approaching aircraft when monitoring the sky [Skolnik 1980]. Since these origins, SDT has come to be used in a wide variety of fields such as defense, psychology, medicine and engineering.

In SDT, a *signal* is the physical representation of any kind of information that varies with time, frequency, energy or any other explanatory variable. Given a set of observed data, a general signal detection approach consists in stating assumptions (or *hypotheses*) allowing to explain or interpret the observations, and then decide which one between these hypotheses is (or seems to be) true. Under each hypothesis, the observed data can be statistically described by probability density functions (or probability mass functions, in the discrete case). Depending on the hypotheses made and on the nature of the data, these densities can be completely or partially known.

The reasons that may make signal detection problems hard to solve are multiple. First of all, the observations on which we seek to make decision are usually affected by *noise*. The noise can take many forms, it is not necessarily a stochastic signal, and it is generally represented by any signal that masks the information we seek to acquire. In the example of the medical diagnosis considered above, the noise could be represented by the presence of misleading symptoms that do not allow the doctor to diagnose the disease with certainty; for an antenna pointing towards the sky, as for example a radio-telescope, any other existing transmitter is a nuisance for the radio astronomer who is interested in receiving signals from the stars; other noise examples are for instance detection and calibration errors, and the atmospheric effects affecting sensor-based systems that provide the information.

Second, the informative signal we seek to find may be unknown: the detection of signals of unknown shape or position (e.g. signals coming from sensor devices whose location is not known) is more difficult with respect to the case of perfect knowledge. Third, the informative signal may have amplitude much lower than the noise level characterizing the considered observations. This could be due to the presence of very strong perturbations, to the intrinsically weak nature of the signal of interest or, in particularly unfortunate circumstances, to both these events.

Most often, the signals we seek to find can be compactly modeled as sparse. A signal is called sparse when most of its coefficients are (possibly approximatively) null. Many signals are not sparse in their nature but often show up to be sparse in the domain of appropriate transforms. According to this, sparse representations are based on the assumption that

a considered set of observations (for instance a time series, an image or a spectrum) can be described by a very reduced number of parameters defined in an adequate space of representation. For example, audio signals are sparse in the time-frequency domain, while images can be often represented by a sparse decomposition in an appropriate transform basis (Discrete Wavelet Transforms, for instance).

The question addressed in this thesis is challenging because it considers cases that combine the difficulties described above. We consider methods accounting for strong variable noise level in the observations and for unknown signals assumed to be sparse, with sometimes very low amplitudes. In this framework, the devising of efficient detection techniques will call for expertise from several related scientific domains such as statistics, sparse representations and approximation algorithms, and the theories of detection and of estimation.

## Thesis context

The research work presented in this manuscript has been carried out from October 2010 to – 2013, at the J.L. Lagrange Laboratory (UMR7293, University of Nice - Sophia Antipolis, CNRS, Observatoire de la Côte d'Azur). All along the thesis, an enriching collaboration has been held within the ANR DAHLIA Project[1], dedicated to the development of new algorithms for hyperspectral imaging in astronomy.

## Objectives

The main target of this work is to define and study methods for the detection of unknown signals in highly noisy data. In particular, our research focuses on signals that are weak and sparse that is, characterized by a very low number of informative components. Given a set of noisy observations, we thus seek to discern only noise signals to those that presumably carry the information of interest. The difficulty in this kind of problem is twofold:

1. Methodological: definition of statistical tests appropriate to the considered framework of weak and sparse signals, and definition of new contributions and their connections with respect to the existing literature (Part II of the manuscript);

2. Applicative: determination of consistent and realistic statistical data models; design, study and implementation of efficient detection techniques in the context of astrophysical hyperspectral data for an existing instrument (Part III of the manuscript).

## Contributions

The works presented in this thesis have led to several scientific publications in the general field of signal detection, with particular attention to the problem of detecting faint sparse sources in highly noisy multi-dimensional data. All the contributions are listed below.

1. In [Paris 2011a], a precise study of the connections existing between the sparse estimation and the sparse detection frameworks has been proposed. Two new detection

[1]The DAHLIA Project partially supported this work through founding the attendance to the EUSIPCO 2011 conference for the presentation of the work [Paris 2011c]

tests have been presented: the Posterior Density Ratio (PDR) test, and a Likelihood Ratio (LR) test using a Maximum A Posteriori (MAP) estimate with Generalized Gaussian priors. We called this last approach the LRMAP test. The tests have been analyzed and compared to classical detection strategies such as the Generalized Likelihood Ratio test and the Bayes Factor, in the case of scalar data model first and then in the vector case. In the vector case, the two tests based on MAP estimates have shown to outperform the classical approaches thanks to their test statistics tending to focus on the informative components of the parameters vector, in accordance to the thresholding properties of the scalar MAP estimate.

2. In [Paris 2011c], the PDR and the LRMAP detection tests were set in the context of a dictionary-based data model for astrophysical hyperspectral data. The considered dictionary was designed by finely discretising elementary spectral features (lines with various widths, Heaviside steps, and continuum parameterisation). A study on the relative detection performances of the two approaches in function of the tests' parameters $(\eta, \gamma)$ was proposed. The tests were then analysed in the particular case of setting the test threshold $\gamma$ to zero. This setting allowed to easily implement the tests and to univocally compute the corresponding Probability of False alarm. A first *toy*-model designed for the exploitation of the spatio-spectral dependencies of neighbors pixels in a given data set was introduced.

3. In [Paris 2011d], an asymptotic analysis (in the vector size) and a finite size analysis of the performances of the PDR and the LRMAP tests was proposed.

4. In [Paris 2011b], the detection performances of the PDR and the LRMAP tests have been improved introducing refined spatio-spectral models allowing to take into account spectral similarities existing between neighbors pixels in hyperspectral data cubes.

5. In [Paris 2012], a precise analysis of the detection performances of the PDR and the LRMAP tests in terms of $p$-values was proposed. This analysis allowed to assess post-data statistical significance of the proposed tests.

6. In [Paris 2013b], a new method aimed at detecting weak, sparse signals in highly noisy three-dimensional (3D) data has been introduced. Such data suffer from information leakage caused by the acquisition system's point spread functions. This function may be different and variable in both the two image dimensions and the third dimension (temporal, spectral or energy dimension). The detection test proposed in this work was based on dedicated 3D dictionaries, and exploited both the sparsity of the data along the third direction and the information spread in the three dimensions. This method revealed to be efficient for the detection of very faint and sparse objects such as Ly$-\alpha$ objects and evidenced detection limits of instrument MUSE to which it was applied.

7. In [Paris 2013a], we deepened the study of the PDR and the LRMAP testing procedures. The behaviors and differences between the two tests were here investigated through a detailed geometrical study of their structural characteristics. As a result, we found out that the PDR test tended to better detect very sparse signals while the LRMAP test was more adapted for less sparse signals.

8. A journal paper is currently in preparation for the presentation of the detection methods developed in this work to the astrophysics community.

## Organization of the manuscript

This manuscript is organized in three principal parts:

1. In the first part, Chapter 1 is dedicated to the description of the basic principles of statistical hypothesis testing for decision making.

   Chapter 2 introduces general principles in the sparse representation and sparse estimation problems, and provides an overview of major previous works existing in the sparse detection literature. This analysis allows to define the general issues of this thesis and to justify the choices made for the rest of the manuscript.

2. The second part of the manuscript, which contains the main contributions of this research work, is divided into two sub-parts.

   - Sub-part II-A: in Chapter 3, two statistical hypotheses tests are proposed and studied, both adapted to the detection of sparse signals and based on the Maximum A Posteriori estimate of the unknown vector of parameters. The first approach is the Posterior Density Ratio test, which computes the ratio of the a posteriori distribution under each hypothesis of the data model. The second is a Likelihood Ratio test in which the Maximum A Posteriori estimate (or in another interpretation, a Penalized Least Square estimate) replaces the Maximum Likelihood estimate. After the definition of the two approaches, the behaviors and the relative differences between these tests are theoretically investigated and the tests' detection performances compared with those of classical frequentist and Bayesian methods such as the Generalized Likelihood Ratio test and the Bayes Factor.

     In order to state which one of the two approaches better adapts to the detection of faint sparse signals, a geometrical study of the structural differences between the two detection techniques is addressed in Chapter 4.

   - Sub-part II-B: According to the three-dimensional data sets considered in the applicative part, and to be closer to realistic scenarios involving data acquisition systems, this sub-part presents detection methods seeking to exploit both the spectral similarities of neighbors pixels in the spatial domain, and the greater accuracy of dictionary-based models that take into account the spatio-spectral blur of information caused by instrumental Point Spread Functions.

     In Chapter 5, the detection tests of Chapter 3 are reformulated and adapted to the case of a dictionary-based spectral model. A special case of equivalence of the proposed tests is found. In this case, the tests are equivalent to the *Max* detection approach introduced in [Arias-Castro 2010]. A simple alternative formulation of the *Max* detector as a constrained Generalized Likelihood Ratio test is proposed. Numerical comparisons between the constrained Generalized Likelihood Ratio test and the Higher Criticism detector of [Donoho 2004] are made.

Detection strategies for the identification of weak, sparse signals in highly noisy three-dimensional data are introduced in Chapter 6. Here, the developed approaches seek to exploit both the spectral similarities of neighbors pixels in the spatial domain, and the greater accuracy of dictionary-based models that take into account the spatio-spectral blur of information caused by instrumental Point Spread Function. With respect to Chapter 5, improved detection performances are reached through the introduction of a multiple-round detection approach based on a simple model of spatial dependencies. A more accurate and more easily exploitable 3D (spatio-spectral) model-based detection test, which accounts explicitly for the instrument's Point Spread Function model, is then considered.

3. The third part concerns the application of the detection approaches to the identification of weak and sparse sources in astrophysical three-dimensional data sets. After a brief introduction to hyperspectral imaging and some of its applications to astrophysics, detection results are shown in Chapter 7 and Chapter 8 for the detection tests implemented in the second part (sub-part II-B) of the manuscript.

# Notations

## General rules

| | |
|---|---|
| $a$ | Scalar |
| $\mathbf{a}$ | Vector |
| $\mathbf{A}$ | Matrix |
| $\approx$ | Approximated to |
| $a \sim F$ | $a$ follows distribution $F$ |
| $a \gtrless T$ | $a$ greater or less than $T$ |
| $\sum_{i=1}^{N}$ | Summation |
| $\|\|\mathbf{a}\|\|_p$ | $l_p-$ norm of vector $\mathbf{a}$ |
| $\mathbb{E}(\mathbf{a})$ | Expectation of $\mathbf{a}$ |

## Sets and Spaces

| | |
|---|---|
| $\mathcal{H}_0, \mathcal{H}_1$ | Hypothesis 0, Hypothesis 1 |
| $\{\mathbf{x} : P\}$ | set of $\mathbf{x}$ with property $P$ |
| $\{i, \ldots, j\}$ | set of integers between $i$ and $j$ |
| $\langle \mathbf{T} \rangle$ | Subspace spanned by the columns of $\mathbf{T}$ |
| $\mathbb{R}^N$ | $N-$dimensional Euclidian space |

## Matrices and Vectors

| | |
|---|---|
| $[a_1, a_2, \ldots, a_N]$ | $(1 \times N)$ vector with components $a_i, i = 1, \ldots, N$ |
| $[\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_N]$ | Matrix with columns $\mathbf{a}_i, i = 1, \ldots, N$ |

| | |
|---|---|
| $\mathbf{I}$ | Identity matrix |
| $\mathbf{0}$ | Vector of zeros |
| $\mathbf{1}$ | Vector of ones |
| $\boldsymbol{A}^t$ | Transpose of a matrix A |
| $\boldsymbol{A}^{-1}$ | Inverse of a matrix A |
| $\det(\boldsymbol{A})$ | Determinant of a matrix A |
| $\mathrm{Tr}\{\boldsymbol{A}\}$ | Trace of a matrix A |
| $\mathrm{diag}[a_1, \ldots, a_N]$ | Diagonal matrix with non-zero elements $a_i, i = 1, \ldots, N$ |

## Acronyms

| | |
|---|---|
| AUC | Area Under the Curve |
| BF | Bayes Factor |
| BPDN | Basis Pursuit DeNoising |
| CDF | Cumulative Distribution Function |
| ESO | European Southern Observatory |
| GLR | Generalized Likelihood Ratio |
| HC | Higher Criticism |
| HT | Hard Thresholding |
| LASSO | Least absolute shrinkage and selection operator |
| LR | Likelihood Ratio |
| LRMAP | Likelihood Ratio with Maximum A Posteriori estimate |
| MAP | Maximum A Posteriori |
| ML | Maximum Likelihood |
| MP | Matching Pursuit |
| MSE | Mean Square Error |
| MUSE | Multi Unit Spectroscopic Explorer |
| NP | Neyman-Pearson |
| PDF | Probability Density Function |
| PDR | Posterior Density Ratio |
| PLS | Penalized Least Square |
| PSF | Point Spread Function |
| ROC | Receiver Operating Characteristic |
| SCAD | Smoothly Clipped Absolute Deviation |
| SSF | Spatial Spread Function |
| ST | Soft Thresholding |
| SVD | Singular Values Decomposition |
| VLT | Very Large Telescope |
| WSF | Wavelength Spread Function |

# Part I

# Statistical hypothesis testing and sparsity: a survey

Chapter 1

# Statistical hypothesis testing: a general overview

**Contents**


**1.1 Introduction . . . . . 13**
**1.2 Building a hypothesis test . . . . . 14**
1.2.1 Definition of the hypotheses . . . . . 14
1.2.2 The test statistic . . . . . 15
1.2.3 Probability of error and significance level of a test . . . . . 16
1.2.4 The $p$-value . . . . . 19
1.2.5 Performance of a test: ROC curves and AUC . . . . . 20
**1.3 Major approaches to hypothesis testing . . . . . 23**
1.3.1 The Likelihood Ratio test . . . . . 23
1.3.2 The Generalized Likelihood Ratio test . . . . . 26
1.3.3 Bayesian approach to hypothesis testing and the Bayes Factor . . . . 30
**1.4 Conclusion . . . . . 32**


## 1.1 Introduction

Every time we face situations in which, from one or more observations of a physical phenomenon, described through the use of an appropriate (probabilistic) model, we seek to derive information on it, we find ourselves confronted to a *statistical inference* problem. We call statistical inference the process allowing to draw reliable statistical conclusions on the basis of experimental (observation) data. The type of information we seek to derive may relate to either the determination (or estimation) of the parameters of the model, or to the validity of scientific hypotheses on which this same model is built. In the latter case, the question about which one of the stated hypotheses best matches the data is answered through *hypothesis testing* methods. In hypothesis testing methods the number of tested hypotheses must be at least of two, and varies in relation to the problem we consider. We call *binary* tests all methods testing only *two* hypotheses, whilst for a number of hypotheses greater than two we refer to *multiple* hypothesis tests.

Another distinction we make concerns the partial or the complete knowledge of the Probability Density Functions (PDFs) of data under the considered hypotheses. If the PDFs are completely known we refer to *simple* hypothesis tests. On the contrary, *composite*

hypothesis tests have to manage with PDFs that are not completely specified under one or more hypotheses of the model, because of their dependence on unknown parameters.

Hypothesis testing is a widely used topic in many different contexts of signal processing and communications, as well as in biology, medicine, and other fields in which, given a set of noisy data, we usually have to choose between two (or more) alternative hypotheses. To make some examples, we may need to determine: whether or not a considered population shows a particular characteristic; whether a person does or does not have a particular disease; whether or not a received noisy signal (or observation) contains salient information.

This chapter, discusses first the basic steps we need to follow to design and then assess the statistical significance of a hypothesis test. This is a very general introduction presenting the very first principles of hypothesis testing. Secondly, an overview on the general main approaches to hypothesis testing is given. For a more exhaustive analysis, we refer to the comprehensive work of [Kay 1998a].

## 1.2 Building a hypothesis test

Classical hypothesis tests take account of four main parts, which always come in the following order:

- State hypotheses;
- Identify a test statistic;
- Specify a significance level;
- Make a decision.

To make clear the discussion on the structure of a hypothesis test, we shall keep in mind the general signal detection problem of determining the presence or absence of significant information over a noisy observation. This example could reflect, for instance, the particular situation where from a received signal of a radar device should be stated the presence or absence of an aircraft.

### 1.2.1 Definition of the hypotheses

A hypothesis (or claim) is a statement, or tentative explanation, about the available data. In a hypothesis test, the number of considered hypotheses must be at least of two. In this and the following sections we consider the case of *binary hypothesis tests* that is, tests characterized by only two (different) claims, called

$$\begin{aligned} \mathcal{H}_0 &: \text{ the } \textit{null hypothesis}; \\ \mathcal{H}_1 &: \text{ the } \textit{alternative}. \end{aligned}$$

A hypothesis test quantifies evidence for the alternative hypothesis $\mathcal{H}_1$, by setting up the null hypothesis $\mathcal{H}_0$ as the opposite of $\mathcal{H}_1$, and tries to find and quantify evidence to disprove it. The final decision always provides two possible conclusions: accept $\mathcal{H}_0$ as the true model, or accept $\mathcal{H}_1$ (reject the null). The hypotheses of a test are formulated on the parameter(s) from which the model may depend. For hypotheses defined on a given scalar parameter

$\theta \in \Theta$, where $\Theta$ represents the space in which $\theta$ takes values, we are interested in hypotheses tests of the form:

$$\mathcal{H}_0 : \quad \theta = \theta_0 \in \Theta_0; \tag{1.1}$$
$$\mathcal{H}_1 : \quad \theta = \theta_1 \in \Theta_1, \tag{1.2}$$

where $\theta_0$ is the value of $\theta$ under $\mathcal{H}_0$ taken in $\Theta_0$, $\theta_1$ the value of $\theta$ under $\mathcal{H}_1$ taken in $\Theta_1$, the union of $\Theta_0$ and $\Theta_1$ defines the whole space $\Theta$ ($\Theta_0 \cup \Theta_1 = \Theta$) and the intersection of $\Theta_0$ and $\Theta_1$ is the empty space ($\Theta_0 \cap \Theta_1 = \emptyset$). The $\theta$ parameter on which the model depends may be known or unknown. If unknown, it may be assumed as *deterministic* or *stochastic* (random) leading to statistical inference in the *frequentist* or in the *Bayesian* frameworks, respectively. Unless clearly specified, when unknown, the case of deterministic parameters is considered in this chapter.

### 1.2.2 The test statistic

Once the hypotheses have been fixed, the very next question we need to solve is: "*How do we decide whether to accept* $\mathcal{H}_0$?" To answer this question, we need to focus on the problem of selecting an appropriate *decision criterion*, according to which decision is made on the veracity of the assertions we test. In practice, the selection of the decision criterion concerns

- the identification of a discriminating function;
- the choice of a test threshold to fix the significance level of the test.

The discriminating function, also known as *test statistic*, is a mathematical function of the data built according to the nature of the problem and to the hypotheses we want to test. Taking for example a data vector $\mathbf{x} = [x_1, \ldots, x_N]^t$, we define the test statistic as the function

$$T(\mathbf{x}) = T(x_1, \ldots, x_N). \tag{1.3}$$

This is a random variable if $\mathbf{x}$ is drawn from a random process. In a hypothesis test, $T(\mathbf{x})$ plays the fundamental role of decision variable. To be correctly designed, a test statistic $T(\mathbf{x})$ should provide as much information as possible on the considered problem, should be characterized by a probability distribution that is as different as possible under $\mathcal{H}_0$ and under $\mathcal{H}_1$, and should be completely known at least under $\mathcal{H}_0$.

All these points are not simple questions to solve, and the identification of a good test statistic sometimes requires a lot of effort and imagination, as it will be illustrated for instance in this work (see chapter 3).

When considering a testing problem, the discriminating function has no reason to be unique. The choice between several statistical models represents another point of interest, influencing for example the discriminating power of the test, that is, the probability of correctly rejecting the null hypothesis $\mathcal{H}_0$ under the alternative $\mathcal{H}_1$ (see subsection 1.2.3 for further details).

Once fixed, the decision variable $T(\mathbf{x})$ is compared to a *test threshold*, say $\xi \in \mathbb{R}$. In a compact mathematical form, we write:

$$T(\mathbf{x}) \underset{H_0}{\overset{H_1}{\gtrless}} \xi, \tag{1.4}$$

where the null hypothesis is rejected if $T(\mathbf{x})$ falls above the test threshold $\xi$, while it is accepted when $T(\mathbf{x})$ falls below $\xi$. Expression (1.4) represents a *decision rule*. The importance of the threshold $\xi$ resides in the fact that it allows to control the probability of making wrong decisions. In fact, when models for $\mathcal{H}_0$ and $\mathcal{H}_1$ involve noise, at any taken decision always corresponds a probability to make the right decision and a probability of being wrong. The definition and the study of such probabilities is presented in the next section.

We finally note that according to the hypotheses and to the test statistic's nature, three different types of hypothesis tests exist: the *two-tailed* (or two-sided) test, the *left-tailed* test, or the *right-tailed* test. The left- and the right-tailed tests are often referred to as *one-tailed* (or one-sided) tests. Examples of right-, left- and two-sided parameter testing problems in the case of simple hypotheses are given below.

**Example 1.2.1. - Right-, left- and two-sided parameter testing problems**

Given a parameter $\theta$ on which the models depend, we consider the following same formulation of the null hypothesis for the three types of model:

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

According to (1.2), we define the alternative hypothesis as the set of hypotheses

$$\mathcal{H}_1 : \theta \in \Theta_1, \text{ with } \theta_0 \notin \Theta_1.$$

On this basis, examples of alternatives leading to one-sided (right- and left- tailed) and of two-sided models are:

$$\begin{aligned} &\mathcal{H}_1 : \theta > \theta_0 && \text{one-tailed, right-sided models;} \\ &\mathcal{H}_1 : \theta < \theta_0 && \text{one-tailed, left-sided models;} \\ &\mathcal{H}_1 : \theta \neq \theta_0 && \text{two-tailed models.} \end{aligned} \tag{1.5}$$

Examples of right-tailed and two-tailed hypothesis tests in the case of Gaussian data distributions are given in subsection 1.3.1 and in subsection 1.3.2. ■

### 1.2.3 Probability of error and significance level of a test

Evaluating a decision rule of the type (1.4) consists in choosing between the following four possibilities: accepting $\mathcal{H}_i$ when $\mathcal{H}_j$ is true, with $i, j \in \{0, 1\}$. To each of those possibilities always corresponds a probability to make the right decision or a probability of being wrong. Of the four possible decision outcomes, the two leading to right decisions are those choosing the $\mathcal{H}_i$ hypothesis when $\mathcal{H}_i$ is indeed true (with $i = 0$ or $i = 1$).

In the particular case of having $i = 1$, we refer to the conditional *Probability of detection* (or *power* of a test)

$$\mathrm{P_{DET}} = Pr\left(\text{“accept } \mathcal{H}_1\text{”} \mid \text{“}\mathcal{H}_1 \text{ is true”}\right), \tag{1.6}$$

or, written more simply,

$$\mathrm{P_{DET}} = Pr\left(\mathcal{H}_1 \mid \mathcal{H}_1\right). \tag{1.7}$$

On the other hand, the two possible outcomes leading to wrong decisions are:

| | DECISION | |
|---|---|---|
| **TRUTH** | **Retain $\mathcal{H}_0$** | **Reject $\mathcal{H}_0$** |
| **$\mathcal{H}_0$ True** | **CORRECT** $(1 - \mathrm{P_{FA}})$ | **TYPE I ERROR** $(\mathrm{P_{FA}})$ |
| **$\mathcal{H}_0$ False** | **TYPE II ERROR** $(\mathrm{P_M})$ | **CORRECT** $(1 - \mathrm{P_M})$ |

Figure 1.1: Right and wrong decisions made evaluating a decision rule and corresponding probabilities.

- to accept $\mathcal{H}_1$ when $\mathcal{H}_0$ is true;
- to accept $\mathcal{H}_0$ when $\mathcal{H}_1$ is true.

The first case yields the conditional *Probability of false alarm*

$$\mathrm{P_{FA}} = Pr\left(\mathcal{H}_1 \mid \mathcal{H}_0\right), \tag{1.8}$$

while in the second case we obtain the *Probability of missed detection*

$$\mathrm{P_M} = Pr\left(\mathcal{H}_0 \mid \mathcal{H}_1\right). \tag{1.9}$$

The $\mathrm{P_{FA}}$ and the $\mathrm{P_M}$ are also respectively known as Type I and Type II errors. Note that, since

$$\mathrm{P_M} + \mathrm{P_{DET}} = 1, \tag{1.10}$$

the probability of detection can also be expressed as

$$\mathrm{P_{DET}} = 1 - \mathrm{P_M}. \tag{1.11}$$

Figure 1.1 shows the four different alternatives we can take about the truth or the falsity of the decision made on the null hypothesis.

**Example 1.2.2. - Illustration of the distribution of the test statistics**

Let us consider the case of a realization $x$ of a scalar random variable $X$ (that is, for example, an observation received from a radar device), which shows Gaussian PDF $p(x \mid \mathcal{H}_0) \sim \mathcal{N}(\theta_0, \sigma^2)$ under $\mathcal{H}_0$, and $p(x \mid \mathcal{H}_1) \sim \mathcal{N}(\theta_1, \sigma^2)$ under $\mathcal{H}_1$. The two PDFs here depend on the values $\theta_0$ and $\theta_1$ (with $\theta_1 > \theta_0$) which we assume to be known. Based on the single information $x$, we want to determine if $\theta = \theta_0$ (no signal present) or $\theta = \theta_1$ (signal present). On the basis of this single realization, assume we decide to reject $\mathcal{H}_0$ if our sample falls above a test threshold $\xi$ (this test actually corresponds to the scalar Likelihood Ratio test, whose detailed analysis is reported in subsection 1.3.1). With reference to the

decision rule defined in (1.4), where $T(x) = x$ in this case, and assuming we are examining a one-sided right-tailed test, with $\theta_1 > \theta_0$ and $T(x) \underset{H_0}{\overset{H_1}{\gtrless}} \xi$, the expression of the $\mathrm{P_{DET}}$ gives

$$\begin{aligned} \mathrm{P_{DET}} &= Pr\left(T(x) > \xi \mid \mathcal{H}_1\right) \\ &= \int_{\xi}^{+\infty} p(x \mid \mathcal{H}_1)\, \mathrm{d}x, \end{aligned} \tag{1.12}$$

with $p(x \mid \mathcal{H}_1)$ the data distribution under $\mathcal{H}_1$.

On the other hand, the two errors that we can make in this kind of situation are: a Type I error (occurring with probability $\mathrm{P_{FA}}$), expressed as

$$\begin{aligned} \mathrm{P_{FA}} &= Pr\left(T(x) > \xi \mid \mathcal{H}_0\right) \\ &= \int_{\xi}^{+\infty} p(x \mid \mathcal{H}_0)\, \mathrm{d}x, \end{aligned} \tag{1.13}$$

where $p(x \mid \mathcal{H}_0)$ is the data distribution under $\mathcal{H}_0$, or a Type II error (occurring with probability $\mathrm{P_M}$)

$$\begin{aligned} \mathrm{P_M} &= Pr\left(T(x) < \xi \mid \mathcal{H}_1\right) \\ &= \int_{-\infty}^{\xi} p(x \mid \mathcal{H}_1)\, \mathrm{d}x. \end{aligned} \tag{1.14}$$

This is illustrated in Figure 1.2. The blue and dark green curves here represent the data distributions under the $\mathcal{H}_0$ and the $\mathcal{H}_1$ hypotheses. If we assume these two distributions to be Gaussian, we have

$$p(x \mid \mathcal{H}_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\theta_i)^2} \tag{1.15}$$

with $i = 0, 1$, for the two hypotheses. Given a fixed test threshold $\xi$, the $\mathrm{P_{FA}}$ (1.13) is graphically shown by the region colored in yellow. The $\mathrm{P_M}$ is highlighted in magenta. As clearly visible, setting the value of the test threshold $\xi$ plays a critical role in defining the two probability regions. ■

Note that since it represents the risk we want to control, the $\mathrm{P_{FA}}$ covers a fundamental role in a binary hypothesis test. In particular, the $\mathrm{P_{FA}}$ tells us how extreme the observed results must be in order to reject the null hypothesis. To tune the statistical significance of a test, the $\mathrm{P_{FA}}$ value is usually set in relation to the number $N$ of observations *and* the context characterizing the considered application. For instance, we consider the case of an industrial robot finding defective bolts in the controlling process of a factory. $\mathrm{P_{FA}} = 5\%$ means in this case that on 100 tested bolts we expect to have 5 pieces wrongly detected as defective. For a company producing $20 \times 10^4$ pieces/day the chosen $\mathrm{P_{FA}}$ level leads to $10^4$ pieces wrongly claimed as defective over $20 \times 10^4$, which could be cheaply and rapidly further controlled by eye. But if instead of simple bolts we consider the case of testing high performance computers, setting $\mathrm{P_{FA}} = 5\%$ is not appropriate in this case. In fact, over an average of 100 machines produced per day, 5 of these will wrongly be labeled as defective, requiring further control more involved with consequent financial loss for the company.

By definition of the $\mathrm{P_{FA}}$, the lower the value fixed for the $\mathrm{P_{FA}}$, the less unlikely the data to be generated under $\mathcal{H}_0$. The $\mathrm{P_{FA}}$ is also called *significance level* (or size) of a test (often indicated with the symbol $\alpha$).

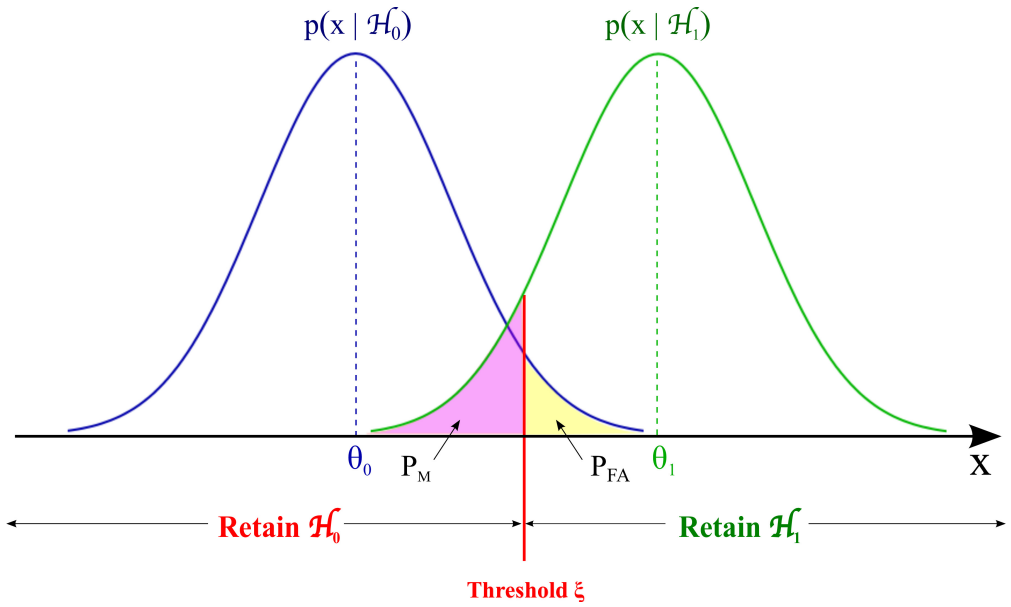



Figure 1.2: Distributions of the data under the $\mathcal{H}_0$ and $\mathcal{H}_1$ hypotheses of the model and decision regions for a one-sided right-tailed test, in a one-dimensional case.

### 1.2.4 The $p$-value

The significance level of a hypothesis test should not be confused with another well known statistical figure: the *p-value*. The $p$-value is one of the most conventional *post-data* measure of confidence to grant to a hypothesis. It represents a probability and thus takes values in the $[0, 1]$ interval. Given an observed value of the test statistic under $\mathcal{H}_0$, say $t_{obs}$, for a one-sided right-tailed test we write it as

$$p = Pr(T(x) \geq t_{obs} \mid \mathcal{H}_0). \tag{1.16}$$

Since it represents the smallest significance level at which we would reject the hypothesis $\mathcal{H}_0$ for the data at hand, the $P_{FA}$ is also called *observed significance level*. The $p$-value thus represents a kind of "*observed* $P_{FA}$": it is the probability that the random variable $T(x)$ is at least as "surprising" (that is, as extreme) as the observed value under the null hypothesis.

Under the null hypothesis, $p$-values $p$ are uniformly distributed. This can be demonstrated by the "Probability integral transform theorem" as follows. From (1.16) we have that

$$p = 1 - Pr(T(x) < t_{obs} \mid \mathcal{H}_0). \tag{1.17}$$

Denoting by $F_T(\cdot)$ the Cumulative Distribution Function (CDF) of the test statistic $T(x)$ under the null hypothesis, (1.17) becomes

$$p = 1 - F_T(t_{obs}). \tag{1.18}$$

We assume $p$ to be a realization of the random variable $P = 1 - F_T(T)$ for $T = t_{obs}$. We want to show that the CDF of $P$, say $F_P(\cdot)$, follows a distribution $\mathcal{U}(0, 1)$. For a given

threshold $x$, we define $F_P(\cdot)$ as

$$\begin{aligned} F_P(x) &= Pr(P < x) \\ &= Pr(1 - F_T(T) < x) \\ &= Pr(F_T(T) > 1 - x) \end{aligned} \tag{1.19}$$

If $F_T(\cdot)$ is an increasing function, then

$$\begin{aligned} F_P(x) &= Pr(T > F_T^{-1}(1 - x)) \\ &= 1 - F_T(F_T^{-1}(1 - x)) \\ &= x. \end{aligned} \tag{1.20}$$

Hence, $F_P(x) = x \ \forall x \in (0, 1)$ and $P \sim \mathcal{U}(0, 1)$.

Very low values of the $p$-value constitute post-data evidence against $\mathcal{H}_0$. The lower the $p$-value, the less likely such data are to arise under $\mathcal{H}_0$. Figure 1.3 shows an example of comparison between the $\mathrm{P_{FA}}$ at a given threshold $\xi$ (yellow region) and the $p$-value, calculated in function of an observed value of the test statistic $t_{obs}$ (dashed green region). The larger $t_{obs}$ with respect to $\xi$, the higher the post-data confidence in rejecting $\mathcal{H}_0$. The blue curve represents the distribution of the data under the null hypothesis.

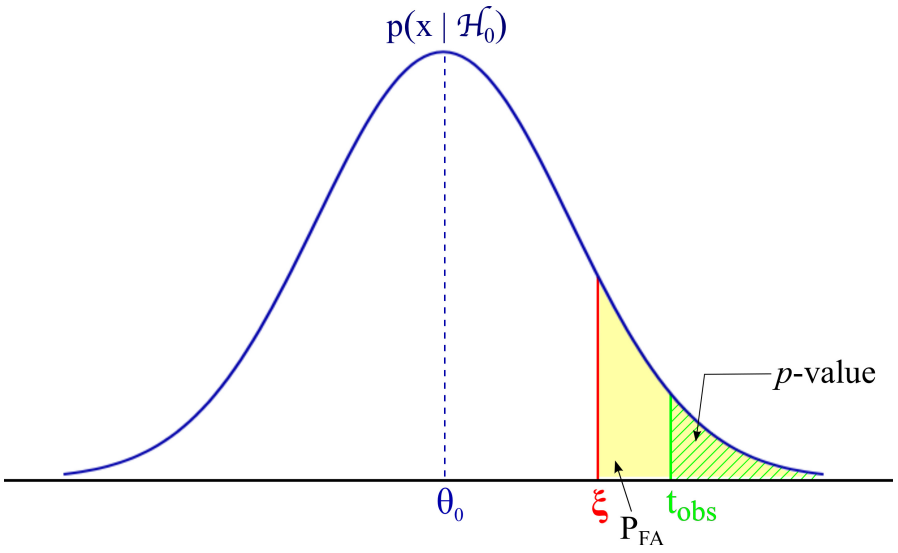



Figure 1.3: Illustration of the $\mathrm{P_{FA}}$ and the $p$-value for a one-sided, right-tailed test.

### 1.2.5 Performance of a test: ROC curves and AUC

The Receiver Operating Characteristic (ROC) is a curve that helps in summarizing the detection power of a test, by expressing the $\mathrm{P_{DET}}$ as a function of the $\mathrm{P_{FA}}$. In particular, each point of a ROC curve plots the $\mathrm{P_{DET}}$ versus the $\mathrm{P_{FA}}$ for a given value of the test threshold $\xi$. The analytical expressions of the $\mathrm{P_{FA}}$ and $\mathrm{P_{DET}}$, for the right-, left- and two-tailed tests of Example 1.2.1 are reported in the next example.

**Example 1.2.3. - Computation of the $\mathrm{P_{FA}}$ and $\mathrm{P_{DET}}$, for right-, left- and two-tailed tests**

Let us consider again the case of a realization $x$ of a scalar random variable $X$ that shows Gaussian PDF $p(x \mid \mathcal{H}_0) \sim \mathcal{N}(\theta_0, \sigma^2)$ under $\mathcal{H}_0$, and $p(x \mid \mathcal{H}_1) \sim \mathcal{N}(\theta_1, \sigma^2)$ under $\mathcal{H}_1$. In this example the test statistic $T(x) = x$. As previously defined in (1.13) and (1.12), the $\mathrm{P_{FA}}$ and $\mathrm{P_{DET}}$ of a one-sided, right-tailed test $T(x) \underset{H_0}{\overset{H_1}{\gtrless}} \xi$ write

$$\begin{aligned} \mathrm{P_{FA}} &= Pr\left(T(x) > \xi \mid \mathcal{H}_0\right) \\ &= \int_{\xi}^{+\infty} p(x \mid \mathcal{H}_0)\, \mathrm{d}x; \end{aligned} \tag{1.21}$$

$$\begin{aligned} \mathrm{P_{DET}} &= Pr\left(T(x) > \xi \mid \mathcal{H}_1\right) \\ &= \int_{\xi}^{+\infty} p(x \mid \mathcal{H}_1)\, \mathrm{d}x, \end{aligned} \tag{1.22}$$

or, equivalently,

$$\mathrm{P_{FA}} = 1 - \int_{-\infty}^{\xi} p(x \mid \mathcal{H}_0)\, \mathrm{d}x; \tag{1.23}$$

$$\mathrm{P_{DET}} = 1 - \int_{-\infty}^{\xi} p(x \mid \mathcal{H}_1)\, \mathrm{d}x. \tag{1.24}$$

Considering the definition of the CDF of a standard normal distribution

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, \mathrm{d}t, \tag{1.25}$$

equations (1.23) and (1.24) become

$$\mathrm{P_{FA}} = 1 - \Phi\left(\frac{\xi - \theta_0}{\sigma}\right); \tag{1.26}$$

$$\mathrm{P_{DET}} = 1 - \Phi\left(\frac{\xi - \theta_1}{\sigma}\right). \tag{1.27}$$

Alternatively, for a one-sided, left-tailed test $T(x) \underset{H_1}{\overset{H_0}{\gtrless}} \xi$ we have

$$\mathrm{P_{FA}} = Pr\left(T(x) < \xi \mid \mathcal{H}_0\right) \tag{1.28}$$

$$= \int_{-\infty}^{\xi} p(x \mid \mathcal{H}_0)\, \mathrm{d}x = \Phi\left(\frac{\xi - \theta_0}{\sigma}\right); \tag{1.29}$$

$$\mathrm{P_{DET}} = Pr\left(T(x) < \xi \mid \mathcal{H}_1\right) \tag{1.30}$$

$$= \int_{-\infty}^{\xi} p(x \mid \mathcal{H}_1)\, \mathrm{d}x = \Phi\left(\frac{\xi - \theta_1}{\sigma}\right). \tag{1.31}$$

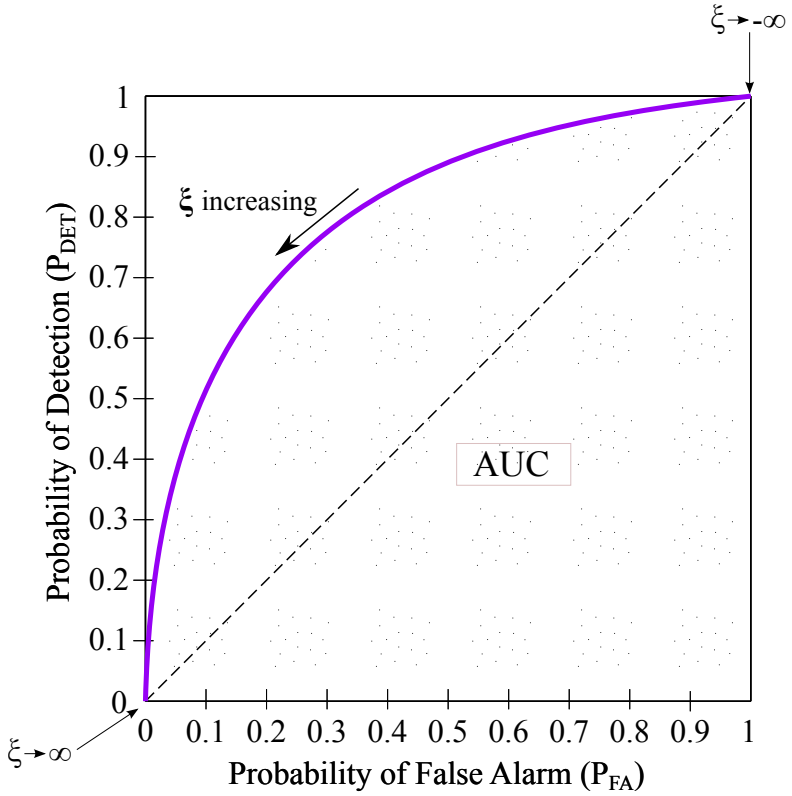



Figure 1.4: ROC curve: $\mathrm{P_{DET}}$ versus $\mathrm{P_{FA}}$ for an increasing value of the test threshold $\xi$. Two limit cases are shown: $\xi \to \infty$ for which $\mathrm{P_{DET}} = \mathrm{P_{FA}} = 0$, and $\xi \to -\infty$ for which $\mathrm{P_{DET}} = \mathrm{P_{FA}} = 1$.

In the case of a two-tailed test $|T(x)| \underset{H_0}{\overset{H_1}{\gtrless}} \xi$ finally, the $\mathrm{P_{FA}}$ and $\mathrm{P_{DET}}$ are defined as

$$\mathrm{P_{FA}} = Pr(|T(x)| > \xi \mid \mathcal{H}_0); \tag{1.32}$$

$$\mathrm{P_{DET}} = Pr(|T(x)| > \xi \mid \mathcal{H}_1). \tag{1.33}$$

These two expressions lead to

$$\mathrm{P_{FA}} = 2 - \Phi\left(\frac{\xi - \theta_0}{\sigma}\right) - \Phi\left(\frac{\xi + \theta_0}{\sigma}\right) \tag{1.34}$$

$$\mathrm{P_{DET}} = 2 - \Phi\left(\frac{\xi - \theta_1}{\sigma}\right) - \Phi\left(\frac{\xi + \theta_1}{\sigma}\right). \tag{1.35}$$

An example of ROC curve, for a one-sided right-tailed test, is given in Figure 1.4. ■

We see that when the threshold $\xi$ increases both the $\mathrm{P_{DET}}$ and the $\mathrm{P_{FA}}$ decrease, and vice-versa. Note that in the limit case of $\xi \to \infty$, $\mathrm{P_{DET}} = \mathrm{P_{FA}} = 0$, while when $\xi \to -\infty$, $\mathrm{P_{DET}} = \mathrm{P_{FA}} = 1$.

A ROC curve should always be above the 45° line. This line, commonly called the *random line*, depicts the detection behavior of a test that randomly decides which one of

the model hypotheses accept, independently to all other observations (dashed line in Figure 1.4). For all the points on this line, we have that $P_{DET} = P_{FA}$ (same probability to choose between $\mathcal{H}_0$ and $\mathcal{H}_1$).

Finally, another means that allows to assess the performance of a test is given by the computation of the Area Under the Curve (AUC, dotted region in Figure 1.4). As its name indicates, the AUC of a test simply corresponds to the value of the area under the associated ROC curve, and takes consistent values between $0.5 < \text{AUC} \leq 1$: the larger the AUC, the more powerful the test in a global sense (AUC = 1 corresponds to a perfect test). Note that AUC = 0.5 corresponds to the performance of a useless test, with random decision rule and ROC curve on the 45° line. The AUC is thus a measure of the global (over all $P_{FA}$) power of a test.

## 1.3 Major approaches to hypothesis testing

To conclude our general overview on statistical hypothesis testing, we discuss some simple but widely known hypothesis tests approaches. The main purpose of the following sections is to give a glimpse on largely used hypotheses tests, and highlight the differences between frequentist methods based on likelihood functions and a Bayesian alternative.

Given an observed data vector $\mathbf{x} = [x_1, \ldots, x_N]^t$ in $\mathbb{R}^N$, the model we consider is

$$\begin{cases} \mathcal{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0, \\ \mathcal{H}_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_1, \end{cases} \tag{1.36}$$

where the $\mathcal{H}_0$ and $\mathcal{H}_1$ hypotheses are described in terms of the parameter vector $\boldsymbol{\theta}$ (with $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_0$).

In the frequentist context, the first test we introduce is the Likelihood Ratio (LR) test. The second test we consider is a straightforward generalization of the LR test, termed the Generalized Likelihood Ratio (GLR) test. We finally analyze a Bayesian alternative, introducing the so called Bayes Factor (BF) test.

### 1.3.1 The Likelihood Ratio test

#### 1.3.1.1 The Likelihood function

In statistics, we define the *Likelihood function* (or simply, the Likelihood) a conditional probability function measuring the fit between the observed data and the parameters of the considered statistical model. Mathematically, writing $\mathbf{x} = [x_1, \ldots, x_N]^t$ for a given a set of observations and $\boldsymbol{\theta}$ for a known parameter vector (on which the probability density functions of the observations depend), the Likelihood function is given by the joint probability function

$$p(\mathbf{x} \mid \boldsymbol{\theta}) = Pr(X_1 = x_1, X_2 = x_2, \ldots, X_3 = x_N \mid \boldsymbol{\theta}). \tag{1.37}$$

If the $X_i$ are independently identically distributed (i.i.d.), then the likelihood function simplifies to

$$p(\mathbf{x} \mid \boldsymbol{\theta}) = \prod_{i=1}^{N} p(x_i \mid \boldsymbol{\theta}). \tag{1.38}$$

Of course, the larger the value of the Likelihood, the better the data are in agreement with the model's parameters. This is the reason why many methods rely on the maximization of the Likelihood function. The notion of Likelihood function plays a key role in the statistical signal processing context, and is in particular extensively used to construct the most varied forms of hypothesis tests. The simplest case of test that takes benefit from the use of the Likelihood function is the so called *Likelihood Ratio test* (or Neyman-Pearson test, NP), which we describe below.

#### 1.3.1.2 The Likelihood Ratio test

The LR test is a statistical test expressing the data goodness-of-fit to two simple statistical models like the ones expressed in (1.36). In the LR test the parameter vector $\boldsymbol{\theta}$ has to be known under the two hypotheses. According to the hypotheses in (1.36), the ratio between the likelihood function of the data under $\mathcal{H}_1$, and the likelihood function of the data under $\mathcal{H}_0$, is defined as

$$\Lambda(\mathbf{x}) := \frac{p(\mathbf{x} \mid \boldsymbol{\theta}_1)}{p(\mathbf{x} \mid \boldsymbol{\theta}_0)}. \tag{1.39}$$

The function $\Lambda(\mathbf{x})$ is commonly called the *Likelihood Ratio*. After choosing a test threshold $\xi$, fixed in accordance to the desired significance level of the test (see subsection 1.2.3), we obtain

$$\text{LR}: \ \Lambda(\mathbf{x}) \underset{H_0}{\overset{H_1}{\gtrless}} \xi, \tag{1.40}$$

which is the general expression of the LR test. The LR test rejects the null hypothesis if the test statistic $\Lambda(\mathbf{x})$ falls above the test threshold $\xi$.

The following two examples explicit the computation of the LR test statistic in both the scalar and the vector cases, when considering Gaussian distributed random variables.

**Example 1.3.1. - Computation of the LR test statistic in the scalar case**

Let us consider the case of a realization $x$ of a scalar Gaussian distributed random variable $X$. In particular, we assume that $p(x \mid \mathcal{H}_0) \sim \mathcal{N}(\theta_0, \sigma^2)$ under $\mathcal{H}_0$, and $p(x \mid \mathcal{H}_1) \sim \mathcal{N}(\theta_1, \sigma^2)$ under $\mathcal{H}_1$, where $\sigma^2$, $\theta_0$ and $\theta_1$ are known, $\theta_0 = 0$ and $\theta_1 \neq 0$. According to (1.39) and (1.40) we have

$$\Lambda(x) = \frac{\exp\left[-\frac{1}{2\sigma^2}\left(x-\theta_1\right)^2\right]}{\exp\left(-\frac{1}{2\sigma^2}x^2\right)} \underset{H_0}{\overset{H_1}{\gtrless}} \xi. \tag{1.41}$$

Taking the log-version of (1.41) gives

$$\Lambda(x) = -\frac{\theta_1^2}{2\sigma^2} + \frac{x\theta_1}{\sigma^2} \underset{H_0}{\overset{H_1}{\gtrless}} \ln \xi, \tag{1.42}$$

which finally resumes to the test

$$\Lambda(x) = x \underset{H_0}{\overset{H_1}{\gtrless}} \xi_{\text{LR}}, \tag{1.43}$$

where $\xi_{LR} = \frac{\sigma^2}{\theta_1}\left(\ln \xi + \frac{\theta_1^2}{2\sigma^2}\right)$. Hence, the LR test amounts to compare $x$ to a threshold, which controls the $\text{P}_{\text{FA}}$. An illustration of the test (1.43), which corresponds to the case considered in Example 1.2.2, has been shown in Figure 1.2. ■

The computation of the LR test statistic in the $N-$dimensional case derives directly from expressions (1.38), (1.39) and (1.40).

**Example 1.3.2. - Computation of the LR test statistic in the vector case**

We assume that the considered data vector $\mathbf{x}$ shows Gaussian PDF $p(\mathbf{x} \mid \mathcal{H}_0) \sim \mathcal{N}(\boldsymbol{\theta}_0, \mathbf{C})$ under $\mathcal{H}_0$, and $p(\mathbf{x} \mid \mathcal{H}_1) \sim \mathcal{N}(\boldsymbol{\theta}_1, \mathbf{C})$ under $\mathcal{H}_1$. Similarly to the scalar case, the covariance matrix $\mathbf{C}$ and the mean vectors $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ are known, $\boldsymbol{\theta}_0 = \mathbf{0}$ and $\boldsymbol{\theta}_1 \neq \mathbf{0}$. For data that are i.i.d, the LR test (1.40) writes

$$\begin{aligned}
\Lambda(\mathbf{x}) &= \frac{\frac{\det\left(\mathbf{C}^{-1/2}\right)}{(2\pi)^{N/2}} \exp\left[-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\theta}_1\right)^t \mathbf{C}^{-1}\left(\mathbf{x}-\boldsymbol{\theta}_1\right)\right]}{\frac{\det\left(\mathbf{C}^{-1/2}\right)}{(2\pi)^{N/2}} \exp\left(-\frac{1}{2}\mathbf{x}^t\mathbf{C}^{-1}\mathbf{x}\right)} \\
&= \exp\left[-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\theta}_1\right)^t \mathbf{C}^{-1}\left(\mathbf{x}-\boldsymbol{\theta}_1\right) + \frac{1}{2}\mathbf{x}^t\mathbf{C}^{-1}\mathbf{x}\right] \\
&= \exp\left[\boldsymbol{\theta}_1^t\mathbf{C}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\theta}^t\mathbf{C}^{-1}\boldsymbol{\theta}\right]. \qquad (1.44)
\end{aligned}$$

Taking the log-version of (1.44) we finally have

$$\Lambda(\mathbf{x}) = \boldsymbol{\theta}_1^t\mathbf{C}^{-1}\mathbf{x} \underset{H_0}{\overset{H_1}{\gtrless}} \xi_{\mathrm{LR}} \qquad (1.45)$$

with $\xi_{\mathrm{LR}} = \ln\xi + \frac{1}{2}\boldsymbol{\theta}^t\mathbf{C}^{-1}\boldsymbol{\theta}$. Test (1.45) corresponds to a matched filter in which data weighted by $\mathbf{C}^{-1}$ are correlated to the signal under $\mathcal{H}_1$. ■

**Example 1.3.3. - LR test statistic in the vector case: a numerical example**

We consider the case of a $N = 2$ dimensional Gaussian data vector $\mathbf{x}$. For the mean vectors $\boldsymbol{\theta}_0 = [\theta_{0_1},\ \theta_{0_2}]^t = [0,\ 0]^t$ under $\mathcal{H}_0$ and $\boldsymbol{\theta}_1 = [\theta_{1_1},\ \theta_{1_2}]^t = [2,\ 2]$ under $\mathcal{H}_1$, and for a covariance matrix $\mathbf{C} = \mathbf{I}$, the LR test (1.45) writes

$$\begin{aligned}
\Lambda(\mathbf{x}) &= \theta_{1_1}x_1 + \theta_{1_2}x_2 \underset{H_0}{\overset{H_1}{\gtrless}} \xi_{\mathrm{LR}} \\
&= 2x_1 + 2x_2 \underset{H_0}{\overset{H_1}{\gtrless}} \xi_{\mathrm{LR}}. \qquad (1.46)
\end{aligned}$$

This example is illustrated in Figure 1.5. The data distribution under $\mathcal{H}_0$ is plotted in blue, while the one under $\mathcal{H}_1$ in green. The circles represent the iso-values of the PDF under both hypotheses. The test detection edge (surface, for N>2) represented by the line of equation $2x_1 + 2x_2 = \xi_{\mathrm{LR}}$ ($\xi_{\mathrm{LR}} = 4$) and separating the two decision regions (retain $\mathcal{H}_0$ vs. retain $\mathcal{H}_1$), is plotted in dashed red. In general ($N > 2$), $\boldsymbol{\theta}_1^t\mathbf{x} = \xi$ is a hyperplane, which represents the *detection surface* of the LR test and $\frac{\xi}{||\boldsymbol{\theta}_1||}$ is its distance to $\boldsymbol{\theta}_0$. A study of the detection surfaces of the detection approaches proposed in chapter 3 will be held in chapter 4. ■

#### 1.3.1.3 The LR test as optimal detector: the Neyman-Pearson lemma

Named after Jerzy Neyman and Egon Sharpe Pearson, who introduced it in 1933, the Neyman-Pearson (NP) lemma, for the LR function described in (1.39), states that (the Proof is given in Appendix A.1)

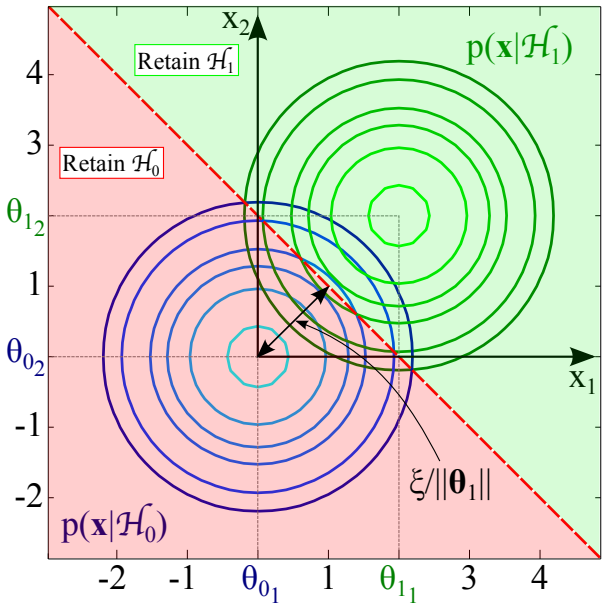



Figure 1.5: Data distributions under the hypotheses of the model and decision regions (retain $\mathcal{H}_0$, in red; retain $\mathcal{H}_1$, in green) for a one-sided right-tailed test in the two-dimensional case. Here, $\boldsymbol{\theta}_0 = [\theta_{0_1}\ \theta_{0_2}]^t = [0\ 0]^t$, $\boldsymbol{\theta}_1 = [\theta_{1_1}\ \theta_{1_2}]^t = [2\ 2]$ and $\xi_{\text{LR}} = 4$.

**Lemma 1.3.1.** *In order to maximize the $P_{DET}$ for a fixed $P_{FA} = \alpha$, we reject $\mathcal{H}_0$ if*

$$\Lambda(\mathbf{x}) > \xi, \tag{1.47}$$

*where*

$$P_{FA} = Pr(\Lambda(\mathbf{x}) > \xi \mid \mathcal{H}_0) = \alpha. \tag{1.48}$$

For model (1.36), in the case of simple hypotheses ($\boldsymbol{\theta}$ value known under $\mathcal{H}_0$ and $\mathcal{H}_1$), this means that, with respect to a significance level $\text{P}_{\text{FA}}(\xi) = \alpha$, the LR test represents the test with greatest detection power compared to any other test for that level of $\text{P}_{\text{FA}}$.

### 1.3.2 The Generalized Likelihood Ratio test

The Neyman-Pearson approach requires that the values taken by the $\boldsymbol{\theta}$ parameter under the two hypotheses of the considered model are known. When the parameter values are not available, an intuitive, well known and widely-applicable decision strategy is the Generalised Likelihood Ratio test. Like the LR test approach, the GLR test also computes the ratio of the likelihoods under the model hypotheses, with the only difference that it involves the Maximum Likelihood (ML) estimate of the parameter vector $\boldsymbol{\theta}$ ($\hat{\boldsymbol{\theta}}_{\text{ML}}$): the definition of the GLR test is

$$\text{GLR}(\mathbf{x}) : \frac{\sup\limits_{\boldsymbol{\theta}\in\Theta_1} p(\mathbf{x} \mid \boldsymbol{\theta})}{\sup\limits_{\boldsymbol{\theta}\in\Theta_0} p(\mathbf{x} \mid \boldsymbol{\theta})} \underset{H_0}{\overset{H_1}{\gtrless}} \xi. \tag{1.49}$$

An example of computation of the ML estimate for an unknown vector of parameters $\boldsymbol{\theta}$ is firstly shown. The definition of the GLR test together with the computation of the expression of its test statistic will follow immediately after.

**Example 1.3.4. - An example of computation of the GLR test statistic**

We show the ML estimate $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ of the unknown vector of parameter $\boldsymbol{\theta}$ in the case of the hypotheses model

$$\begin{cases} \mathcal{H}_0 : \boldsymbol{\theta} = \mathbf{0}, \\ \mathcal{H}_1 : \boldsymbol{\theta} \neq \mathbf{0}, \end{cases} \tag{1.50}$$

and for normally distributed data $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\theta}, \mathbf{C})$, where $\mathbf{C}$ represents the covariance matrix.

By maximizing the likelihood function $p(\mathbf{x} \mid \boldsymbol{\theta})$, with respect to $\boldsymbol{\theta}$ we obtain

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{\mathrm{ML}} &= \arg\max_{\boldsymbol{\theta}} p(\mathbf{x} \mid \boldsymbol{\theta}) \\ &= \arg\max_{\boldsymbol{\theta}} \frac{\det\left(\mathbf{C}^{-1/2}\right)}{(2\pi)^{N/2}} \exp\left[\sum_{i=1}^{N} -\frac{1}{2\sigma_i^2}(x_i - \theta_i)^2\right]. \end{aligned} \tag{1.51}$$

Taking the logarithm of (1.51), computing the derivative with respect to the $\boldsymbol{\theta}$ components and equating to 0, we obtain

$$\left.\frac{\partial \ln p(\mathbf{x} \mid \boldsymbol{\theta})}{\partial \theta_i}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{\mathrm{ML}}} = \frac{x_i}{\sigma_i^2} - \frac{\hat{\theta}_{ML_i}}{\sigma_i^2} = 0, \qquad \text{for } i = 1, \ldots, N \tag{1.52}$$

from which is easy to see that

$$\hat{\theta}_{ML_i} = x_i, \qquad \text{for } i = 1, \ldots, N \tag{1.53}$$

or, equivalently,

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \mathbf{x}. \tag{1.54}$$

For model (1.50), the definition of the GLR test (1.49) leads to

$$\mathrm{GLR}(\mathbf{x}) : \frac{\max_{\boldsymbol{\theta}} p(\mathbf{x} \mid \boldsymbol{\theta})}{p(\mathbf{x} \mid \mathbf{0})} \underset{H_0}{\overset{H_1}{\gtrless}} \xi. \tag{1.55}$$

Consequently, the computation of the $T_{\mathrm{GLR}}$ test statistic gives

$$\begin{aligned} T_{\mathrm{GLR}}(\mathbf{x}) &= \frac{\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2\sigma_i^2}\left(x_i - \hat{\theta}_{ML_i}\right)^2\right]}{\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{1}{2\sigma_i^2} x_i^2\right)} \\ &= \prod_{i=1}^{N} \exp\left[-\frac{1}{2\sigma_i^2}\left(x_i - \hat{\theta}_{ML_i}\right)^2 + \frac{1}{2\sigma_i^2} x_i^2\right] \\ &= \exp\left[\sum_{i=1}^{N} -\frac{1}{2\sigma_i^2}\left(\hat{\theta}_{ML_i}^2 - 2x_i\hat{\theta}_{ML_i}\right)\right] \end{aligned} \tag{1.56}$$

Inserting in the log-version of (1.56) the value of the $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ estimate computed in (1.53), and denoting $x_i/\sigma_i = u_i$, we finally have that

$$\begin{aligned} T_{\mathrm{GLR}}(\mathbf{u}) &= \sum_{i=1}^{N} u_i^2 \underset{H_0}{\overset{H_1}{\gtrless}} \xi_{\mathrm{GLR}} \\ &= ||\mathbf{u}||_2^2 \underset{H_0}{\overset{H_1}{\gtrless}} \xi_{\mathrm{GLR}} \ , \end{aligned} \tag{1.57}$$

where $\xi_{\mathrm{GLR}} = 2\ln\xi$.

The computation of the GLR test statistic in (1.57) highlights the fact that this test acts as an *Energy Detector*. An Energy Detector computes the energy of the signal $\mathbf{u}$ (sum of the square values of its components) and compares it to an appropriate threshold. Note that the data components in the GLR test statistic are all weighted by $\sigma_i^2$ reflecting that all components do not present the same Signal-to-Noise Ratio (SNR). This test will be taken into account all along the manuscript as a reference for comparison purposes.

Note that in the scalar case ($\mathcal{H}_0 : \theta = 0$, $\mathcal{H}_1 : \theta \neq 0$), the GLR test (1.57) resumes to

$$\begin{aligned} T_{\mathrm{GLR}}(u) &= u^2 \underset{H_0}{\overset{H_1}{\gtrless}} \xi_{\mathrm{GLR}}, \\ &= \frac{x^2}{\sigma^2} \underset{H_0}{\overset{H_1}{\gtrless}} \xi_{\mathrm{GLR}}, \\ &= |x| \underset{H_0}{\overset{H_1}{\gtrless}} \xi'_{\mathrm{GLR}}, \end{aligned} \tag{1.58}$$

with $\xi'_{\mathrm{GLR}} = \sqrt{\sigma^2 \xi_{\mathrm{GLR}}}$. ■

The scalar (1.58) and the vector (1.57) GLR tests are illustrated in Figure 1.6 and Figure 1.7, for the same parameter values used for the LR test in subsubsection 1.3.1.2 (see Figure 1.5 and Figure 1.2) and for $\xi = 3$. We note that in the scalar case (1.58) the absolute value on the data $x$ leads to a two-sided detection test, because of the unknown sign of the parameter $\theta$ under $\mathcal{H}_1$. For a two-sided tests, we have reported the $\mathrm{P_{FA}}$ and $\mathrm{P_{DET}}$ expressions in (1.32)−(1.35).

On the other hand, in the $N = 2$ vector case of Figure 1.7 the GLR detection edge separating the two decision regions, is represented by a circle centered in the origin and with radius $r = \sqrt{\xi_{\mathrm{GLR}}}$.

Finally, a comparison between the LR and GLR tests detection performances in terms of ROC curves is shown in Figure 1.8. According to model (1.50), we have considered a $N = 2$ dimensional data vector $\mathbf{x}$, characterized by a parameter vector $\boldsymbol{\theta} = [2,\ 0]$ under $\mathcal{H}_1$. As expected, the LR test (in blue) outperforms the GLR (in green), because while the LR test has perfect knowledge of the true value of $\boldsymbol{\theta}$ under $\mathcal{H}_1$, the GLR estimates it by ML estimate leading to $\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \mathbf{x}$. When computing $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}^t \mathbf{x}$, which can be interpreted as an approximation of a matched filter, the parameter estimate is thus inaccurate in correlating with the data. This situation is obviously frequent in practice.

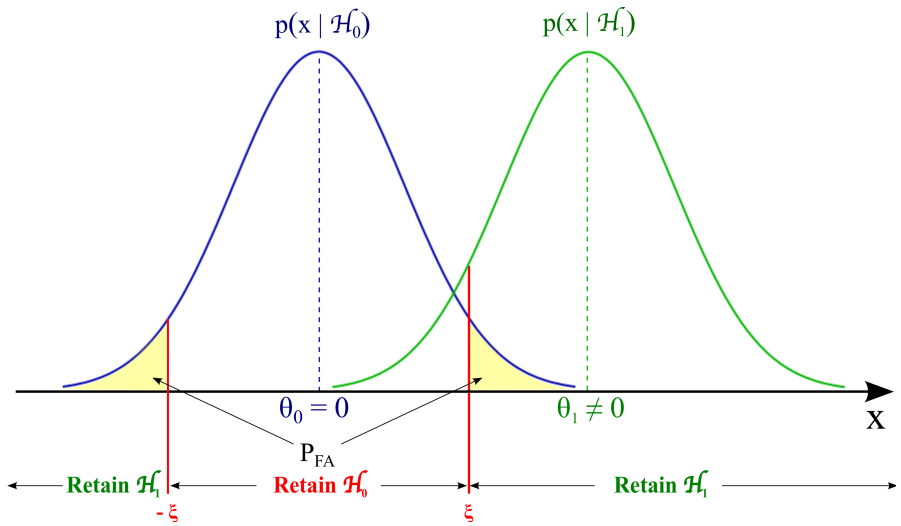



Figure 1.6: Scalar GLR test for model $\mathcal{H}_0 : \theta_0 = 0$ and $\mathcal{H}_1 : \theta_1 \neq 0$: distributions of the data under the hypotheses of the model and decision regions. The GLR test represents a two-sided test.

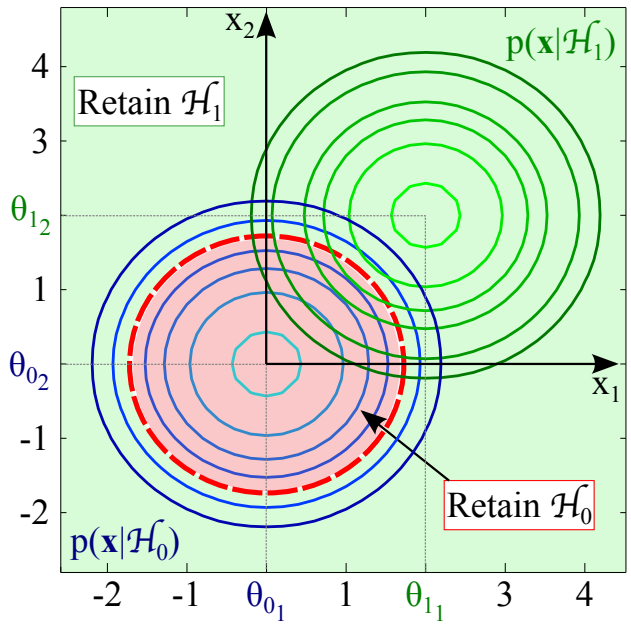



Figure 1.7: Vector GLR test for model $\mathcal{H}_0 : \boldsymbol{\theta}_0 = \mathbf{0}$ and $\mathcal{H}_1 : \boldsymbol{\theta}_1 \neq \mathbf{0}$: distributions of the data under the hypotheses of the model and decision regions (retain $\mathcal{H}_0$, in red; retain $\mathcal{H}_1$, in green) for $\boldsymbol{\theta}_0 = [\theta_{0_1},\ \theta_{0_2}]^t = [0,\ 0]^t$, $\boldsymbol{\theta}_1 = [\theta_{1_1},\ \theta_{1_2}]^t = [2,\ 2]$ and $\xi_{\text{GLR}} = 3$. The detection surface of the test is shown in dashed red.

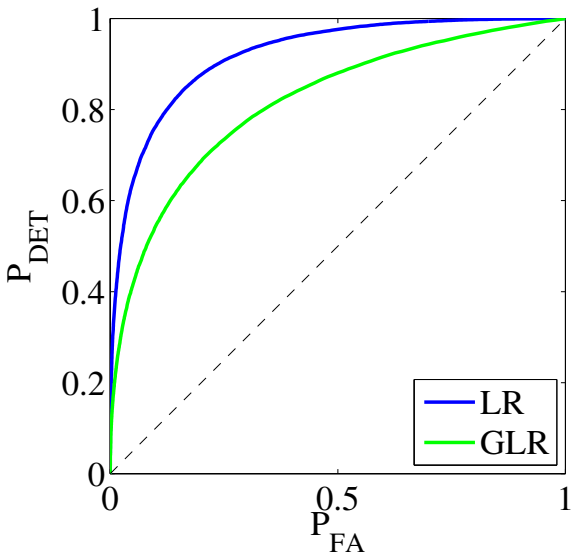



Figure 1.8: Comparison of the LR and GLR ROC curves for a $N = 2$ parameter vector $\boldsymbol{\theta} = [2,\ 0]$ under $\mathcal{H}_1$.

### 1.3.3 Bayesian approach to hypothesis testing and the Bayes Factor

#### 1.3.3.1 The Bayes' theorem and the Bayes' Factor

In the Bayesian approach to statistical inference, the unknown model parameters are considered random variables. The basis to this approach is given by the *Bayes' theorem*, called after Thomas Bayes who first introduced it. For a data vector $\mathbf{x}$ and a vector of parameters $\boldsymbol{\theta}$, the Bayes' theorem states that

$$p\left(\mathbf{x}, \boldsymbol{\theta}\right) = p\left(\boldsymbol{\theta} \mid \mathbf{x}\right) \pi(\mathbf{x}) = p\left(\mathbf{x} \mid \boldsymbol{\theta}\right) \pi\left(\boldsymbol{\theta}\right), \tag{1.59}$$

where the joint probability $p\left(\mathbf{x}, \boldsymbol{\theta}\right)$ is obtained as the product of the *posterior probability* $p\left(\boldsymbol{\theta} \mid \mathbf{x}\right)$, which gives the probability of the parameters $\boldsymbol{\theta}$ given the data $\mathbf{x}$, times the *prior probability* on the parameters $\pi(\boldsymbol{\theta})$ or, equivalently, as the product of the likelihood $p\left(\mathbf{x} \mid \boldsymbol{\theta}\right)$ times the prior probability on the data $\pi(\mathbf{x})$. From (1.59) we have,

$$p\left(\boldsymbol{\theta} \mid \mathbf{x}\right) = \frac{p\left(\mathbf{x}, \boldsymbol{\theta}\right)}{\pi(\mathbf{x})} = \frac{p\left(\mathbf{x} \mid \boldsymbol{\theta}\right) \pi\left(\boldsymbol{\theta}\right)}{\pi(\mathbf{x})}. \tag{1.60}$$

Because of its independence from the parameters $\boldsymbol{\theta}$, $\pi(\mathbf{x})$, which acts like a normalizing constant ensuring the posterior probability to be between 0 and 1, is often omitted. However, if considered in the context of Bayesian hypothesis testing, this term is of crucial importance. Under the two hypotheses of a given model, say $\mathcal{H}_i$ with $i = 0, 1$, $\pi(\mathbf{x})$ can be obtained as

$$\pi(\mathbf{x} \mid \mathcal{H}_i) = p(\mathbf{x} \mid \mathcal{H}_i) = \int p\left(\mathbf{x} \mid \boldsymbol{\theta}\right) \pi_i\left(\boldsymbol{\theta}\right) d\boldsymbol{\theta}. \tag{1.61}$$

This expression computes the *marginal probability* (since we marginalize over $\boldsymbol{\theta}$) of the data and can be interpreted as a weighted average likelihood, where the weights are given by the prior distribution $\pi_i(\boldsymbol{\theta})$.

This quantity is the basis for the construction of a simple and widely used Bayesian alternative to frequentist hypothesis tests, called the Bayes' Factor (BF). According to the likelihood $p(\mathbf{x} \mid \boldsymbol{\theta})$ and to the hypotheses of the following model

$$\begin{cases} \mathcal{H}_0 : \boldsymbol{\theta} \sim \pi_0(\boldsymbol{\theta}) \\ \mathcal{H}_1 : \boldsymbol{\theta} \sim \pi_1(\boldsymbol{\theta}), \end{cases} \tag{1.62}$$

where $\pi_0(\boldsymbol{\theta})$ and $\pi_1(\boldsymbol{\theta})$ are the prior probabilities on the random parameter under $\mathcal{H}_0$ and $\mathcal{H}_1$ respectively, the BF is defined as

$$\begin{aligned} \mathrm{BF}(\mathbf{x}) &= \frac{p(\mathbf{x} \mid \mathcal{H}_1)}{p(\mathbf{x} \mid \mathcal{H}_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \xi \\ &= \frac{\int p(\mathbf{x} \mid \boldsymbol{\theta}) \, \pi_1(\boldsymbol{\theta}) \, d\boldsymbol{\theta}}{\int p(\mathbf{x} \mid \boldsymbol{\theta}) \, \pi_0(\boldsymbol{\theta}) \, d\boldsymbol{\theta}} \underset{H_0}{\overset{H_1}{\gtrless}} \xi. \end{aligned} \tag{1.63}$$

Introduced by Harold Jeffreys in the middle of the last century [Jeffreys 1935], the BF computes the marginal probabilities of the data under the two hypotheses of the considered model giving a way to say which of two competing models better provides an accurate explanation of the data. The uncertainty about the parameter is thus encapsulated in the prior laws $\pi_i(\boldsymbol{\theta})$.

**Example 1.3.5. - Computation of the BF test statistic**

We consider the statistical model $\mathbf{x} = \boldsymbol{\theta} + \mathbf{n}$ with the following hypotheses

$$\begin{cases} \mathcal{H}_0 : \boldsymbol{\theta} = \mathbf{0} \\ \mathcal{H}_1 : \boldsymbol{\theta} \neq \mathbf{0}, \quad \text{and} \quad \boldsymbol{\theta} \sim \pi_1(\boldsymbol{\theta}). \end{cases} \tag{1.64}$$

In this case, definition (1.63) becomes

$$\mathrm{BF}(\mathbf{x}) = \frac{\int p(\mathbf{x} \mid \boldsymbol{\theta}) \, \pi_1(\boldsymbol{\theta}) \, d\boldsymbol{\theta}}{p(\mathbf{x} \mid \mathbf{0})} \underset{H_0}{\overset{H_1}{\gtrless}} \xi. \tag{1.65}$$

For a i.i.d noise $\mathbf{n} \in \mathbb{R}^N$, a data vector $\mathbf{x} \in \mathbb{R}^N$ and a separable prior $\pi_1(\boldsymbol{\theta}) = \prod_{i=1}^{N} \pi_{1_i}(\theta_i)$, expression (1.65) leads to the test statistic

$$T_{\mathrm{BF}}(\mathbf{x}) = \prod_{i=1}^{N} \frac{\int_{\mathbb{R}} p(x_i \mid \theta_i) \, \pi_{1_i}(\theta_i) \, d\theta_i}{p(x_i \mid 0)}, \tag{1.66}$$

where $T_{\mathrm{BF}}(\mathbf{x})$ is simply obtained by the product of the BF test expression for each scalar component $x_i$ in $\mathbf{x}$. As the GLR test, the BF will be taken into account in next chapters for comparison purposes. ■

#### 1.3.3.2 The minimum $\mathrm{P_E}$ detector

From the Bayesian framework considered in the sections above, the *Probability of error* $\mathrm{P_E}$ is defined as

$$\begin{aligned} \mathrm{P_E} &= Pr\left\{(\text{“accept } \mathcal{H}_1\text{”}, \text{“}\mathcal{H}_0 \text{ is true”}) \cup (\text{“accept } \mathcal{H}_0\text{”}, \text{“}\mathcal{H}_1 \text{ is true”})\right\} \\ &= Pr\left(\mathcal{H}_1, \mathcal{H}_0\right) + Pr\left(\mathcal{H}_0, \mathcal{H}_1\right) \\ &= Pr\left(\mathcal{H}_1 \mid \mathcal{H}_0\right) Pr\left(\mathcal{H}_0\right) + Pr\left(\mathcal{H}_0 \mid \mathcal{H}_1\right) Pr\left(\mathcal{H}_1\right) \\ &= \mathrm{P_{FA}} Pr\left(\mathcal{H}_0\right) + \mathrm{P_M} Pr\left(\mathcal{H}_1\right), \end{aligned} \tag{1.67}$$

where $Pr\left(\mathcal{H}_0\right)$ and $Pr\left(\mathcal{H}_1\right)$ are probabilities assigned on the hypotheses, which express a prior belief in the likelihood of the hypotheses. According to this, we have the following result: the detector that minimizes the $\mathrm{P_E}$ is the one that decides $\mathcal{H}_1$ if

$$\frac{p(\mathbf{x} \mid \mathcal{H}_1)}{p(\mathbf{x} \mid \mathcal{H}_0)} > \frac{p(\mathcal{H}_1)}{p(\mathcal{H}_0)} = \xi. \tag{1.68}$$

This detector corresponds to the BF (1.63). Thus, similarly to the LR test of Lemma 1.3.1, the BF (or *conditional* likelihood ratio [Kay 1998a]) is compared to a threshold, computed this time as the ratio of the prior probabilities of the model's hypotheses. The Proof of this result is reported in Appendix A.2.

## 1.4 Conclusion

In this chapter we have firstly introduced the basic steps that need to be followed to design a hypothesis test. We have seen that these main steps are in a number of four. In fact, to correctly build and perform a test we have to: state the hypotheses of the test, identify an appropriate test statistic, specify the significance level of the test by selecting a test threshold, and finally make a decision. These steps will indeed be considered for the design of the detection tests introduced in the next chapters. We presented some statistical figures helping to evaluate a decision rule, such as the decision/error probabilities (for instance, Figure 1.2 and Figure 1.5), the $p$-value (Figure 1.3), the ROC and the AUC (Figure 1.4).

Secondly, we have given some examples on the principal methods used in hypothesis testing highlighting the differences existing between frequentist methods, such as the LR and the GLR tests, and Bayesian methods, such as the BF test. While in the former two cases the considered vector of parameters is deterministic, in the latter it assumes random values. The GLR and the BF tests will be taken into account as reference tests for purposes of comparison.

The 1D and 2D analysis of the LR and the GLR tests' detection surfaces illustrated in Figure 1.5 and Figure 1.7 and in various examples, will help in chapter 4 to understand the reasoning at the basis of the study of the detection surfaces of the approaches proposed in chapter 3.

Chapter 2

# Sparse representations, estimation and detection

**Contents**


**2.1 Introduction . . . . . 33**
**2.2 Sparse representations of signals . . . . . 34**
2.2.1 Sparsity of signals in a dictionary of atoms . . . . . 34
2.2.2 The choice of the dictionary . . . . . 35
**2.3 Sparse estimation . . . . . 36**
2.3.1 Penalty functions and thresholding estimators . . . . . 37
2.3.2 Bayesian framework for sparsity-promoting estimators . . . . . 41
2.3.3 MAP sparsity-promoting estimate of unknown parameters . . . . . 44
**2.4 Sparse detection . . . . . 47**
2.4.1 Previous works in sparse detection . . . . . 47
2.4.2 The Higher Criticism . . . . . 48
2.4.3 A glance at sparse detection in Hyperspectral Imaging . . . . . 51
**2.5 Conclusion . . . . . 54**


## 2.1 Introduction

Signal processing is a research area that proposes to conceive, study and design systems for the exploitation of signals. Generally speaking, a signal describes an evolving phenomenon (e.g. over time) and usually carries a specific set of information.

Over the last decades, a new way of representing signals has experienced a growing trend in the community of signal and image processing: the *sparse representations*. Since the advent of the need of new strategies to store and analyze rising amounts of data, methodologic and algorithmic advances have been proposed and this research theme still keeps the interest of researchers very active.

Although the discussion cannot be complete owing to the large theoretical and applicative extent of the subject, in this chapter we aim to summarize the main principles of sparse signals representations. In section 2.2 the problem of representing sparse signals over redundant dictionaries is addressed together with the question of the choice of a suitable dictionary. In section 2.3, we present two main approaches to signals' sparse estimation leading to estimators with *thresholding* characteristics: the Penalized Least Squares and the Bayesian Maximum A Posteriori. The chapter ends with a review of existing works

on sparse signal detection, which lays the bases to the introduction of the sparsity-based detection tests of chapter 3.

## 2.2 Sparse representations of signals

### 2.2.1 Sparsity of signals in a dictionary of atoms

By definition, a signal is called *sparse* when most of its coefficients are (approximatively) null. Typically, if we consider a signal $\mathbf{x} \in \mathbb{R}^N$ and its support $\mathcal{I}(\mathbf{x})$, that is the set of non-zero components in $\mathbf{x}$, the signal is said to be $K-$sparse when $\text{card}(\mathcal{I}(\mathbf{x})) = K \ll N$.

Many signals are not sparse in their nature but show up to be sparse (or approximately sparse) in the domain of an appropriate transform. This means that for all these signals suitable spaces of representation exist in which they can be considered as sparse. To make an example, a sinusoidal signal, not sparse in the canonical basis, reveals to be sparse in the domain of the Fourier transform.

In the sparsity vocabulary the basis representation is called *dictionary*. The dictionary, hereafter used as a synthesis dictionary, say $\mathbf{R} = [\boldsymbol{r}_1, \ldots, \boldsymbol{r}_L]$, is a matrix containing the set of all the prototype functions (elementary signals) $\boldsymbol{r}_i$ through which our signal is modeled. Each one of these functions, also called *atoms*, corresponds to a column of $\mathbf{R}$. Fundamental references on this subject are [Mallat 2008, Elad 2010, Starck 2010]. When a (full column rank) dictionary contains more atoms than the signal length, we talk about *redundant* (or over-complete) dictionaries. In this case, a signal $\mathbf{x} \in \mathbb{R}^N$ is represented in the dictionary $\mathbf{R} \in \mathbb{R}^{N \times L}$, with $N \leq L$, by a linear superposition of atoms $\boldsymbol{r}_i$ in accordance with the following linear model:

$$\mathbf{x} = \sum_{i=1}^{J} \alpha_i \boldsymbol{r}_i, \qquad J \leq N. \tag{2.1}$$

We say that the signal $\mathbf{x}$ is *synthesized* in the dictionary $\mathbf{R}$.

When $L > N$, the atoms of the redundant dictionary are linearly dependent and no unique representation of a given signal $\mathbf{x}$ in $\mathbf{R}$ exists. A signal $\mathbf{x}$ is thus $K-$sparse in a dictionary $\mathbf{R}$ if the vector of coefficients $\boldsymbol{\alpha}$ presents $K$ non-zero components. The signal model becomes

$$\mathbf{x} = \mathbf{R}\boldsymbol{\alpha} = \sum_{i \in \mathcal{I}(\boldsymbol{\alpha})} \alpha_i \boldsymbol{r}_i, \qquad \text{with} \;\; \text{card}(\mathcal{I}(\boldsymbol{\alpha})) = ||\boldsymbol{\alpha}||_0 = K < N \ll L, \tag{2.2}$$

where $\mathcal{I}(\boldsymbol{\alpha})$ represents the *support* of $\boldsymbol{\alpha}$, that is the set of indices corresponding to the non-zero components in the coefficients vector. An example is illustrated in Figure 2.1. Here, a signal vector $\mathbf{x}$ is synthesized in a dictionary $\mathbf{R}$ of dimensions $N \times L$ through the representation vector $\boldsymbol{\alpha}$. In the coefficients vector $\boldsymbol{\alpha}$, only three components are different from zero (highlighted in green). Those components correspond to the three active atoms in the matrix $\mathbf{R}$, which provide the reconstruction of the signal $\mathbf{x}$. Note that Figure 2.1 shows the ideal case in which $\mathbf{x} = \mathbf{R}\boldsymbol{\alpha}$. In practice, due to the complexity of the signal and/or to the dictionary's characteristics, signals $\mathbf{x}$ are often only approximately reconstructed from atoms of $\mathbf{R}$ ($\mathbf{x} \approx \mathbf{R}\boldsymbol{\alpha}$). This is the case we will encounter in the sequel (see chapter 5 and chapter 6).

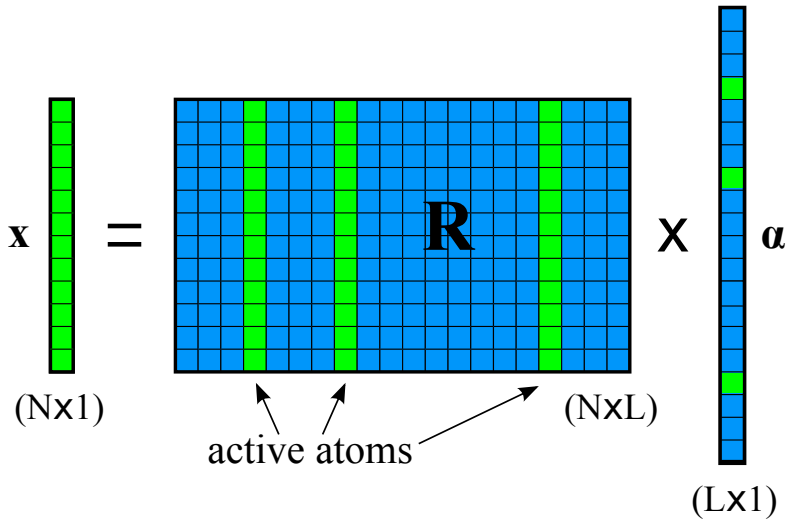



Figure 2.1: Synthesis of a $N-$dimensional vector $\mathbf{x}$ in a redundant dictionary $\mathbf{R}$ of dimensions $N \times L$, through a vector of coefficients $\boldsymbol{\alpha}$ of length $L$. Here $||\boldsymbol{\alpha}_0|| = 3$ and $\mathbf{x}$ is $3-$sparse in $\mathbf{R}$.

### 2.2.2 The choice of the dictionary

Dictionaries are one central piece in signals' sparse representations. The quality of the obtained representation, in terms of sparsity as well as in terms of approximation error, directly depends on the choice of the number and of the nature of the elementary functions that compose the dictionary.

In the redundant case, that is the one considered later on in this manuscript (see chapter 5), the dictionary generally forms a large catalog containing precise signal's forms. The intuition is that the larger the number of elementary atoms in the dictionary, the better the considered signal will be explained as the linear combination of a very few elementary functions, leading to a sparse but very accurate reconstruction. If we want the signal (or a class of signals) that we observe to be well represented, we need a dictionary that is suitable for them. A straightforward approach is thus to put in the dictionary all the forms or structures that may appear in our data. Since real signals can have very different structures, this reasoning could however lead to extremely large dictionaries, the use of which would make the computation of the signal representation difficult in a reasonable time for a sometimes negligible gain in presence of noise.

To overcome this problem, the overall dictionary is usually designed as the union of several sub-dictionaries, say $\{\mathbf{R}_i\}_{i=1,\dots,P}$ with $P < L$, where each sub-dictionary $\mathbf{R}_i$ represents a basis (or a set of simple functions) on which fast transforms are easily applicable. Examples of frequently used sets of functions are, for instance, the *Discrete Cosine* and *Sine Transform* bases (DCT and DST, respectively) to reproduce signal's oscillatory behaviors, the *wavelet* basis, to represent localized signals in time and frequency [Mallat 2008], and the canonical basis obviously ($\mathbf{I}$).

Despite their simplicity and their redundancy, those *fixed* and generic transform matrices are not always well adapted to the particular and specific behavior of the processed data. In

that case, two common issues are the following: i) to construct specific dictionaries, in accordance to the processed signal (known) characteristics and to the application framework, such as in [Starck 2010, Bourguignon 2011] for instance, in the field of astrophysics; ii) to resort to *dictionary learning* techniques that allow to obtain very well suited dictionaries for the signals we are considering. For dictionaries' design purposes (see section 6.3 and section 8.3), two different approaches to dictionary learning will be considered in this work: the first is the K - Singular Values Decomposition (K-SVD) approach introduced in [Aharon 2006], a brief presentation of which is proposed in Appendix B.3; the second is a minimax approach, which given a known set of possible signal features $\mathcal{L} \in \mathbb{R}^{N\times L} = [\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_L]$ with $L \gg N$ representing all possible alternatives under $\mathcal{H}_1$, determines the optimized atom $\boldsymbol{r}^*_{\text{minimax}} \in \mathbb{R}^N$, which maximizes the $\text{P}_{\text{DET}}(\boldsymbol{\ell}_i)$ occurring for the worst case of signal feature $\boldsymbol{\ell}_i$ in $\mathcal{L}$ [Lehmann 2005, Suleiman 2013]

$$\boldsymbol{r}^*_{\text{minimax}} = \underset{\boldsymbol{r},\ ||\boldsymbol{r}||_2^2=1}{\arg\max}\ \min_i\ \text{P}_{\text{DET}}(\boldsymbol{\ell}_i, \boldsymbol{r}), \tag{2.3}$$

with $\text{P}_{\text{DET}}(\boldsymbol{\ell}_i, \boldsymbol{r}) = Pr(\left|\boldsymbol{r}^t\mathbf{x}\right| > \xi \mid \mathcal{H}_1)$. This method will be used in section 8.3 to find the spectral signature that maximizes the worst probability of detection under $\mathcal{H}_1$ over a known library $\mathcal{L}$ of possible spectral line profiles.

## 2.3 Sparse estimation

Signal and image processing theories and methods have a long tradition of "sparsity promoting/aware" approaches. Most of the corresponding literature can be related, widely speaking, to an *estimation* framework and as opposed to the framework of *detection* – which is the main theme of this work.

State-of-the-art signal and image models are often based on dictionaries of elementary features (atoms) through which, as explained in the previous section, the information of interest can be synthesized with a good approximation by a linear combination of a relatively low number of atoms, with appropriate weights: most signal energy is living in a subspace (spanned by these atoms), whose dimension is much smaller than that of the signal. In this framework, and as we will see in a few lines, sparse estimation methods most often use sparsity-inducing (that is, thresholding) functions [Antoniadis 2001, Moulin 1999, Nikolova 2000], whose goal is to localize where the relevant information (the non-zero weights) concentrates.

In this section, we consider the problem of estimating a sparse deterministic unknown signal $\boldsymbol{\theta} \in \mathbb{R}^N$ based on a noisy observation modeled as

$$\mathbf{x} = \boldsymbol{\theta} + \boldsymbol{\epsilon}, \tag{2.4}$$

where $\mathbf{x} \in \mathbb{R}^N$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is an additive white Gaussian noise vector. Note that (2.4) also corresponds to the orthogonal case

$$\mathbf{y} = \mathbf{W}\boldsymbol{\theta} + \boldsymbol{\epsilon}', \tag{2.5}$$

with $\boldsymbol{\epsilon}' \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathbf{W} \in \mathbb{R}^{N\times N}$ so that $\mathbf{W}^t\mathbf{W} = \mathbf{I}$. Multiplying both terms in (2.5) by $\mathbf{W}^t$ we obtain

$$\mathbf{W}^t\mathbf{y} = \boldsymbol{\theta} + \mathbf{W}^t\boldsymbol{\epsilon}'. \tag{2.6}$$

By noting $\mathbf{W}^t\mathbf{y} = \mathbf{x}$, $\mathbf{W}^t\boldsymbol{\epsilon}' = \boldsymbol{\epsilon}$, the last equation reverts back to (2.4).

This kind of issue, also known as *denoising* problem, is often solved by estimating $\boldsymbol{\theta}$ through *Penalized Least Squares* (PLS). In a PLS problem, one wants to find $\boldsymbol{\theta}$ so as to solve a function of the form

$$\hat{\boldsymbol{\theta}}_{\text{PLS}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{2} ||\mathbf{x} - \boldsymbol{\theta}||_2^2 + \mu\varphi(\boldsymbol{\theta}). \tag{2.7}$$

We thus seek to minimize (with respect to $\boldsymbol{\theta}$) a particular energy function that is the sum of a data-fidelity term $\frac{1}{2} ||\mathbf{x} - \boldsymbol{\theta}||_2^2$ and a penalty term $\mu\varphi(\boldsymbol{\theta})$. Solutions to the PLS problem represent a trade-off between the data-proximity term (Least Squares for Gaussian noise) and the a priori signal characteristics taken into account through $\varphi(\boldsymbol{\theta})$. $\mu > 0$ is called *regularization parameter* and takes an important role in balancing between the penalty function's smoothness and the goodness-of-fit in the data fidelity term. While larger values of $\mu$ will put more weight in the regularization, smaller values of $\mu$ will put more emphasis on the data fidelity term. With reference to its dependence on the regularization parameter, we hereafter refer to the penalty function as $\varphi_\mu(\boldsymbol{\theta})$.

The performance of the estimator strongly depends on the nature of the penalty function and on the cost function $\mathcal{C}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{\text{PLS}})$, where $\boldsymbol{\theta}$ is the true value of the unknown parameter.

Next section illustrates the cases in which particular structures of the penalty function $\varphi(\boldsymbol{\theta})$ lead to *shrinkage* techniques, allowing to find sparse solutions to the signal denoising problem (2.7).

### 2.3.1 Penalty functions and thresholding estimators

We illustrate some examples of PLS estimators which, thanks to the characteristics of the penalty term injected in the minimization problem, show up solutions with interesting thresholding properties. For a more detailed and exhaustive investigation of the problem, we make reference to the works of [Antoniadis 2001, Moulin 1999, Nikolova 2000].

In order to produce sparse solutions, a necessary condition is that the penalty function $\varphi(\boldsymbol{\theta})$ is non-smooth at the origin [Antoniadis 2001] (examples are shown in Figures 2.2(a), 2.3(a) and 2.4(a)).

The choice of an additive function makes the estimation problem (2.7) separable. This means that for each of the $\boldsymbol{\theta}$ components, the corresponding PLS estimate can be computed as $\hat{\boldsymbol{\theta}}_{\text{PLS}} = [\hat{\theta}_{\text{PLS}_1}, \ldots, \hat{\theta}_{\text{PLS}_N}]$, with

$$\hat{\theta}_{\text{PLS}_i} = \arg\min_{\theta_i} \frac{1}{2} (x_i - \theta_i)^2 + \varphi_\mu(\theta_i). \tag{2.8}$$

To simplify the notation, we hereafter refer to this scalar case as

$$\hat{\theta}_{\text{PLS}} = \arg\min_{\theta} \frac{1}{2} (x - \theta)^2 + \varphi_\mu(\theta). \tag{2.9}$$

In the literature, many penalty functions are used. We illustrate three of them: the $l_1$ penalty function, leading to the *soft-thresholding* (ST) estimator [Donoho 1994]; the clipped $l_1$ penalty function, leading to the *hard-thresholding* (HT) estimator [Donoho 1994]; and the *smoothly clipped absolute deviation* (SCAD) penalty function [Antoniadis 1997], leading to an intermediate between the soft- and the hard-thresholding estimators.

#### 2.3.1.1 The Soft-Thresholding estimator

When considering the PLS problem (2.9), the ST estimator arises when choosing an $l_1$ penalty function of the form [Antoniadis 2001]

$$\varphi_\mu(\theta) = \mu\,|\theta|\,. \tag{2.10}$$

Indeed, replacing (2.10) in (2.9) we have

$$\hat{\theta}_{\text{PLS}} = \arg\min_{\theta} \frac{1}{2}\,(x-\theta)^2 + \mu\,|\theta|\,. \tag{2.11}$$

Since $\hat{\theta}_{\text{PLS}}$ is an even function in $\theta$, we consider $\theta \geq 0$. Thus

$$\hat{\theta}_{\text{PLS}} = \arg\min_{\theta} \frac{1}{2}\,(x-\theta)^2 + \mu\theta, \tag{2.12}$$

which is the equation of a parabola. Computing the derivative with respect to $\theta$ and equating it to zero yields

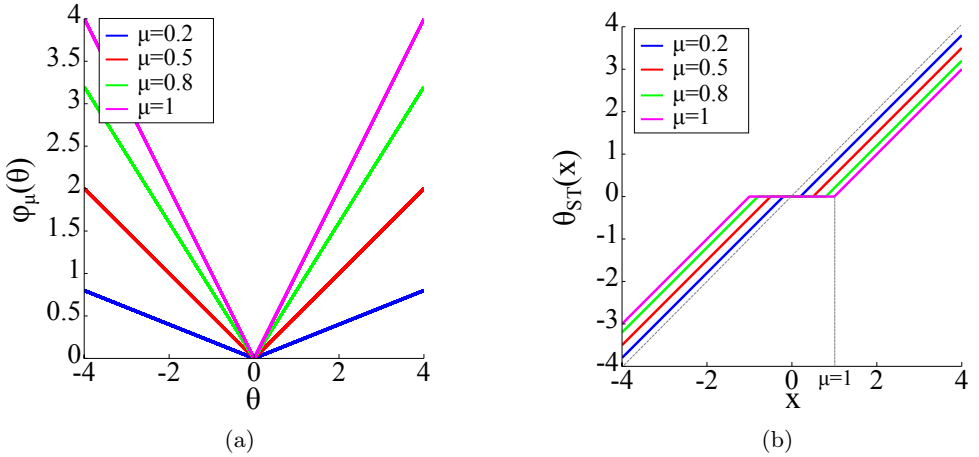
$$\theta = x - \mu. \tag{2.13}$$

(2.13) holds for $x > \mu$. If $x > \mu$, $x - \mu > 0$, so the minimum of the parabola for $\theta \in \mathbb{R}^+$ is $\hat{\theta}_{\text{PLS}} = x - \mu$. If $x \leq \mu$, the minimum is negative, so the minimum of the parabola for $\theta \in \mathbb{R}^+$ is obtained in $\hat{\theta}_{\text{PLS}} = 0$. According to this, we finally obtain

$$\hat{\theta}_{\text{PLS}} = \theta_{\text{ST}_\mu}(x) = \begin{cases} 0, & \text{if } |x| < \mu \\ x - \text{sgn}(x)\mu, & \text{if } |x| > \mu, \end{cases} \tag{2.14}$$

where the $\mu$ parameter takes the function of a threshold in the thresholding operator $\theta_{\text{ST}}$.

The ST estimator provides a continuous function of the data $x$ that is biased in the sense that the non-zero value of the estimate $\hat{\theta}_{\text{PLS}}$ always differ from the signal $x$ of a quantity




(a) (b)

Figure 2.2: 2.2(a) Examples of $l_1$ penalty functions. 2.2(b) Corresponding ST operators for four different values of $\mu \in [0.2, 0.4, 0.8, 1]$.

function of the threshold $\mu$, which penalizes large coefficients values. Figure 2.2 illustrates the $l_1$ penalty functions (2.10) (in Figure 2.2(a)) and the corresponding ST operators (2.14) (in Figure 2.2(b)), obtained for different values of the parameter $\mu \in [0.2, 0.4, 0.8, 1]$. Note that the stronger the $\mu$ threshold, the more sparse the $\hat{\theta}_{\text{PLS}}$ estimate we get, but also the stronger the estimation bias.

#### 2.3.1.2 The Hard-Thresholding estimator

The HT estimator can be derived from problem (2.9) by choosing a non-convex clipped $l_1$ penalty function of the form [Antoniadis 2001]

$$\varphi_\mu(\theta) = \mu^2 - \frac{1}{2}(|\theta| - \mu)^2 I(|\theta| < \mu), \tag{2.15}$$

where $I(\cdot)$[2] represents the indicator function. In accordance to the value of the indicator function, expression (2.15) can be rewritten as

$$\varphi_\mu(\theta) = \begin{cases} \mu^2, & \text{if } |\theta| \geq \mu \\ \mu^2 - \frac{1}{2}(|\theta| - \mu)^2, & \text{if } |\theta| < \mu. \end{cases} \tag{2.16}$$

Replacing (2.16) in (2.9) and solving with respect to the $\theta$ parameter (see proof in Appendix B.1), we obtain the expression of the $\hat{\theta}_{\text{PLS}}$ as

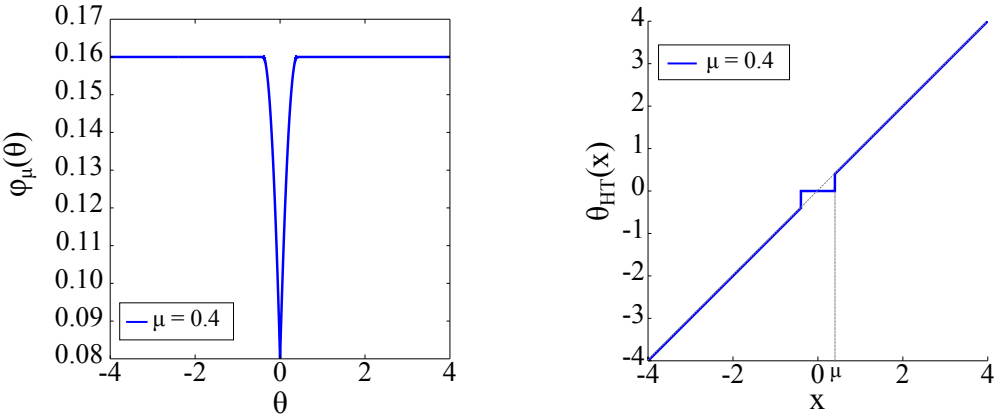
$$\hat{\theta}_{\text{PLS}}(x) = \theta_{\text{HT}_\mu}(x) = \begin{cases} 0, & \text{if } |x| < \mu \\ x, & \text{if } |x| > \mu, \end{cases} \tag{2.17}$$

which is the hard thresholding operator, also compactly, noted as [Donoho 1994]

$$\hat{\theta}_{\text{PLS}} = xI(|x| > \mu). \tag{2.18}$$

This is a very simple and intuitive thresholding rule: once the threshold $\mu$ has been fixed, the data coefficient is taken as the estimate of the parameter if its value falls above $\mu$. On the contrary case (coefficient value smaller or equal to $\mu$), the estimate is set to zero. In contrast to the ST operator, the HT solution (2.18) is unbiased. The clipped $l_1$ penalization function (2.15) and the corresponding HT operator (2.18) are shown for $\mu = 0.4$ in Figure 2.3(a) and Figure 2.3(b), respectively.

[2] By definition $I(x > a) = 1$ if $x > a$ and 0 otherwise.

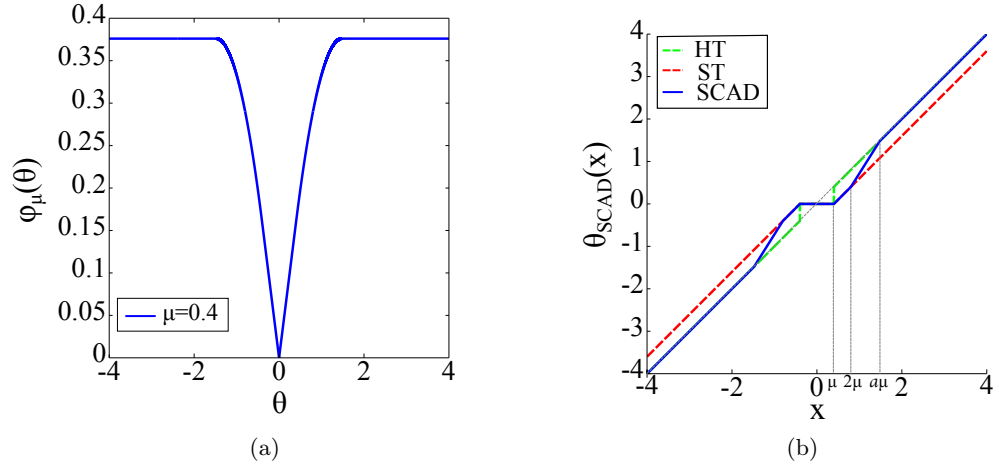



(a) (b)

Figure 2.3: 2.3(a) An example of clipped $l_1$ penalty function for $\mu = 0.4$. 2.3(b) Corresponding HT operator.

#### 2.3.1.3 The penalized SCAD thresholding estimator

We finally report an example of thresholding function that is intermediate between the ST and HT operators respectively seen in (2.14) and (2.18). The SCAD thresholding operator allows to overcome HT and ST drawbacks: HT always yields solutions that are not continuous with respect to the data $x$. On the other hand, while providing continuous solutions, the ST operator gives biased estimates because of the shifting of $\mu$.




(a) (b)

Figure 2.4: 2.4(a) Example of SCAD penalty function for $\mu = 0.4$ and $a = 3.7$. 2.4(b) Corresponding SCAD thresholding operator (in blue). The HT and ST functions are reported in dashed green and dashed red, respectively.

Using in the PLS problem (2.9) a SCAD penalty function of the form [Fan 1997,

Fan 2001]

$$\varphi_\mu(\theta) = \begin{cases} \mu\,|\theta|\,, & \text{if } |\theta| \leq \mu \\ -\left(\frac{|\theta|^2 - 2a\mu|\theta| + \mu^2}{2(a-1)}\right), & \text{if } \mu < |\theta| \leq a\mu \\ \frac{(a+1)\mu^2}{2}, & \text{if } |\theta| > a\mu \end{cases} \tag{2.19}$$

for $a > 2$, leads to the linear thresholding function

$$\hat{\theta}_{\text{PLS}} = \theta_{\text{SCAD}}(x) = \begin{cases} \text{sgn}(x)(|x| - \mu), & \text{if } |x| \leq 2\mu, \\ +\frac{(a-1)x - a\mu\,\text{sgn}(x)}{a-2}, & \text{if } 2\mu < |x| \leq a\mu, \\ x, & \text{if } |x| > a\mu. \end{cases} \tag{2.20}$$

Figure 2.4 reports the penalized SCAD function (in Figure 2.4(a)) and the corresponding continuous and asymptotically unbiased estimator (in Figure 2.4(b)) for $a = 3.7$ [Fan 2001] and for $\mu = 0.4$.

#### 2.3.1.4 Case of $l_p$ penalty functions, with $p \leq 1$

Another important particular case is when $\varphi(\boldsymbol{\theta})$ is chosen as

$$\varphi_\mu(\boldsymbol{\theta}) = \mu\,||\boldsymbol{\theta}||_p^p = \mu \sum_{i=1}^{N} |\theta_i|^p\,, \tag{2.21}$$

with $0 < p \leq 1$. According to (2.21), the scalar PLS problem (2.9) becomes

$$\hat{\theta}_{\text{PLS}} = \arg\min_\theta \frac{1}{2}(x-\theta)^2 + \mu\,|\theta|^p\,. \tag{2.22}$$

The solution of this problem also yields thresholding estimators, a study of which can be found in [Moulin 1999]. The particular case of $p = 1$, leading to the ST estimator, was reported in subsubsection 2.3.1.1.

### 2.3.2 Bayesian framework for sparsity-promoting estimators

We now aim at investigating the sparse estimation problem from a Bayesian point of view. In this purpose, we consider the case of the following scalar model

$$x = \theta + \epsilon, \tag{2.23}$$

where a noisy data component $x$ is function of an unknown stochastic parameter $\theta$, which we want to estimate. Here, the noise component is $\epsilon \sim \mathcal{N}(0,1)$.

#### 2.3.2.1 Bayesian Risk Functions

A Bayesian estimator can always be considered as the minimizer of some *Bayesian risk function* $\mathcal{R}$ ([Kay 1998b], from which the results below are drawn). The Bayesian risk function $\mathcal{R}$ gives a measure of the performance of the considered estimator. Denoting

$$\epsilon = \theta - \hat{\theta}, \tag{2.24}$$

the error between the estimator $\hat{\theta}$ and $\theta$, and $\mathcal{C}(\epsilon)$ the *cost function* dependent on the error $\epsilon$, the risk is defined as

$$\mathcal{R} = \mathbb{E}\left[\mathcal{C}(\epsilon)\right], \tag{2.25}$$

representing, in accordance with a Bayesian point of view, the expectation value of the cost function $\mathcal{C}(\epsilon)$ with reference to the *joint* PDF $p(x, \theta)$ of $x$ and $\theta$.

Several types of cost functions exist. In this section we focus on two of the most commonly used ones: the *quadratic* loss function

$$\mathcal{C}(\epsilon) = \epsilon^2, \tag{2.26}$$

shown in Figure 2.5(a), and the *hit-or-miss* loss function illustrated in Figure 2.5(b)

$$\mathcal{C}(\epsilon) = \begin{cases} 0, & \text{if } |\epsilon| < \delta \\ 1, & \text{if } |\epsilon| > \delta, \text{ with } \delta > 0. \end{cases} \tag{2.27}$$

Definition (2.25) yields (summations are over the appropriate domains of $x$ and $\theta$)

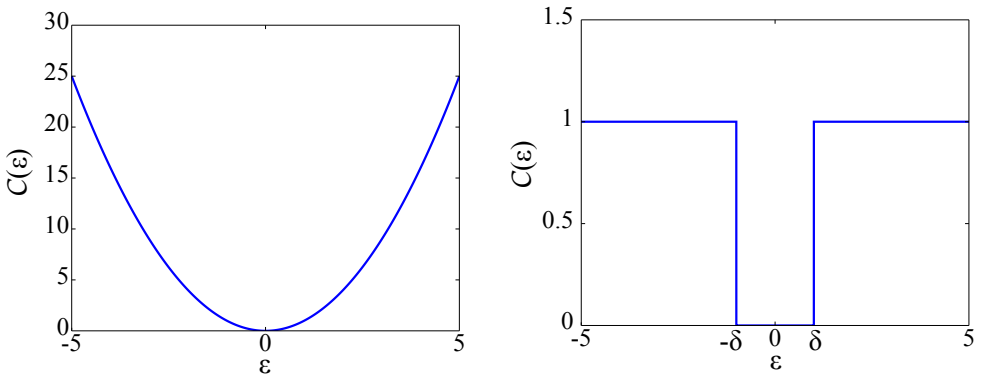
$$\mathcal{R} = \mathbb{E}\left[\mathcal{C}(\epsilon)\right] = \int\int \mathcal{C}(\epsilon) p(x, \theta)\, \mathrm{d}x\mathrm{d}\theta. \tag{2.28}$$

According to the Bayes' theorem (1.60), which states that

$$p(x, \theta) = p(\theta \mid x)\pi(x), \tag{2.29}$$

the expression of the $\mathcal{R}$ function in (2.25) becomes

$$\begin{aligned} \mathcal{R} &= \int\int \mathcal{C}(\epsilon) p(\theta \mid x)\pi(x)\, \mathrm{d}x\mathrm{d}\theta \\ &= \int \left[\int \mathcal{C}(\epsilon) p(\theta \mid x)\mathrm{d}\theta\right] \pi(x)\, \mathrm{d}x. \end{aligned} \tag{2.30}$$




(a) (b)

Figure 2.5: 2.5(a) Example of quadratic function. 2.5(b) Example of Hit-or-Miss loss function.

We make the following important observation: since $\pi(x) \geq 0$, minimizing the $\mathcal{R}$ function (2.30) is equivalent to minimize the integral in brackets over $\theta$. This integral is hereafter denoted as the function

$$g(\hat{\theta}) = \int \mathcal{C}(\epsilon) p(\theta \mid x) \mathrm{d}\theta. \tag{2.31}$$

For the two $\mathcal{C}(\epsilon)$ error functions above, the minimization of the corresponding risk function $\mathcal{R}$ leads to two well known optimal estimators: the Minimum Mean Square Error and the Maximum A Posteriori estimators. The explicit computation of their expression is addressed below.

#### 2.3.2.2 The Minimum Mean Square Error estimator

We replace in (2.30) the quadratic loss function introduced in (2.26). Recalling that $\mathcal{C}(\epsilon) = \epsilon^2 = (\theta - \hat{\theta})^2$, we seek to find the estimator $\hat{\theta}$ that minimizes the Mean Square Error (MSE) risk function

$$\mathcal{R} = \mathbb{E}\left[(\theta - \hat{\theta})^2\right] = \int \int (\theta - \hat{\theta})^2 p(\theta \mid x) \mathrm{d}\theta \pi(x) \, \mathrm{d}x. \tag{2.32}$$

According to what stated above, this is equivalent to minimize

$$g(\hat{\theta}) = \int (\theta - \hat{\theta})^2 p(\theta \mid x) \mathrm{d}\theta. \tag{2.33}$$

Taking the partial derivative of (2.33) with respect to $\hat{\theta}$ we have

$$\begin{aligned} \frac{\partial}{\partial \hat{\theta}} \int (\theta - \hat{\theta})^2 p(\theta \mid x) \mathrm{d}\theta &= \int \frac{\partial}{\partial \hat{\theta}} (\theta - \hat{\theta})^2 p(\theta \mid x) \mathrm{d}\theta \\ &= \int -2(\theta - \hat{\theta}) p(\theta \mid x) \mathrm{d}\theta \\ &= -2 \int \theta p(\theta \mid x) \mathrm{d}\theta + 2\hat{\theta} \int p(\theta \mid x) \mathrm{d}\theta. \end{aligned} \tag{2.34}$$

Since

$$\int p(\theta \mid x) \mathrm{d}\theta = 1, \tag{2.35}$$

setting (2.34) to zero finally gives

$$\begin{aligned} \hat{\theta}_{\mathrm{MMSE}} &= \int \theta p(\theta \mid x) \mathrm{d}\theta \\ &= \mathbb{E}(\theta \mid x). \end{aligned} \tag{2.36}$$

This expression tells us that the Bayesian estimator that minimizes the MSE is equal to the mean a posteriori.

The expression of the MMSE estimator can be easily generalized to the vectorial case. In fact, for a noisy data vector $\mathbf{x}$, function of a parameter vector $\boldsymbol{\theta}$, the MMSE estimator resumes to compute the mean of the posterior PDF for *each* component in the vector of parameters, leading to

$$\hat{\boldsymbol{\theta}} = \mathbb{E}(\boldsymbol{\theta} \mid \mathbf{x}). \tag{2.37}$$

We refer to [Kay 1998b] for the proof details.

#### 2.3.2.3 The Maximum A Posteriori estimator

We now replace in (2.30) the hit-or-miss loss function introduced in (2.27). In this case, we have $\mathcal{C}(\epsilon) = 1$ for $\epsilon > \delta$ and $\epsilon < -\delta$ (or equivalently for $\theta > \hat{\theta} + \delta$ and $\theta < \hat{\theta} - \delta$, respectively). The optimal Bayesian estimator is found minimizing the function

$$g(\hat{\theta}) = \int_{-\infty}^{\hat{\theta}-\delta} p(\theta \mid x)\mathrm{d}\theta + \int_{\hat{\theta}+\delta}^{+\infty} p(\theta \mid x)\mathrm{d}\theta. \tag{2.38}$$

Since

$$\int_{-\infty}^{+\infty} p(\theta \mid x)\mathrm{d}\theta = 1, \tag{2.39}$$

the $g(\hat{\theta})$ can be written as

$$g(\hat{\theta}) = 1 - \int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} p(\theta \mid x)\mathrm{d}\theta, \tag{2.40}$$

from which we have that minimizing (2.40) actually corresponds to *maximize* the integral

$$\int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} p(\theta \mid x)\mathrm{d}\theta. \tag{2.41}$$

The Bayesian estimator that minimizes the risk function $\mathcal{R}$ (2.30) for a hit-or-miss loss function of type (2.27) with $\delta \to 0$ is thus the one that maximizes the posterior PDF of $\theta$ given the data $x$. We call this estimator the Maximum A Posteriori (MAP) estimator, and we mathematically write it as

$$\hat{\theta}_{\mathrm{MAP}} = \arg\max_{\theta} p(\theta \mid x). \tag{2.42}$$

In the next section, we introduce and study the advantageous sparsity promoting properties of the MAP estimate using appropriate prior laws.

### 2.3.3 MAP sparsity-promoting estimate of unknown parameters

The question of finding an estimate of the unknown deterministic parameter vector $\boldsymbol{\theta}$ has been addressed in subsection 1.3.2 in the context of the GLR test. As specified by equations (1.53) and (1.54), the ML estimate of $\boldsymbol{\theta}$ for model (1.50), leads to $\hat{\boldsymbol{\theta}}_{ML} = \mathbf{x}$. Nevertheless, when taking into account sparse parameter vectors, as it is the case in our work, the unconstrained ML estimate reveals not to be in agreement with the considered model because it is not sparse. In this section we seek to find sparse estimates of $\boldsymbol{\theta}$ that will be further injected in the detection approaches later discussed. We focus on the MAP estimate of $\boldsymbol{\theta}$ because it is associated to the Hit-or-Miss cost function (see subsection 2.3.2), which is appropriated for sparse vectors. In particular we focus on the appropriate priors which will induce sparsity.

As already defined in the previous section, for a scalar parameter $\theta$ the MAP estimate writes

$$\begin{aligned} \hat{\theta}_{\mathrm{MAP}} &= \arg\max_{\theta} p(\theta \mid x) \\ &= \arg\max_{\theta} \; p\left(x \mid \theta\right) \pi\left(\theta\right). \end{aligned} \tag{2.43}$$

With $-\ln p\,(x \mid \theta) = \frac{1}{2}(x-\theta)^2$ and $-\ln \pi\,(\theta) = \varphi_\mu(\theta)$, (2.43) reverts back to a PLS thresholding problem (see subsection 2.3.1 and Example 2.3.1). From [Moulin 1999, Antoniadis 2001], prior distributions $\pi(\theta)$ that are mono-modal and not differentiable at the origin, will yield sparse estimates $\hat{\theta}_{\text{MAP}}$.

For example, Laplacian priors indeed lead to soft-thresholding functions (see Example 2.3.1 and Figure 2.2(b)). Similarly, Generalized Gaussians (GG) with $p-$parameter smaller than 1 correspond to the $l_p$ norms of subsubsection 2.3.1.4, which yield thresholding functions that are well studied in terms of bias and threshold coordinates (as discussed in [Moulin 1999, Antoniadis 2001]). Thresholding functions are also obtained using Bernoulli-Gaussian priors [Pesquet 1996].

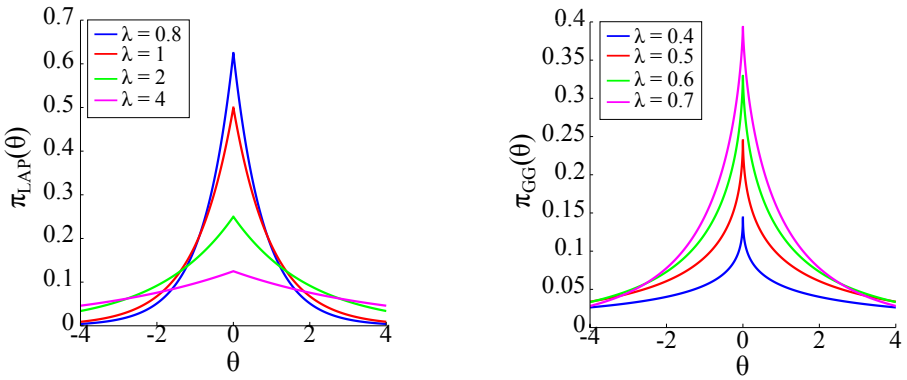
Examples of Laplacian and GG distributions for a scalar parameter $\theta$ are shown in Figure 2.6: in Figure 2.6(a), a Laplacian prior of the form

$$\pi_{\text{LAP}}\,(\theta) = \frac{1}{2\lambda} e^{-\frac{|\theta|}{\lambda}} \tag{2.44}$$

is plotted for different values of the $\lambda$ parameter ($\lambda = 0.8, 1, 2, 4$), while Figure 2.6(b) shows a GG prior of the form

$$\pi_{\text{GG}}\,(\theta) = \frac{p}{2s\Gamma\left(\frac{1}{s}\right)} e^{-\frac{|\theta-\mu|^p}{s^p}}, \tag{2.45}$$

where $\Gamma(\cdot)$ represents the Gamma function and $s, p \in \mathbb{R}^+$ are real positive parameters. In Figure 2.6(b), $s = 1$, $\mu = 0$ and $p = 0.4, 0.5, 0.6, 0.7$.




(a) (b)

Figure 2.6: 2.6(a) Examples of Laplacian prior for three different values of $\lambda \in \{1, 2, 4\}$. 2.6(b) Examples of Generalized Gaussian prior for three different values of $p \in [0, 1]$.

Thanks to the thresholding properties it yields and to the simple computation of the solution it allows, the Laplacian prior encapsulates most of the desirable properties a sparse-detection should possess. This prior law will be considered throughout this work. The derivation of the ST operator when injecting a Laplacian prior in the MAP estimate is reported in the example below.

**Example 2.3.1. - MAP estimate with Laplacian prior and ST operator**

We consider the case of the following vector model

$$\mathbf{x} = \boldsymbol{\theta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{C}) \tag{2.46}$$

where a noisy data vector $\mathbf{x}$ is function of an unknown parameter $\boldsymbol{\theta}$, which we want to estimate. We consider a Laplacian prior of the form

$$\pi(\boldsymbol{\theta}) = \prod_{i=1}^{N} \frac{1}{2\lambda_i} e^{-\frac{|\theta_i|}{\lambda_i}}. \tag{2.47}$$

For model (2.46) and prior distribution (2.47), the computation of the $\hat{\boldsymbol{\theta}}_{\text{MAP}}$ estimate can be made component-wise. We have

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \left[\hat{\theta}_{MAP_1}, \cdots, \hat{\theta}_{MAP_N}\right]^t, \tag{2.48}$$

where each scalar component $\hat{\theta}_{MAP_i}$ (noted $\hat{\theta}_{\text{MAP}}$ below) of $\hat{\boldsymbol{\theta}}_{\text{MAP}}$ is given by

$$\begin{aligned} \hat{\theta}_{\text{MAP}} &= \arg\max_{\boldsymbol{\theta}} \; p\left(\theta \mid x\right) \\ &= \arg\max_{\boldsymbol{\theta}} \; p\left(x \mid \theta\right) \pi\left(\theta\right) \\ &= \arg\min_{\boldsymbol{\theta}} \; -\ln p\left(x \mid \theta\right) \pi\left(\theta\right). \end{aligned} \tag{2.49}$$

For a Gaussian Likelihood function $p\left(x \mid \theta\right) \sim \mathcal{N}\left(0, \sigma^2\right)$ and a scalar Laplacian prior, definition (2.49) becomes

$$\begin{aligned} \hat{\theta}_{\text{MAP}} &= \arg\min_{\boldsymbol{\theta}} \; -\ln \left\{ \frac{1}{2\sigma\lambda\sqrt{2\pi}} \exp\left[ -\frac{(x-\theta)^2}{2\sigma^2} - \frac{|\theta|}{\lambda} \right] \right\} \\ &= \arg\min_{\boldsymbol{\theta}} \; -\ln \left( \frac{1}{2\sigma\lambda\sqrt{2\pi}} \right) + \left[ \frac{(x-\theta)^2}{2\sigma^2} + \frac{|\theta|}{\lambda} \right] \\ &= \arg\min_{\boldsymbol{\theta}} \; \frac{(x-\theta)^2}{2\sigma^2} + \frac{|\theta|}{\lambda}. \end{aligned} \tag{2.50}$$

Expression (2.50) is equivalent to the PLS problem (2.11) with $\frac{\sigma^2}{\lambda} = \mu$. Thus, the MAP estimate is obtained by the soft-thresholding operator:

$$\hat{\theta}_{\text{MAP}} = \begin{cases} 0, & \text{if } -\frac{\sigma^2}{\lambda} < x < \frac{\sigma^2}{\lambda} \\ x - \operatorname{sgn}(x) \frac{\sigma^2}{\lambda}, & \text{otherwise.} \end{cases} \tag{2.51}$$

Examples of soft thresholding functions of the form (2.51) have already been shown in Figure 2.2(b). Note that, the lower the $\lambda$ parameter, which corresponds to a sharper prior function (see Figure 2.6(a)), the higher the $\sigma^2/\lambda$ threshold. Consequently, the higher the value of the threshold, the stricter the condition on data values $\mathbf{x}$ to obtain a nonzero MAP estimate for the $\boldsymbol{\theta}$ components.

## 2.4 Sparse detection

All the works presented in this manuscript generally deals with the topic of *sparse detection*. We aim at setting tests that are powerful for the detection of signals having many zeros or, more generally, many very weak components.

In the particular but extremely frequent case of sparse parameters (as for instance in data models using dictionaries), estimation and detection share a common goal: finding spaces of reduced dimensions where useful information is located and discard the others. This was illustrated through the thresholding functions of section 2.3. Despite this connection and the popularity of sparsity, the signal processing literature in sparse detection is far less abundant than that of sparse estimation. A brief survey is proposed below.

### 2.4.1 Previous works in sparse detection

In the seminal paper of Donoho and Johnstone [Donoho 1994], near optimality[3] of hard- and soft- thresholding operators is established in an estimation (denoising) framework using orthogonal wavelets. These adaptive operators are introduced using a simple and powerful heuristic argument: while all empirical wavelets coefficients contribute noise, only very few contribute signal. Connection with detection is evoked in the same paper by relating thresholding operators to an automated version of the model selection rule, using $z$-scores[4] in statistics.

In the statistical literature, Fan [Fan 1996] makes more explicit the connection between estimation of sparse parameters using thresholding operators and detection tests adapted to such parameters. The introductory examples analyzed in [Fan 1996] show that classical detection tests based on the empirical data distribution functions (like the Kolmogorov-Smirnov and the Cramér-Von Mises tests[5]) are not appropriate for such parameters. One existing test (the Neyman's truncation[6]) would however be appropriate by relaxing the assumption made in this test that the information is concentrated in the first components of the parameter vector. Fan [Fan 1996] thus proposes a straightforward improvement to adaptively select the components used in Neyman's test, using thresholding operators. Hence, the resulting test statistics are derived from the same heuristic as in [Donoho 1994], rather than from the "canonical" form of explicit density or probability ratios – in which we are interested, as this provides a precise ground for the interpretation of such tests.

As will be shown in section 3.3, the test described by Fan with soft-thresholding is actually a special instance of another test, the Posterior Density Ratio (PDR) [Basu 1996], introduced in the sixties by I. J. Good [Good 1965]. The PDR is discussed in [Basu 1996]

[3] With respect to an ideal estimator.

[4] For a realization $x$ of a random variable $X$ following distribution $\mathcal{D}(\mu, \sigma)$, standard scores, also called $z$-values or $z$-scores are are defined as $z(x) = \frac{x-\mu}{\sigma}$.

[5] The Kolmogorov-Smirnov (KS) test is a hypothesis test used to determine the goodness of fit beween a sample's empirical CDF and a reference distribution, or between the empirical CDF of two samples. The Cramér-Von Mises (CVM) test has the same application of the Kolmogorov-Smirnov test. The main difference between the two tests is that while the KS test focuses only on the maximum difference between the empirical and the reference distributions is taken into consideration, the CVM test takes better account of the data set involving the sum of those differences.

[6] Neyman's truncation or Adaptive Neyman's Test (ANT) [Fan 1996]: given an N-dimensional data vector $\mathbf{x}$, the ANT proposes to use the testing procedure $||\mathbf{x}||_2^2 \underset{H_0}{\overset{H_1}{\gtrless}} \gamma$ only on the first $m$ components of $\mathbf{x}$.

in the case of scalar parameters, outside the framework of sparsity. This test seems absent from the subsequent literature of the "sparse representation era" [Fuchs 2010] despite its good behavior for sparse signals.

In the Signal Processing literature, a few works [Haupt 2007, Wang 2008, Paredes 2009, Fuchs 2010, Malloy 2011, Bajwa 2013] have recently proposed detection tests exploiting sparsity, mostly in the Compressed Sensing framework. In [Haupt 2007], a best $k$-term approximation of the parameter vector is injected in the Likelihood Ratio. This is similar in spirit to the approach called LRMAP that we introduce in the next chapter. In these works, however, no analytical study is attempted to investigate the effects of the randomness of the estimate injected in the test. We propose such an analysis in section 3.5. The same idea appears in [Wang 2008], where a MAP estimate is used (as a solution of the Basis Pursuit DeNoising (BPDN) or LASSO (Least absolute shrinkage and selection operator) [Tibshirani 1996, Chen 2001, Fuchs 2004]), and the tests efficiency is also evaluated numerically. In [Paredes 2009], the sensing matrix is designed to correspond to the subspace of the sparse signals, which is assumed to be known. The works presented in [Bajwa 2013] propose to optimize Parseval frames for signal detection using the GLR. In [Malloy 2011], sparse recovery is investigated in the framework of multiple testing where the number of measurements per components is allowed to be optimized, with the constraint that their total number is fixed. In [Fuchs 2010], Fuchs describes the connection between the support detection performed by the BPDN and a GLR, testing whether data contain a signal that has sparsity one in some dictionary. The model selection approach for sparse detection is adopted in the paper of Larsson and Selen [Larsson 2007]. In this work, the considered sparse models correspond to all the possible sparse configurations of the parameter vector. In the MMSE estimation framework, a greedy algorithm is proposed in order to reduce the combinatorial complexity of this approach.

Back to the statistical literature, the test called Higher Criticism (HC) is set in [Donoho 2004] in order to detect faint sparse vector means of multivariate normal vectors, whose covariates under $\mathcal{H}_0$ are zero-mean, independent and identically distributed. In [Arias-Castro 2010], this test is applied to the case where data are observed through a sensing matrix, which is a random and redundant dictionary. Another test, based on a simple Bonferroni-type correction[7] (the *Max* test), is also considered. The *Max* test corresponds actually to a particular case of the proposed PDR and LRMAP approaches introduced in the next chapter. Both the HC and the *Max* test will be shown to be powerful for highly sparse signals (cf. chapter 5). For comparison purposes, these tests will be taken into account in section 5.2 and section 6.3, in the context of dictionary model-based detection tests. A brief description of HC procedure is thus proposed below.

### 2.4.2 The Higher Criticism

Introduced by J. W. Tukey in 1976 [Tukey 1976], the notion of *Higher Criticism* or second-level significance testing, refers to a multiple comparison (multiple testing) problem that occurs when a set of statistical inferences are considered simultaneously to discriminate between two hypotheses.

[7] Bonferroni-type correction is a way to control the Type I error rate in multiple testing problems. When testing $n$ independent hypotheses on a set of data, each individual hypothesis is set to a statistical significance level $\alpha/n$, where $\alpha$ is the Type I error rate that would be set if only one hypothesis was considered.

The following example introduces the problem: we want to state the efficiency of a new medical treatment on a given set of $n$ patients. For each patient we are given an observation of the form

$$z_i = \theta_i + w_i, \qquad i = 1, \dots, n \tag{2.52}$$

where the index $i$ designs the $i-$th patient, $\theta_i$ represents the effect of the treatment on the $i-$th patient and $w_i$ is a measurement error assumed to be $\mathcal{N}(0,1)$. The corresponding testing model becomes

$$\begin{cases} \mathcal{H}_0 : \theta_1 = \theta_2 = \cdots = \theta_n = 0 \\ \mathcal{H}_1 : \text{At least one } \theta_i \neq 0, \end{cases} \tag{2.53}$$

which means to test the absence of the treatment effect on the overall joint null hypothesis $\mathcal{H}_0$ versus its presence in at least one of the tested patients. To make the test valid, we set a significance level $\alpha$ (say $\alpha = 5\%$) for each test performed on a single patient.

As seen in chapter 1, the significance level conditions the validation of the effects of the treatment, allowing to quantify the risk of wrong conclusions, as for example Type I errors. When multiple statistical tests are performed simultaneously however, the global risk of getting wrong conclusions obviously rises. This is for example the case if the multiple testing procedure involves $n$ consecutive independent tests. In this case, the overall risk to wrongly conclude on the effectiveness of the treatment is not 5% (the threshold for each single test), but it is $n$ times higher. The significance level $\alpha$ is said to be subject to inflation.

With his *second-level significance testing* [Tukey 1976], Tukey proposed a method to control the inflation of the risk when multiple comparisons are needed. In 2004, Tukey's work has been further developed by D.L. Donoho and J. Jin [Donoho 2004], who applied the HC procedure to the detection of weak sparse signals in white noise. The authors were precisely interested in (2.53): *testing whether* n *normal means are all zero versus the alternative that a small fraction is positive and nonzero* [Donoho 2004].

To implement the HC, the authors consider $n$ independent models

$$\begin{cases} \mathcal{H}_{0_i} : z_i \sim \mathcal{N}(0,1) \\ \mathcal{H}_{1_i} : z_i \sim \mathcal{N}(\mu_i, 1) \qquad \mu_i > 0, \ i = 1, \dots, n. \end{cases} \tag{2.54}$$

Similarly to the example illustrated above, it is expected that in the majority of the tests $\mu_i = 0$, while for a small fraction of the tests, $\mu_i > 0$. At this point, the data $z_i$ are converted in terms of $p$-values (defined in subsection 1.2.4) under $\mathcal{H}_0$

$$p_i = Pr\left\{\mathcal{N}(0,1) > z_i\right\}. \tag{2.55}$$

Denoting by $p_{(i)}$ the resulting $p$-values sorted in increasing order ($p_{(1)} < p_{(2)} < \cdots < p_{(n)}$) the HC test statistic writes:

$$\mathrm{HC}_n^* = \underset{\{1 \geq i \geq \alpha_0 n\}}{\arg\max}\ \mathrm{HC}_{n,i}, \tag{2.56}$$

where the $i$-th normalized $z$-score $\mathrm{HC}_{n,i}$, defined as

$$\mathrm{HC}_{n,i} = \sqrt{n} \cdot \left[\left|i/n - p_{(i)}\right|\right] / \left[\sqrt{p_{(i)}\left(1 - p_{(i)}\right)}\right], \tag{2.57}$$

is maximized over a considered range of $p$-values set by $\alpha_0$. The parameter $\alpha_0$ can be set in accordance to the data degree of sparsity under $\mathcal{H}_1$. For instance, if $\alpha_0 = 1/n$, only the

smallest $p$-value (the largest component) $p_{(1)}$ is accounted for. Choosing $\alpha_0 = 1/2$ allows to test only on the half smallest $p$-values (the largest components).

The HC approach can be interpreted as follows:

- The test compares the maximal deviation of the empirical CDF $F_n(t)$ of $p$-values $p_i$, evaluated at $p_{(1)}, \ldots, p_{(\alpha_0 n)}$ to their true CDF under $\mathcal{H}_0$, $F(t)$. Figure 2.7 illustrates an example of the two CDFs $F_n(t)$ and $F(t)$ when taking $n = 5$ and $\alpha_0 = 1$.

- As demonstrated in subsection 1.2.4, under $\mathcal{H}_0$, the $p$-values are uniformly distributed as $\mathcal{U}(0,1)$, and their CDF is $F(t) = t$ (green plot). On the other hand, the empirical CDF $F_n(t)$ (blue plot) is a random variable that may take values $0, \frac{1}{n}, \ldots, \alpha_0$. For example, the realization $F_n(t) = \frac{k}{n}$ means that we found $k$ $p$-values (out of $\alpha_0 n$) having magnitude less or equal to $t$. Hence, $F_n(t) = \frac{K(t)}{n}$, where $K$ is the number of $p$-values less or equal to $t$.

  Since the $p$-values are independent, $F_n(t) = \frac{K(t)}{n} \sim \mathcal{B}\left(\mu = t, \sigma = \sqrt{nt(1-t)}\right)$. Asymptotically (when $n \to \infty$), $\mathcal{B}\left(\mu = t, \sigma = \sqrt{nt(1-t)}\right) \xrightarrow{d} \mathcal{N}\left(t, \sqrt{nt(1-t)}\right)$[8].

At this point, computing the $z$-scores (see subsection 2.4.1) of $F_n(t)$ at values $t = p_{(i)}$, for $i = 0, \ldots, \alpha_0 n$, yields (2.57) and the HC procedure thus works at the maximal $z$-score in (2.56).

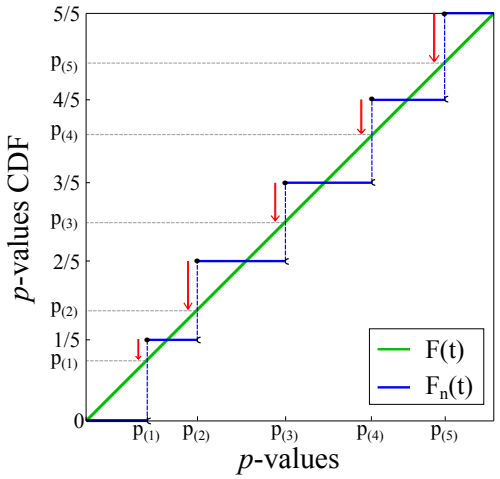



Figure 2.7: True (in green) and Empirical (in blue) CDF of $p$-values $p_i$. The $z$-scores $\mathrm{HC}_{5,i}$ are indicated by red arrows (up to factor $\sqrt{n}/\left[\sqrt{p_{(i)}\left(1-p_{(i)}\right)}\right]$ in (2.57)). The HC statistic is the max of those weighted deviations to $F(t)$.

[8] Convergence in distribution ($\xrightarrow{d}$): A sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$ is said to converge in distribution to a random variable $X$ if $\lim_{n\to\infty} F_n(x) = F(x)$, for every number $x \in \mathbb{R}$ at which $F$ is continuous. Here, $F_n$ and $F$ are the CDFs of $X_n$ and $X$, respectively.

### 2.4.3 A glance at sparse detection in Hyperspectral Imaging

The main application of our work concerns the detection of low Signal-to-Noise Ratio and sources sparse in dictionary in multi-dimensional noisy data. In particular, we focus on hyperspectral three-dimensional data. As it will be detailed in chapter 6 and chapter 7, to each pixel of a hyperspectral image is associated a vector, which reflects the pixel's characteristics in function of different spectral bands. In this context, a multitude of general principles and methods have been proposed.

#### 2.4.3.1 Spectral-based detection techniques

One of the most widespread problems in hyperspectral imaging is target detection. According to the spectral characteristics of a pixel under test, target detection algorithms aim to find out if this pixel contains a target signature or not. Hyperspectral target detection algorithms can be divided into two groups: *anomaly detection algorithms* and *spectral matching detection algorithms*. In anomaly detection algorithms, no prior knowledge about the target spectrum is required. Since target and background spectra usually show distinct spectral characteristics, the main goal is to identify pixels that substantially differ from the image background. On the other hand, spectral matching detection algorithms assume the target spectral signature to be known. In this case, all those pixels whose spectrum is highly correlated with the reference one are detected as targets.

Statistical hypothesis testing techniques have been extensively used for target detection [Manolakis 2002]. Two first examples are the *Matched Filter* (MF) and the *Adaptive Matched Filter* (AMF) approaches introduced in [Trees 1968]. For a given $N-$dimensional observation vector $\mathbf{x}$, we consider the two following hypotheses

$$\begin{cases} \mathcal{H}_0 : \mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_b, \mathbf{C}_b) \\ \mathcal{H}_1 : \mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_t, \mathbf{C}_t), \end{cases} \tag{2.58}$$

where $\mathbf{x}$ is normally distributed (with mean vector and covariance matrix that depend on the considered hypothesis) and assumed to belong to the background under $\mathcal{H}_0$ ($\boldsymbol{\mu} = \boldsymbol{\mu}_b$ and $\mathbf{C} = \mathbf{C}_b$) or to the target under $\mathcal{H}_1$ ($\boldsymbol{\mu} = \boldsymbol{\mu}_t$ and $\mathbf{C} = \mathbf{C}_t$). For $\mathbf{C} = \mathbf{C}_b = \mathbf{C}_t$, computing the LR (1.39) yields the following detector statistic

$$T_{\text{MF}} = k(\boldsymbol{\mu}_t - \boldsymbol{\mu}_b)^t \mathbf{C}^{-1}(\mathbf{x} - \boldsymbol{\mu}_b) \tag{2.59}$$

where $k$ is a normalization constant. (2.59) represents the test statistic of the (linear) MF for model (2.58). The MF detector requires the mean vectors and the covariance matrices under the two hypotheses of the model to be known. When not available, these quantities need to be estimated. In this case, for known target spectral signatures, and for background mean and covariance matrix estimated by ML, one refer to AMF detectors [Robey 1992].

A more general theory for hypothesis testing based techniques for target detection is the *Matched Subspace* (MS), introduced by [Scharf 1994]. In this framework, a $N-$dimensional observation vector $\mathbf{x}$ (i.e. a pixel vector in Hyperspectral Imaging) is modeled as

$$\begin{cases} \mathcal{H}_0 : \mathbf{x} = \mathbf{B}\boldsymbol{\zeta} + \boldsymbol{\epsilon}, \\ \mathcal{H}_1 : \mathbf{x} = \mathbf{T}\boldsymbol{\theta} + \mathbf{B}\boldsymbol{\zeta} + \boldsymbol{\epsilon}, \end{cases} \tag{2.60}$$

where $\mathbf{T} \in \mathbb{R}^{N \times t}$ and $\mathbf{B} \in \mathbb{R}^{N \times b}$, with $b < N - t$, are two known matrices whose columns span the target spectra $\langle \mathbf{T} \rangle$ and the background spectra $\langle \mathbf{B} \rangle$ subspaces respectively, $\boldsymbol{\theta} \in \mathbb{R}^t$ and $\boldsymbol{\zeta} \in \mathbb{R}^b$ are the associated unknown coefficient vectors, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$ is an $N$-component Gaussian noise vector. In the most simple case, the covariance matrix $\mathbf{C}$ is known. Other detection scenarios (e.g. known signal in subspace with unknown noise level) are also studied in [Scharf 1994].

The target and background subspaces are assumed to be linearly independent in the sense that no signal in $\langle \mathbf{T} \rangle$ can be obtained as a linear combination of vectors in $\langle \mathbf{B} \rangle$ and vice-versa. According to this, the goal of a MS detector is to find out whether the observation $\mathbf{x}$ belongs to the target or to the background subspace. The authors perform subspace detection by using the GLR test for model (2.60). In particular, the GLR test is derived and its properties studied for four different detection problems, which take into account the complete or partial specification of the tested signal and the known or unknown noise level. These scenarios resume the characteristics of the majority of practical detection problems in time series analysis and multi-sensor array processing, and can be indeed applied to Hyperspectral Imaging assuming $\mathbf{T}$ and $\mathbf{B}$ to be known.

In [Kraut 2001], MS detectors are adapted to the case of noise covariance matrix $\mathbf{C}$ with unknown structure. Similarly to MS detectors, *Adaptive Subspace* (AS) detectors are designed using the GLR test for the four detection problems mentioned above, estimating the covariance matrix $\mathbf{C}$ by a given set of training data. An application of AS detectors to pixel and sub-pixel target detection (with known spectral signature) in hyperspectral images is proposed in [Manolakis 2001].

The same idea of [Haupt 2007], described in previous section, appears also in [Theiler 2012], applied this time in order to perform gaseous plumes detection in noisy hyperspectral images.

Numerous other works exist. We shall restrict to cite the sparsity-based algorithm for target detection developed in the first part of [Chen 2011b] because it uses sparse models and techniques connected to (but different from) what we propose in chapter 5. In this study, the model used to describe data is somewhat similar to the one defined in (2.60) for MS detectors. Each observation $\mathbf{x}$ is modeled using a structured dictionary $\mathbf{R} = [\mathbf{R}_b \ \mathbf{R}_t]$ consisting in the concatenation of a set of background ($\mathbf{R}_b$) and training ($\mathbf{R}_t$) samples

$$\mathbf{x} = \mathbf{R}\boldsymbol{\alpha} = [\mathbf{R}_b \ \mathbf{R}_t] \begin{bmatrix} \boldsymbol{\alpha}_b \\ \boldsymbol{\alpha}_t \end{bmatrix} = \mathbf{R}_b\boldsymbol{\alpha}_b + \mathbf{R}_t\boldsymbol{\alpha}_t. \tag{2.61}$$

Here, $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_b \ \boldsymbol{\alpha}_t]^t$ represents the corresponding unknown vector of sparse coefficients. After recovering the $\hat{\boldsymbol{\alpha}} = [\hat{\boldsymbol{\alpha}}_b \ \hat{\boldsymbol{\alpha}}_t]^t$ sparse vector using greedy pursuit techniques (see section 6.2), the background $r_b(\mathbf{x}) = ||\mathbf{x} - \mathbf{R}_b\hat{\boldsymbol{\alpha}}_b||_2$ and target $r_t(\mathbf{x}) = ||\mathbf{x} - \mathbf{R}_t\hat{\boldsymbol{\alpha}}_t||_2$ residuals are computed and the following detector' statistic is established

$$T(\mathbf{x}) = r_b(\mathbf{x}) - r_t(\mathbf{x}). \tag{2.62}$$

If $T(\mathbf{x}) > \xi$, with $\xi$ a prescribed threshold, then $\mathbf{x}$ is labeled as a target pixel.

#### 2.4.3.2 Spatio-spectral based detection techniques

When applied to Hyperspectral Imaging, the approaches described above perform detection for each pixel in the image independently, without taking into account eventual spectral

correlations existing between pixels in a same spatial neighborhood. A brief survey on some works that propose spatio-spectral based detection techniques is proposed below. Contributions for the particular case of 3D spatio-spectral model-based techniques will be proposed in section 6.3.

In the second part of the paper [Chen 2011b], the authors suggest to exploit inter-pixel correlation to perform detection of targets containing multiple pixels. To improve detection performances, a spatial smoothing constraint is introduced in the detector to force spectral similarity between spatially correlated pixels.

An alternative way to exploit spatio-spectral dependences between pixels has been presented in [Chen 2011a]. Based on the same detector of (2.62), a joint sparsity model is introduced through which pixels in a same small neighborhood are approximated by a sparse linear combination of a few common dictionary atoms, using weighting coefficient vectors with same support but different values of the non-zero components.

In [Schweizer 2001], voxel's inter-correlation is exploited performing a ML anomaly detector based on spatio-spectral Gauss-Markov Random Field modeling.

A "top-to-bottom" $l_1-$minimization algorithm for the detection of non-overlapping (spatially sparse) point sources in additive noise is introduced in [Herranz 2010]. After filtering the entire image with matched filter, point sources detection is performed downwards from the brightest (top) to the faintest (bottom) picks in the filtered image.

In [Bourguignon 2012], a multi-step approach for astrophysical hyperspectral data processing is presented. The whole hyperspectral cube is considered as a collection of spectra in which data are sparsely modeled as a superposition of a line spectrum and a continuum spectrum (modeled in the DCT domain). After performing data restoration from noisy observations by using a PLS estimator with $l_1-$norm penalization, an "object detection strategy" is introduced. Based on the estimated data, pixels' aggregation into objects is performed connecting pixels with similar spectral characteristics as belonging to the same object. More precisely, two neighbors spectra are aggregated if they share at least one detected spectral line at the same wavelength. In the last step, estimation is addressed again. For each object found, a separable model is considered aiming at representing the corresponding part of the data cube as a unique spectrum weigthed by spatial amplitude map.

In line with the works listed above, the spatio-spectral detection techniques presented in Part II-B of this manuscript and then applied to astrophysical hyperspectral data in Part III, will be based on sparse data models. These models will account for known noise characteristics and will take advantage from adapted redundant dictionaries, designed according to the characteristics of the sought target signatures. Moreover, the proposed approaches will account for instrument's specificities (see chapter 5 and chapter 6) not only in the spectral domain (as in the works [Bourguignon 2012]), but also in the spatial domain. This will substantially help in improving detection performances toward extremely faint objects (see chapter 8).

## 2.5 Conclusion

We have exposed some of the main basic concepts of sparse representations, sparse estimation and sparse detection. By definition, a signal is called sparse when most of its coefficients are (possibly approximatively) null. Many signals are not sparse in their nature but show up to be sparse in the domain of an appropriate space of representation. In many sparsity-based approaches, these spaces are represented by synthesis dictionaries, which form a large catalog containing precise signal's features (atoms). Dictionaries can be pre-specified or learned. In the context of the presented work, the K-SVD dictionary learning approach (see Appendix B.3), will be used in chapter 8.

In the estimation (denoising) framework, we presented the Soft-, Hard-, and SCAD thresholding operators derived as solutions of Penalized Least Square problems. We presented their connection to sparsity-promoting (thresholding) properties of the Bayesian Maximum A Posteriori estimator. Such thresholding properties are taken into account in the design of the sparse signal detectors introduced in the next chapter.

Finally, a review of important previous works existing in the sparse detection literature has been proposed including the HC procedure that will be used for comparison purposes in chapter 5. The end of the chapter focuses on a quick survey of detection methods in hyperspectral imaging, which is the application domain of our work (see chapter 7).

# Part II

# Sparsity-based Detection Tests: Theory and Methods

Subpart A

# Description and study of the proposed approaches

Chapter 3

# Sparsity-based detection tests

Contents


**3.1 Introduction . . . . . 59**
**3.2 A first approach: the LRMAP detection test . . . . . 60**
**3.3 A second approach: the PDR detection test . . . . . 61**
**3.4 Comparison of the tests performances in the scalar case . . . . . 63**
3.4.1 Scalar model and test statistics . . . . . 63
3.4.2 Analytical properties of the tests . . . . . 64
3.4.3 Numerical simulations . . . . . 68
**3.5 Comparison of the tests performances in the vector case . . . . . 69**
3.5.1 $T_{\text{PDR}}$ and $T_{\text{LRMAP}}$ parameter dependency and asymptotic results . . . 69
3.5.2 On the relative power of the PDR and LRMAP tests . . . . . 70
3.5.3 Numerical simulations . . . . . 73
**3.6 Conclusion . . . . . 74**


## 3.1 Introduction

The main aim of this work is to design and study new detection tests that are powerful when the alternative is sparse (i.e., when the parameter vector $\boldsymbol{\theta}$ presents many zeros under the $\mathcal{H}_1$ hypothesis of the considered model). In order to achieve this goal, we start putting into practice all the concepts illustrated in the previous chapters.

The first thing we stress on concerns the model that we will use throughout this part. All along this chapter we will consider the following elementary composite hypothesis test:

$$\begin{cases} \mathcal{H}_0 : \mathbf{x} = \boldsymbol{\epsilon}, & \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \\ \mathcal{H}_1 : \mathbf{x} = \boldsymbol{\theta} + \boldsymbol{\epsilon}, \end{cases} \tag{3.1}$$

where a given signal $\mathbf{x}$ is supposed to be characterized by noise only ($\boldsymbol{\epsilon}$) under the $\mathcal{H}_0$ hypothesis, and by noise plus a signal component $\boldsymbol{\theta} \neq \mathbf{0}$ under $\mathcal{H}_1$. This model actually corresponds to the one described in (1.50). We assume $\mathbf{x}, \boldsymbol{\theta}$ and $\boldsymbol{\epsilon}$ to be real $N$-vectors. $\boldsymbol{\theta}$ represents an unknown deterministic vector of parameters with few nonzero components. As many applications involve non-identically distributed noise [Kay 1998a, Bourguignon 2011], $\boldsymbol{\epsilon}$ is considered as Gaussian, with zero mean and known diagonal covariance matrix $\boldsymbol{\Sigma} = \text{diag}\left[\sigma_1^2, \ldots, \sigma_N^2\right]$.

In the following sections, we present two new approaches for the detection of sparse signals: the LRMAP test, consisting in a LR test in which we inject a MAP estimate,

and the PDR test, which computes the ratio of the posterior distributions under the two hypotheses of the model. These two new methods are deeply investigated: after defining the tests for $N-$dimensional data vectors in section 3.2 and section 3.3, we open a parenthesis in section 3.4 to define and study the characteristics of the tests in the scalar case. In this framework, interesting analytical properties on the test statistics are derived. Asymptotic results and numerical insights are finally given in section 3.5. Note that, in both the scalar and vectorial cases, the detection performances of the proposed approaches are compared to that of the GLR and the BF classical methods introduced in chapter 1.

## 3.2 A first approach: the LRMAP detection test

The first approach to sparse signal detection we propose is directly inspired by the GLR test discussed in Sec. 1.3.2. The idea on which the method is based is to improve the detection performances of a GLR test by replacing the ML estimate of the unknown parameter vector $\boldsymbol{\theta}$ in the likelihood function under $\mathcal{H}_1$ with a MAP estimate, allowing so to take into account an appropriate prior knowledge on it.

We refer to this strategy as the LRMAP test, and we define it as follows

$$\text{LRMAP}(\mathbf{x}) : \frac{p(\mathbf{x} \mid \hat{\boldsymbol{\theta}}_{\text{MAP}})}{p(\mathbf{x} \mid \mathbf{0})} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma. \tag{3.2}$$

The computation of the $T_{\text{LRMAP}}$ test statistic for model (3.1) gives

$$\begin{aligned} T_{\text{LRMAP}}(\mathbf{x}) &= \frac{p(\mathbf{x} \mid \hat{\boldsymbol{\theta}}_{\text{MAP}})}{p(\mathbf{x} \mid \mathbf{0})} \\ &= \frac{\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2\sigma_i^2}\left(x_i - \hat{\theta}_{MAP_i}\right)^2\right]}{\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{1}{2\sigma_i^2} x_i^2\right)} \\ &= \prod_{i=1}^{N} \exp\left[-\frac{1}{2\sigma_i^2}\left(x_i - \hat{\theta}_{MAP_i}\right)^2 + \frac{1}{2\sigma_i^2} x_i^2\right] \\ &= \exp\left[\sum_{i=1}^{N} -\frac{1}{2\sigma_i^2}\left(\hat{\theta}_{MAP_i}^2 - 2x_i\hat{\theta}_{MAP_i}\right)\right] \end{aligned} \tag{3.3}$$

Taking the logarithm of (3.3) and inserting in it the expression of the MAP estimate computed for a Laplacian prior (2.51), we have

$$\begin{aligned} T_{\text{LRMAP}}(\mathbf{x}) &= \sum_{i=1}^{N} \left[-\frac{1}{2\sigma_i^2}\left(x_i - \text{sgn}(x_i)\frac{\sigma_i^2}{\lambda_i}\right)^2 + \frac{x_i}{\sigma_i^2}\left(x_i - \text{sgn}(x_i)\frac{\sigma_i^2}{\lambda_i}\right)\right] I\left(\frac{|x_i|}{\sigma_i} > \frac{\sigma_i}{\lambda_i}\right) \\ &= \sum_{i=1}^{N} \left(\frac{x_i^2}{\sigma_i^2} - \frac{\sigma_i^2}{\lambda_i^2}\right) I\left(\frac{|x_i|}{\sigma_i} > \frac{\sigma_i}{\lambda_i}\right), \end{aligned} \tag{3.4}$$

where $I(\cdot)$ represents the indicator function and $\gamma_{\text{LRMAP}} = 2\ln\gamma$.

Denoting $x_i/\sigma_i = u_i$ and $\sigma_i/\lambda_i = \eta_i$, the expression of the LRMAP detection test finally becomes

$$T_{\text{LRMAP}}(\mathbf{u}) \;=\; \sum_{i=1}^{N} \left(u_i^2 - \eta_i^2\right) I\left(|u_i| > \eta_i\right) \underset{H_0}{\overset{H_1}{\gtrless}} \gamma_{\text{LRMAP}}. \tag{3.5}$$

When $\hat{\boldsymbol{\theta}}_{\text{MAP}} = \mathbf{0}$, $T_{\text{LRMAP}}(\mathbf{u}) = \mathbf{0}$. According to (2.51), this happens when all the components of the vector $\boldsymbol{u}$ satisfy $-\eta_i < u_i < \eta_i$. On the contrary, when $\hat{\boldsymbol{\theta}}_{\text{MAP}} \neq \mathbf{0}$, which means that the absolute value of one or more components of $\mathbf{u}$ is above the correspondent threshold $\eta_i$, $T_{\text{LRMAP}}(\mathbf{u}) \neq \mathbf{0}$. In this case, if the $T_{\text{LRMAP}}(\mathbf{u})$ value is bigger than the $\gamma_{\text{LRMAP}}$ test threshold, the $\mathcal{H}_1$ hypothesis is accepted against $\mathcal{H}_0$. To make an example, we consider the case of a $N = 2$ dimensional data vector $\mathbf{u}$. For such a vector, we will retain the $\mathcal{H}_1$ hypothesis in the two following different situations:

1. the case in which $u_1 > \eta$ and $u_2 \leq \eta$ (or similarly, permuting the $u_1$ and $u_2$ components);
2. the case in which both components $u_1$ and $u_2$ are above the $\eta$ parameter.

In the first case, detection is claimed if

$$T_{\text{LRMAP}}(\mathbf{u}) = \left(u_i^2 - \eta_i^2\right) > \gamma_{\text{LRMAP}}, \tag{3.6}$$

with $i = 1, 2$. In the second case, we have detection if

$$T_{\text{LRMAP}}(\mathbf{u}) = \left(u_1^2 - \eta_1^2\right) + \left(u_2^2 - \eta_2^2\right) > \gamma_{\text{LRMAP}}. \tag{3.7}$$

We refer to chapter 4 for a thorough study on the characteristics of the $T_{\text{LRMAP}}$ test statistic in the $N = 2$ and $N > 2$ dimensional cases.

## 3.3 A second approach: the PDR detection test

The second sparse signal detection method we introduce in this work is called the Posterior Density Ratio (PDR) test. Differently from the line of thought followed for the construction of the LRMAP test, the basic idea behind the PDR test is to compute the ratio between the suprema of the parameter vector posterior distributions, calculated under each hypothesis of the model. Likelihood distributions are often depicted as posterior distributions with uniform priors. According to this, the PDR test can be seen as a sort of generalization of the LRMAP test: the posterior distributions under $\mathcal{H}_0$ and $\mathcal{H}_1$ are here obtained as the product of the Likelihood (in which, similarly to the LRMAP test, we inject the MAP estimate of $\boldsymbol{\theta}$) times a Laplacian-like prior distribution on the parameter vector. This detection test is not new in the literature. Introduced in the sixties by I. J. Good [Good 1965], the PDR test was also discussed in [Basu 1996], in the case of scalar parameters. The PDR test seems absent from the subsequent literature of the "sparse representation era" [Fuchs 2010], despite its good behavior for detecting sparse signals, as we shall see.

We define the PDR detection test as

$$\text{PDR}(\mathbf{x}) : \frac{\max\limits_{\boldsymbol{\theta}} p\left(\boldsymbol{\theta} \mid \mathbf{x}\right)}{p\left(\mathbf{0} \mid \mathbf{x}\right)} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma. \tag{3.8}$$

This definition leads to the test statistic

$$T_{\mathrm{PDR}}(\mathbf{x}) = \frac{p\left(\mathbf{x} \mid \hat{\boldsymbol{\theta}}_{\mathrm{MAP}}\right)\pi(\boldsymbol{\theta})}{p\left(\mathbf{x} \mid \mathbf{0}\right)\pi(\mathbf{0})}. \tag{3.9}$$

Making explicit the expression of $T_{\mathrm{PDR}}(\mathbf{x})$, using a Laplacian prior of the form seen in (2.47), we obtain

$$\begin{aligned} T_{\mathrm{PDR}}(\mathbf{x}) &= \prod_{i=1}^{N} \exp\left[-\frac{1}{2\sigma_i^2}\left(x_i - \hat{\theta}_{MAP_i}\right)^2 - \frac{\left|\hat{\theta}_{MAP_i}\right|}{\lambda_i} + \frac{1}{2\sigma_i^2}x_i^2\right] \\ &= \prod_{i=1}^{N} \exp\left(\frac{x_i\hat{\theta}_{MAP_i}}{\sigma_i^2} - \frac{\hat{\theta}_{MAP_i}^2}{2\sigma_i^2} - \frac{\left|\hat{\theta}_{MAP_i}\right|}{\lambda_i}\right). \end{aligned} \tag{3.10}$$

Taking the logarithm of (3.10) and factorizing by the quantity $\frac{\hat{\theta}_{MAP_i}}{\sigma_i^2}$, the expression of the PDR test becomes

$$T_{\mathrm{PDR}}(\mathbf{x}) = \sum_{i=1}^{N} \frac{\hat{\theta}_{MAP_i}}{\sigma_i^2}\left[x_i - \frac{\hat{\theta}_{MAP_i}}{2} - \frac{\sigma_i^2 \operatorname{sgn}\left(\hat{\theta}_{MAP_i}\right)}{\lambda_i}\right]. \tag{3.11}$$

Inserting in (3.11) the value of $\hat{\theta}_{MAP_i} \neq 0$ given in (2.51), we have

$$\begin{aligned} T_{\mathrm{PDR}}(\boldsymbol{x}) &= \sum_{i=1}^{N}\left(\frac{x_i}{\sigma_i^2} - \frac{\operatorname{sgn}(x_i)}{\lambda_i}\right)\left[x_i - \frac{1}{2}\left(x_i - \operatorname{sgn}(x_i)\frac{\sigma_i^2}{\lambda_i}\right) - \operatorname{sgn}(x_i)\frac{\sigma_i^2}{\lambda_i}\right] \\ &= \frac{1}{2}\sum_{i=1}^{N}\left(\frac{x_i}{\sigma_i^2} - \frac{\operatorname{sgn}(x_i)}{\lambda_i}\right)\left[x_i - \operatorname{sgn}(x_i)\frac{\sigma_i^2}{\lambda_i}\right] \\ &= \frac{1}{2}\sum_{i=1}^{N}\left(\frac{x_i^2}{\sigma_i^2} - \frac{2\left|x_i\right|}{\lambda_i} + \frac{\sigma_i^2}{\lambda_i^2}\right) \\ &= \frac{1}{2}\sum_{i=1}^{N}\left(\frac{\left|x_i\right|}{\sigma_i} - \frac{\sigma_i}{\lambda_i}\right)^2. \end{aligned} \tag{3.12}$$

According to (3.12) we finally write

$$T_{\mathrm{PDR}}(\mathbf{x}) = \sum_{i=1}^{N}\left(\frac{\left|x_i\right|}{\sigma_i} - \frac{\sigma_i}{\lambda_i}\right)^2 \underset{H_0}{\overset{H_1}{\gtrless}} \gamma_{\mathrm{PDR}}, \tag{3.13}$$

where $\gamma_{\mathrm{PDR}} = 2\ln\gamma$.

Similarly to the LRMAP test, denoting $x_i/\sigma_i = u_i$ and $\sigma_i/\lambda_i = \eta_i$, the PDR test finally gives

$$T_{\mathrm{PDR}}(\mathbf{u}) = \sum_{i=1}^{N}(\left|u_i\right| - \eta_i)^2 I(\left|u_i\right| > \eta_i) \underset{H_0}{\overset{H_1}{\gtrless}} \gamma_{\mathrm{PDR}}. \tag{3.14}$$

As seen for the LRMAP test in (3.5), when all the components of the vector $\mathbf{u}$ satisfy $-\eta_i < u_i < \eta_i$, $\hat{\boldsymbol{\theta}}_{\text{MAP}} = \mathbf{0}$ and, consequently, $T_{\text{PDR}}(\mathbf{u}) = \mathbf{0}$. Taking the same example as before, when a $N = 2$ component vector $\mathbf{u}$ is considered, we have detection with the PDR test when

$$T_{\text{PDR}}(\mathbf{u}) \ = (|u_i| - \eta_i)^2 \ > \ \gamma_{\text{PDR}}, \quad i = 1, 2, \tag{3.15}$$

if only one component of $\mathbf{u}$ is above $\eta_i$, while when both $u_1$ and $u_2$ are bigger than $\eta_i$ detection is claimed when

$$T_{\text{PDR}}(\mathbf{u}) \ = (|u_1| - \eta_1)^2 + (|u_2| - \eta_2)^2 > \ \gamma_{\text{PDR}}. \tag{3.16}$$

We note that both sums in the $T_{\text{PDR}}$ and $T_{\text{LRMAP}}$ test statistics involve only the largest components of the data vector $\mathbf{u}$. For this reason these tests tend under $\mathcal{H}_1$ to preferentially select those components for which $\theta_i \neq 0$. This effect is indeed seen as an advantage with respect to classical approaches such as the GLR, which uses all data components, and was advocated in [Fan 1996] to use thresholding operators for detection purposes.

We also note that $T_{\text{PDR}}$ corresponds to the soft-thresholding statistic introduced in detection by Fan [Fan 1996], with the objective of focusing the test on active components only, in order to mimic an Oracle test. The Oracle test assumes that the support $\mathcal{I} = \{i \mid \theta_i \neq 0\}$ of $\boldsymbol{\theta}$ is known, but not the corresponding amplitudes $\theta_i$, $i \in \mathcal{I}$. These amplitudes are estimated by ML, leading to estimates $\hat{\boldsymbol{\theta}}_{\text{ML}}(\mathcal{I})$. This yields to the test

$$T_{\text{Oracle}}(\mathbf{x}) = \frac{p\left(\mathbf{x} \mid \mathcal{I}, \hat{\boldsymbol{\theta}}_{\text{ML}}(\mathcal{I})\right)}{p\left(\mathbf{x} \mid \mathbf{0}\right)} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma_{\text{Oracle}}. \tag{3.17}$$

Thus, the PDR test introduced by Basu [Basu 1996] furnishes a precise framework to Fan's soft-thresholding in detection. A detailed study of the $T_{\text{LRMAP}}(\mathbf{u})$ and the $T_{\text{PDR}}(\mathbf{u})$ characteristics, and their geometrical comparison, is addressed in chapter 4.

## 3.4 Comparison of the tests performances in the scalar case

### 3.4.1 Scalar model and test statistics

As seen in previous sections, the computation of the test statistic for data vectors having independent components can be done component-wise, involving a test statistic for each one of the vector's components. Inspired by this fact, we make a short digression and begin the analysis and comparison of the detection tests introduced up to now in a very simple case: the scalar case. The insights that can be drawn from such analysis are very useful and lead inter alia to the formulation of some properties on the statistics of the considered tests.

The model we consider is the following composite hypothesis:

$$\begin{cases} \mathcal{H}_0 : x = \epsilon, & \epsilon \sim \mathcal{N}\left(0, \sigma^2\right) \\ \mathcal{H}_1 : x = \theta + \epsilon \end{cases} \tag{3.18}$$

where $\theta$ is an unknown deterministic parameter and $\epsilon$ is a Gaussian noise component with zero mean and variance $\sigma^2$. According to the definitions seen in (3.5) and (3.14), the computation of the LRMAP and PDR tests for model (3.18) yields

- **LRMAP test**:

$$T_{\text{LRMAP}}(u) = (u^2 - \eta^2) I(|u| > \eta) \underset{H_0}{\overset{H_1}{\gtrless}} \gamma_{\text{LRMAP}}; \tag{3.19}$$

- **PDR test**:

$$T_{\text{PDR}}(u) = (|u| - \eta)^2 I(|u| > \eta) \underset{H_0}{\overset{H_1}{\gtrless}} \gamma_{\text{PDR}}, \tag{3.20}$$

where, similarly to previous sections, we have noted $x/\sigma = u$ and $\sigma/\lambda = \eta$.

We compare the two approaches above with the GLR and the BF tests defined in Sec.1.3. For model (3.18), the computation of the GLR test statistic gives

- **GLR test**:

$$T_{\text{GLR}}(u) = u^2 \underset{H_0}{\overset{H_1}{\gtrless}} \gamma_{\text{GLR}}. \tag{3.21}$$

On the other hand, the definition of the BF test for the scalar model (3.18) with Gaussian likelihood and Laplacian prior leads to

- **BF test**:

$$T_{\text{BF}}(x) = e^{\frac{1}{2}\left(\frac{x^2}{\sigma^2} + \frac{\sigma^2}{\lambda^2}\right)} \left[2\cosh\left(\frac{x}{\lambda}\right) - e^{-\frac{x}{\lambda}}\Phi\left(-\frac{x}{\sigma} + \frac{\sigma}{\lambda}\right) - e^{\frac{x}{\lambda}}\Phi\left(\frac{x}{\sigma} + \frac{\sigma}{\lambda}\right)\right] \underset{H_0}{\overset{H_1}{\gtrless}} \gamma_{\text{BF}}, \tag{3.22}$$

where $\gamma_{\text{BF}} = 2\gamma\lambda/\sqrt{2\pi}$ (see Proof in Appendix B.2). Lightening the notation, similarly to what done for the other tests, we obtain

$$T_{\text{BF}}(u) = e^{\frac{1}{2}\left(u^2+\eta^2\right)} \left[2\cosh(u\eta) - e^{-u\eta}\Phi(-u+\eta) - e^{u\eta}\Phi(u+\eta)\right] \underset{H_0}{\overset{H_1}{\gtrless}} \gamma_{\text{BF}}. \tag{3.23}$$

### 3.4.2 Analytical properties of the tests

For the GLR, LRMAP, PDR and BF detection tests the probability of false alarm ($\text{P}_{\text{FA}_{\text{test}}}$, already defined in (1.13)), is equal to

$$\text{P}_{\text{FA}_{\text{test}}} = Pr\left\{T(u) > \gamma_{\text{test}} \mid \mathcal{H}_0\right\}. \tag{3.24}$$

According to this, the following Properties hold (see Fig. 3.1).

**Property 3.4.1.** *Consider a detection test characterized by its test statistic $T(u)$. If the test statistic $T(u)$ is an even function, identically 0 for $|u| < u_0$ and strictly increasing for $u \geq u_0$ with $T(u_0) \geq 0$, then the following properties hold :*

**Prop. 3.4.1.1:** $\forall \gamma_{test} \geq \gamma_{test_0}$, *with* $\gamma_{test_0} = T(u_0)$, *the test*

$$T(u) \underset{H_0}{\overset{H_1}{\gtrless}} \gamma_{test} \tag{3.25}$$

*is equivalent to a GLR with*

$$\gamma_{GLR} = \left[T^{-1}(\gamma_{test})\right]^2. \tag{3.26}$$

*Consequently, the ROC curve of* $T(u)$ *coincides with that of the GLR. Using*

$$\Phi(u) = \int_0^u \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}t^2\right) dt, \tag{3.27}$$

*and noting* $\Phi_{\chi_1^2}(\gamma)$ *the cumulative distribution function of a central* $\chi^2$ *random variable with* 1 *degree of freedom, the false alarm rate of the test is given by*

$$P_{FA_{test}}(\gamma_{test}) = 1 - \Phi_{\chi_1^2}\left(\left[T^{-1}(\gamma_{test})\right]^2\right). \tag{3.28}$$

*In particular, for* $\gamma_{test} = \gamma_{test_0} = T(u_0)$, *we have*

$$P_{FA_0} = 1 - \Phi_{\chi_1^2}\left(u_0^2\right). \tag{3.29}$$

**Prop. 3.4.1.2:** $\forall \gamma_{test} \in [0; \gamma_{test_0}]$, $P_{FA_{test}} = P_{FA_0}$.

*(see Proof in Appendix A.3)*

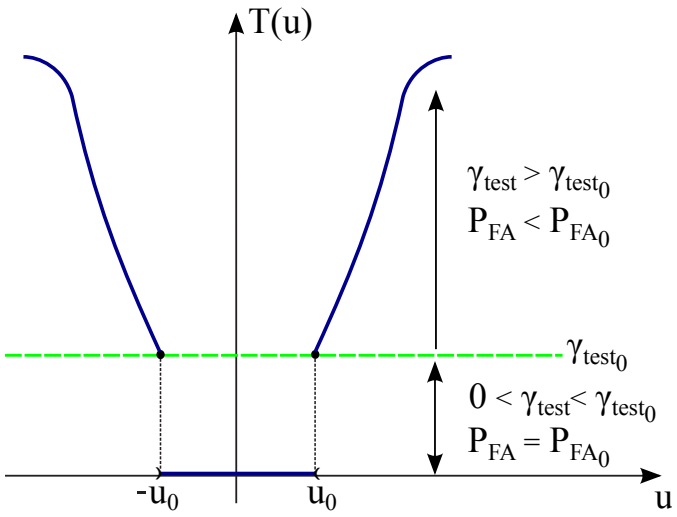



Figure 3.1: Example of a thresholded test statistic considered in Property 3.4.1, which yields the same test as the GLR for thresholds $\gamma_{\text{test}}$ above some value $\gamma_{\text{test}_0}$. If $T(u)$ is continuous, then $\gamma_{\text{test}_0} = 0$.

Property 3.4.1 holds for the LRMAP, PDR and BF test statistics defined in (3.19), (3.20) and (3.23). In fact, for $|u| \geq u_0 = \frac{\sigma}{\lambda} = \eta$ , both the $T_{\text{LRMAP}}(u)$ and $T_{\text{PDR}}(u)$ test

statistics are even and strictly increasing functions. Accordingly, the ROC curves of the LRMAP and PDR tests thus coincide with that of the GLR test, with

$$\begin{aligned} \mathrm{P}_{\mathrm{FA}_{\mathrm{LRMAP}}}(\gamma_{\mathrm{LRMAP}}) &= 1-\Phi_{\chi_1^2}\left(T^{-1}(\gamma_{\mathrm{LRMAP}})^2\right) \\ &= 1-\Phi_{\chi_1^2}\left(\gamma_{\mathrm{LRMAP}}+\eta^2\right), \end{aligned} \tag{3.30}$$

for the LRMAP test and

$$\begin{aligned} \mathrm{P}_{\mathrm{FA}_{\mathrm{PDR}}}(\gamma_{\mathrm{PDR}}) &= 1-\Phi_{\chi_1^2}\left(T^{-1}(\gamma_{\mathrm{PDR}})^2\right) \\ &= 1-\Phi_{\chi_1^2}\left(\left[\sqrt{\gamma_{\mathrm{PDR}}}+\eta\right]^2\right) \end{aligned} \tag{3.31}$$

for the PDR test. For both tests, the value of $\mathrm{P}_{\mathrm{FA}0}$ is equal to

$$\begin{aligned} \mathrm{P}_{\mathrm{FA}_0} &= 1-\Phi_{\chi_1^2}\left(u_0^2\right) \\ &= 1-\Phi_{\chi_1^2}\left(\eta^2\right). \end{aligned} \tag{3.32}$$

Since the probability of detection of the GLR is independent of the hyper-parameter $\eta$, selecting a particular value of $\eta$ for the PDR is equivalent to state a particular compromise detection versus false alarm of the GLR. Moreover, since the value $u_0 = \eta$ is the same for both the PDR and the LRMAP, the ROC curves of the two tests correspond for the same range of false alarm.

Concerning the BF, the statistic function of this test is clearly an even function of $x$. The following property shows that, for a more general class of priors, it is also a strictly increasing function of $x$ for $x \geq x_0 = 0$ and that, consequently, the BF test amounts here to a GLR.

**Property 3.4.2.** *For a Gaussian likelihood and even prior $\pi(\theta)$, the Bayes Factor test is equivalent to a GLR test.*

**Proof of Property 3.4.2:** First we show that $\mathrm{BF}(x)$ is an even function of $x$:

$$\mathrm{BF}(x) = \int_{-\infty}^{0} e^{\frac{x\theta}{\sigma^2}-\frac{\theta^2}{2\sigma^2}}\pi(\theta)\,d\theta + \int_{0}^{+\infty} e^{\frac{x\theta}{\sigma^2}-\frac{\theta^2}{2\sigma^2}}\pi(\theta)\,d\theta$$

Assuming $\theta^t = -\theta$ and consequently $d\theta^t = -d\theta$, we obtain

$$\mathrm{BF}(x) = \int_{0}^{\infty} 2\cosh\left(\frac{x\theta}{\sigma^2}\right) e^{-\frac{\theta^2}{2\sigma^2}}\pi(\theta)\,d\theta,$$

which is even in $x$ because cosh is even.
Secondly, we show that BF is an increasing function of $x$:

$$\frac{\partial \mathrm{BF}}{\partial x} = \int_{-\infty}^{+\infty} \frac{\partial}{\partial x} e^{\frac{x\theta}{\sigma^2}} e^{-\frac{\theta^2}{2\sigma^2}}\pi(\theta)\,d\theta$$

and after some simple algebra we obtain

$$\frac{\partial \mathrm{BF}}{\partial x} = \int_{0}^{\infty} 2\frac{\theta}{\sigma^2}\sinh\left(\frac{x\theta}{\sigma^2}\right) e^{-\frac{\theta^2}{2\sigma^2}}\pi(\theta)\,d\theta,$$

which is strictly positive for $x > 0$. Consequently, $\mathrm{BF}(x)$ is an increasing function for $x > 0$. For any even prior and Gaussian likelihood, $\mathrm{BF}(x)$ obeys the conditions of Property 3.4.1. The BF test is thus equivalent to a GLR for $P_{FA} < P_{FA_0} = 1 - \Phi_{\chi^2_1}(0) = 1$.

The equivalence in the scalar case of the GLR, PDR, LRMAP and BF tests, illustrated for a Laplacian prior, can be finally generalized to the case of Generalized Gaussian priors of parameter $0 < p < 1$ (GG):

**Property 3.4.3.** *For generalized Gaussian prior functions with $0 < p < 1$, the PDR, the LRMAP and the BF tests defined in (3.19), (3.20) and (3.23) are equivalent to a GLR.*

**Proof of Property 3.4.3:** Consider the generalized zero mean Gaussian distribution

$$GG\left(\theta, p\right) = \frac{p}{2s\Gamma\left(\frac{1}{s}\right)} \exp\left(-\frac{|\theta|^p}{s^p}\right), \tag{3.33}$$

where $\Gamma$ represents the Gamma function, $s$ is a real positive parameter, and $p \in ]0;1[$. In the case of the Gaussian likelihood function, such priors are known to produce MAP estimates $\hat{\theta}_{\mathrm{MAP}}(x)$ of the $\theta$ parameter, which are thresholding functions of $x$ [Moulin 1999, Antoniadis 2001]. In other words, similarly to what seen in Sec. 2.3.3, there exists $x_0$ such that

$$\hat{\theta}_{\mathrm{MAP}}(x) = \begin{cases} 0, & \forall x : |x| < x_0, \\ \text{is increasing}, & \forall x : x \geq x_0. \end{cases} \tag{3.34}$$

Moreover, $\hat{\theta}_{\mathrm{MAP}}(x)$ is antisymmetric $\left(\hat{\theta}_{\mathrm{MAP}}\left(-x\right) = -\hat{\theta}_{\mathrm{MAP}}\left(-x\right)\right)$, and

$$\hat{\theta}_{\mathrm{MAP}}\left(x\right) = \mathrm{sgn}\left(x\right)\left(|x| - \zeta\left(x\right)\right), \tag{3.35}$$

where $\zeta\left(x\right)$ is a positive bias with $\lim\limits_{x\to\infty} \zeta(x) = 0$.

From the log-version of the $T_{\mathrm{LRMAP}}$ expression seen in (3.3), considered for only one scalar component $\hat{\theta}_{\mathrm{MAP}}$, we have

$$T_{\mathrm{LRMAP}}(x) = \hat{\theta}_{\mathrm{MAP}}\left(\frac{x}{\sigma^2} - \frac{\hat{\theta}_{\mathrm{MAP}}}{2\sigma^2}\right). \tag{3.36}$$

It is clear that $T_{\mathrm{LRMAP}}(x)$ is even, because $\hat{\theta}_{\mathrm{MAP}}(x)$ is antisymmetric. One also sees that since $\hat{\theta}_{\mathrm{MAP}}(x)$ is a thresholding function with threshold $x_0$, the same will hold for $T_{\mathrm{LRMAP}}(x)$. Denoting by $'$ the partial derivative with respect to $x$, we have

$$\begin{aligned} T'_{\mathrm{LRMAP}}(x) = \frac{\partial T_{\mathrm{LRMAP}}}{\partial x} &= \frac{\hat{\theta}_{\mathrm{MAP}}}{\sigma^2} + \frac{\hat{\theta}'_{\mathrm{MAP}}}{\sigma^2}(x - \hat{\theta}_{\mathrm{MAP}}) \\ &= \frac{\hat{\theta}_{\mathrm{MAP}}}{\sigma^2} + \frac{\hat{\theta}'_{\mathrm{MAP}}}{\sigma^2}\zeta, \end{aligned} \tag{3.37}$$

which is positive for $x > x_0$. Consequently, by Property 3.4.1, $T_{\mathrm{LRMAP}}(x)$ is equivalent to a GLR for any GG with $0 < p < 1$.

The same results hold for the PDR for which, similarly to the expression in (3.11), we can write

$$T_{\mathrm{PDR}}(x) = \hat{\theta}_{\mathrm{MAP}} \left( \frac{x}{\sigma^2} - \frac{\hat{\theta}_{\mathrm{MAP}}}{2\sigma^2} \right) - \frac{\left|\hat{\theta}_{\mathrm{MAP}}\right|^p}{s^p}. \tag{3.38}$$

First, for the same reason as the LRMAP test statistic, $T_{\mathrm{PDR}}$ is an even function. Second, $T_{\mathrm{PDR}}(x)$ is clearly zero $\forall x < x_0$, and the derivative $T'_{\mathrm{PDR}}(x)$ is positive for $x \geq x_0$. So by Property 3.4.1 the PDR yields a GLR, which completes the proof. As far as the BF is concerned, the GG is an even function, so the results of Prop. 3.4.2 hold also for GG priors with $0 < p < 1$.

### 3.4.3 Numerical simulations

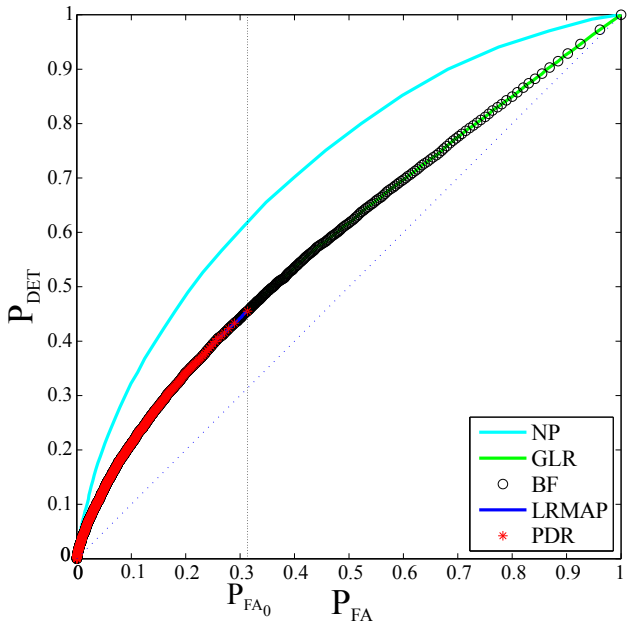



Figure 3.2: NP, GLR, LRMAP and PDR ROC curves in the case of a scalar parameter $\theta$ = 0.8. Here, $\sigma = \lambda = 1$. The dotted vertical line highlights the value of the maximal $\mathrm{P_{FA}}$, $\mathrm{P_{FA0}} \approx 0.32$, for the PDR and the LRMAP.

The results achieved in the previous section are confirmed by the numerical simulations presented below. As shown in Figure 3.2, we have calculated the ROC curves of the GLR, LRMAP and PDR tests in the scalar case for $\theta = 0.8$, $\sigma = 1$ and $\lambda = 1$. The NP test (or LR test (1.40), for which the alternative $\theta$ is known) is reported as a reference. Except for the ROC of the NP, all the curves overlap (Properties 3.4.1.1 and 3.4.1.2). The curves of the NP (in cyan), GLR (in green) and the BF (black circles) cover the whole range of $\mathrm{P_{FA}}$. As expected, the LRMAP (blue continuous line) and the PDR (red stars) curves correspond to that of the GLR for $\mathrm{P_{FA}} \leq \mathrm{P_{FA0}}$. In this case, the maximal value of $\mathrm{P_{FA}}$ for the PDR and the LRMAP is equal to $\mathrm{P_{FA0}} \approx 0.32$, as highlighted in Figure 3.2 by a dotted vertical

line. We reserve a more complete discussion on the maximal false alarm rate of the tests in next section.

## 3.5 Comparison of the tests performances in the vector case

### 3.5.1 $T_{\mathbf{PDR}}$ and $T_{\mathbf{LRMAP}}$ parameter dependency and asymptotic results

We now come back to the analysis of the $T_{\mathrm{PDR}}$ and $T_{\mathrm{LRMAP}}$ test statistics for the vector model defined in (3.1). As clearly visible from expressions (3.5) and (3.14), in this case the LRMAP and PDR tests are both characterized by positive test statistics, which are null if all components $|u_i|$ fall below $\eta_i$. Hence, when the test threshold $\gamma$ is set to 0, a single component $|u_i|$ above $\eta_i$ under $\mathcal{H}_0$ is sufficient to cause a false alarm (FA). In this case, if one wishes all components to have the same probability to cause a FA, the $\eta_i$ should be made equal to some value, say $\eta$.

Note that, if $\gamma \neq 0$, the probability that one particular component $|u_i|$ causes a FA is not the probability that $|u_i| > \eta_i$, see subsection 4.2.2 and Figure 4.1. Hence, $\eta$ generally does not completely characterize the per-component false alarm rate.

We pause to note that the two detection tests,

$$T_{\mathrm{LRMAP}}(\mathbf{u}, \eta) \underset{H_0}{\overset{H_1}{\gtrless}} \gamma \qquad \text{and} \qquad T_{\mathrm{PDR}}(\mathbf{u}, \eta) \underset{H_0}{\overset{H_1}{\gtrless}} \gamma,$$

depend on the two parameters, $\eta$ and $\gamma$. This dependency is different for the two tests, and makes the analysis and comparison of the tests quite involved. For instance, different pairs of values $(\eta, \gamma)$ can yield the same $\mathrm{P_{FA}}$, as shown in Figure 3.3. In these simulations, the empirical distributions of $T_{\mathrm{LRMAP}}$ (on the left) and $T_{\mathrm{PDR}}$ (on the right) under $\mathcal{H}_0$ are obtained for $10^4$ realizations of a random $N = 1000$ component standard normal vector, and presented for three different values of the $\eta$ parameter, respectively $\eta = 1,\ 2.7,\ 3$. The two first figures of the left column illustrate how two different pairs of values ($\eta = 1$,$\gamma_1 = 503$: *red*) and ($\eta = 2.7$,$\gamma_2 = 15.8$: *green*) may lead to the same $\mathrm{P_{FA}}$ ($\mathrm{P_{FA}} = 0.31$, here).

Unfortunately, the analytical computation of the probability of false alarm,

$$\mathrm{P_{FA}}(\eta, \gamma) = Pr(T(\eta) > \gamma; \mathcal{H}_0), \tag{3.39}$$

turns out to be intractable even in the framework of the most simple model (3.1). Looking at Figure 3.3, while for small values of $\eta$ (i.e. $\eta = 1$, here) the two test statistics distributions tend to be Gaussian, as the $\eta$ value is increased the $T_{\mathrm{PDR}}$ and $T_{\mathrm{LRMAP}}$ distributions suddenly become very difficult to analyse. For instance, the distribution $\hat{p}_{T_{\mathrm{LRMAP}}}(t)$ of $T_{\mathrm{LRMAP}}$ under $\mathcal{H}_0$ is the convolution of $N$ $\chi^2$ distributions, truncated on $[\eta, +\infty[$ and shifted back to zero. This implies the presence of a probability mass at 0 of value equal to

$$Pr(0 \leq \chi^2_N \leq \eta) = [2\Phi(\eta) - 1]^N, \tag{3.40}$$

followed by an (analytically unknown) tail.

Similarly to the work of Fan [Fan 1996] for $T_{\mathrm{PDR}}$, an asymptotic approximation of $T_{\mathrm{LRMAP}}$ under $\mathcal{H}_0$ can nevertheless be obtained for $N \to \infty$. However, the Gaussian approximation of the LRMAP test statistic suffers from slow convergence (see Figure 3.3)

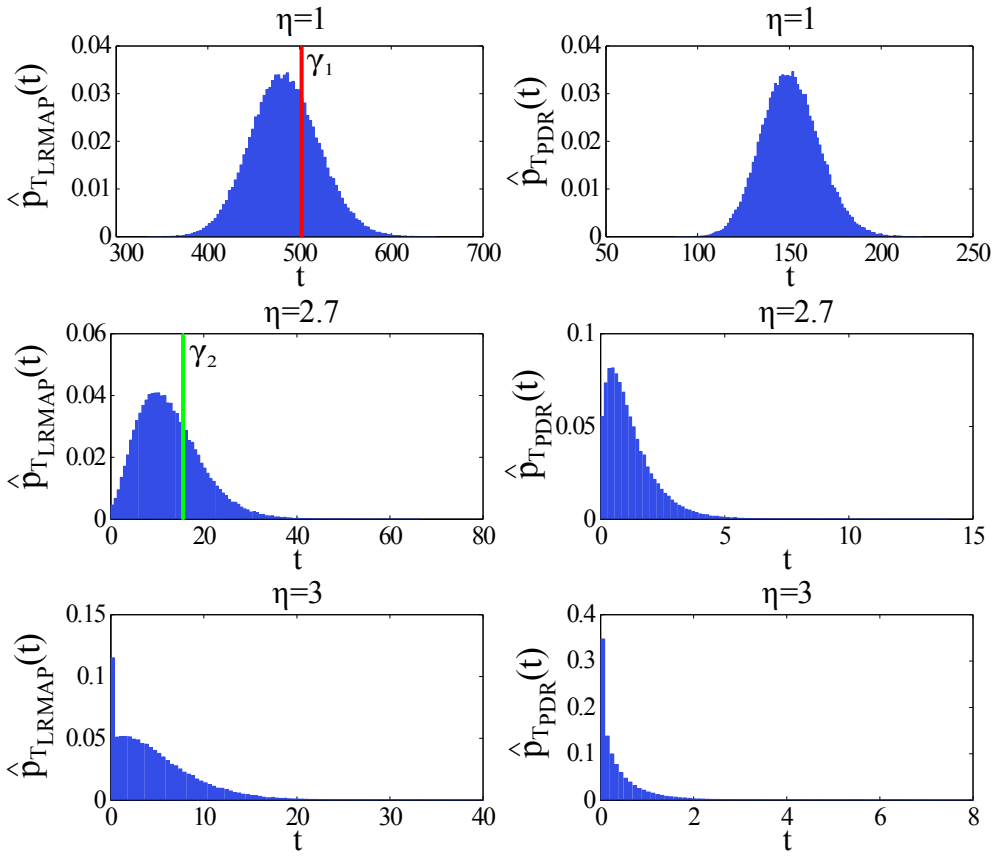



Figure 3.3: Empirical distributions of the test statistics $T_{\mathrm{LRMAP}}$ (left column) and $T_{\mathrm{PDR}}$ (right column) under $\mathcal{H}_0$ for a random $N = 1000$ component vector and different values of the $\eta$ parameter. Top row: $\eta = 1$, middle row: $\eta = 2.7$, bottom row: $\eta = 3$. The two first figures in the left column show how two different values of the couple $(\eta, \gamma)$ can generate the same false alarm rate. In particular, ($\eta = 1$,$\gamma_1 = 503$: *red vertical line*) and ($\eta = 2.7$,$\gamma_2 = 15.8$: *green vertical line*) lead to the same $\mathrm{P_{FA}}$ ($\mathrm{P_{FA}} = 0.31$, here).

and does not allow a reliable computation of the $\mathrm{P_{FA}}$, as it has been emphasized in [Fan 1996] also for the PDR. We thus propose this approximation in Appendix B.4, but we will not pursue asymptotic results in the rest of the manuscript.

### 3.5.2 On the relative power of the PDR and LRMAP tests

The beginning of this section focuses on two limit cases : $\eta \to 0$ and $\gamma \to 0$. Comparing the two expressions in (3.5) and (3.14) we have:

$$T_{\mathrm{LRMAP}} = T_{\mathrm{PDR}} + 2\eta \sum_{i=1}^{N} (|u_i| - \eta)\, I(|u_i| > \eta). \tag{3.41}$$

From this relation we can show that $T_{\text{LRMAP}}$ converges to $T_{\text{PDR}}$ in the sense of the Mean Squared Error (MSE) when $\eta = \frac{\sigma_i}{\lambda_i} \to 0$. We define the error between the two tests statistics as

$$T_{\text{LRMAP}} - T_{\text{PDR}} = 2\eta \sum_{i=1}^{N} (|u_i| - \eta)\, I(|u_i| > \eta). \tag{3.42}$$

Taking the mean $\mathbb{E}(\cdot)$ of the square value of (3.42) we have, for data vector components $|u_i| > \eta$,

$$\mathbb{E}\left[\left(2\eta \sum_{i=1}^{N} (|u_i| - \eta)\right)^2\right] = 4\eta^2 \mathbb{E}\left[\left(\sum_{i=1}^{N} (|u_i| - \eta)\right)^2\right]. \tag{3.43}$$

Using the Cauchy-Schwartz inequality, we can write that

$$\begin{aligned} 4\eta^2 \mathbb{E}\left[\left(\sum_{i=1}^{N} (|u_i| - \eta)\right)^2\right] &\leq 4\eta^2 n \mathbb{E}\left[\sum_{i=1}^{N} (|u_i| - \eta)^2\right] \\ &\leq 4\eta^2 n \sum_{i=1}^{N} \mathbb{E}\left[u_i^2 + \eta^2 - 2\eta |u_i|\right] \\ &\leq 4\eta^2 n \sum_{i=1}^{N} \left(\mathbb{E}\left[u_i^2\right] + \eta^2 - 2\eta \mathbb{E}\left[|u_i|\right]\right). \end{aligned} \tag{3.44}$$

Since $\mathbb{E}\left[u_i^2\right]$ and $\mathbb{E}\left[|u_i|\right]$ are finite quantities ($< \infty$), the second member in (3.44) tends to zero as soon as $\eta \to 0$, showing the convergence of $T_{\text{LRMAP}}$ to $T_{\text{PDR}}$ in the MSE sense.

Having $\eta \to 0$ corresponds to the following cases:

$$\begin{cases} \sigma_i \to 0, & \text{case of high signal-to-noise ratio} \\ \lambda_i \to \infty, & \text{case of uniform prior distribution.} \end{cases} \tag{3.45}$$

In such cases, the two tests distributions tend under $\mathcal{H}_0$ to a $\chi^2_N$ and are equivalent to the test of the GLR. These cases are indeed of limited interest for sparse signals since when $\sigma_i \to 0$ there is no difficulty to locate the support of $\boldsymbol{\theta}$, and the GLR manages to do so as well as the other tests. On the other hand, setting $\lambda_i$ to a large value tends to remove the thresholding effect, leading again to the GLR, that is, a global, non adaptive, energy detector.

Within these limits, both tests always differ from the GLR. Another remarkable case of equivalence between the PDR and the LRMAP tests occurs when, for a given value of $\eta \neq 0$, $\gamma$ is set to 0. Since both test statistics are positive, detection is claimed for the two tests as soon as at least one weighted component $|u_i|$ is larger than $\eta$. At $\gamma = 0$ the PDR and the LRMAP correspond to the same test, which depends on the sole tuning of the $\eta$ parameter. This allows to easily compute the maximal probability of false alarm noted ($\text{P}_{\text{FA0}}$). Since this probability corresponds to the probability to have a strictly positive test statistic under the $\mathcal{H}_0$ hypothesis, $\text{P}_{\text{FA0}}$ is indeed the probability to have at least one component $|u_i|$ above the $\eta$ threshold

$$\text{P}_{\text{FA0}} = Pr(T(\eta) > 0 | \mathcal{H}_0) = Pr(\exists i : |u_i| > \eta; \mathcal{H}_0). \tag{3.46}$$

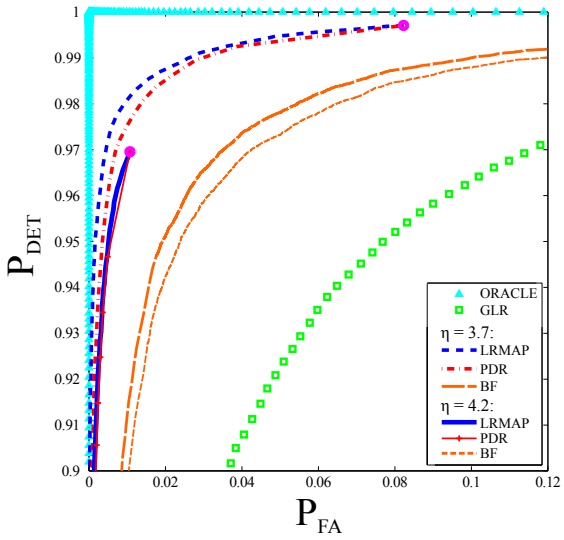



Figure 3.4: PDR, LRMAP, BF and GLR ROC curves for $\eta = 3.7$ (blue dashed line for the LRMAP, red dashed-dotted line for the PDR, orange long-dashed line for the BF) and 4.2 (blue continuous line for the LRMAP, red ticked line for the PDR, orange short-dashed line for the BF), considering a $N = 400$ component vector $\boldsymbol{\theta}$ of which only 10 are different from zero, with linearly increasing values in $[0.5; 5]$. The Oracle test is reported as a reference (cyan triangles). Magenta points correspond to the maximal probability of false alarm and probability of detection at a given $\eta$. The coordinates of those points are written in closed form in (3.48) and (3.49).

Since the probability to have all components $|u_i| < \eta$ is equal to

$$Pr(|u_i| < \eta \; \forall i) = (2\Phi(\eta) - 1)^N, \tag{3.47}$$

$\mathrm{P}_{\mathrm{FA0}}$ is given by

$$\begin{aligned} \mathrm{P}_{\mathrm{FA0}} &= 1 - Pr(|u_i| < \eta \; \forall i) \\ &= 1 - (2\Phi(\eta) - 1)^N. \end{aligned} \tag{3.48}$$

In the same way we obtain the maximal probability of detection for $\gamma = 0$:

$$\mathrm{P}_{\mathrm{DET0}}(\boldsymbol{\theta}) = Pr(T(\eta) > 0; \mathcal{H}_1) = 1 - \prod_{i=1}^{N} [\Phi(\eta - \theta_i) + \Phi(\eta + \theta_i) - 1]. \tag{3.49}$$

When the threshold $\gamma$ is set to 0, the tests are thus strictly equivalent and precisely characterized by (3.48) and (3.49). Figure 3.4 shows the Receiver Operating Characteristic (ROC) of the PDR, LRMAP, BF and GLR tests, plotted taking a $N = 400$ component vector $\boldsymbol{\theta}$ of which only 10 are different from zero, with linearly increasing values in $[0.5; 5]$. Two values of the $\eta$ threshold are considered, precisely $\eta = 3.7$ (blue dashed line for the

LRMAP, red dashed-dotted line for the PDR, orange long-dashed line for the BF) and 4.2 (blue continuous line for the LRMAP, red ticked line for the PDR, orange short-dashed line for the BF). The computation of the BF test statistic in the vector case is the product of the scalar test statistic (3.23) for each one of the $N$ components of the considered data vector

$$T_{\mathrm{BF}}(\mathbf{u}) = \prod_{i=1}^{N} e^{\frac{1}{2}\left(u_i^2+\eta_i^2\right)} \left[2\cosh\left(u_i\eta_i\right) - e^{-u_i\eta_i}\Phi\left(-u_i+\eta_i\right) - e^{u_i\eta_i}\Phi\left(u_i+\eta_i\right)\right] \underset{H_0}{\overset{H_1}{\gtrless}} \gamma_{\mathrm{BF}}. \tag{3.50}$$

As both the ML and the MMSE estimates of $\boldsymbol{\theta}$ are in general not sparse (see chapter 2), the GLR and the BF are not optimal here. For instance, as previously mentioned in subsection 1.3.3, for any considered data vector $\mathbf{u}$, the BF can be seen as the mean of the NP test over all the values of the $\boldsymbol{\theta}$ parameter. We note that while always outperforming the BF and the GLR test (green squares), the LRMAP and the PDR tests generally differ, in line with the considered values of the $\eta$ threshold. Consequently, different values of $\mathrm{P_{FA0}}$ and $\mathrm{P_{DET0}}$ are reached for different $\eta$. These values correspond to the coordinates of the two magenta points in the figure. The Oracle test is here reported as a reference and plotted in cyan (triangles).

### 3.5.3 Numerical simulations

We finally show by numerical simulations that, depending on the parameters values, the relative performances of the PDR and the LRMAP tests can be inverted. Figure 3.5 compares

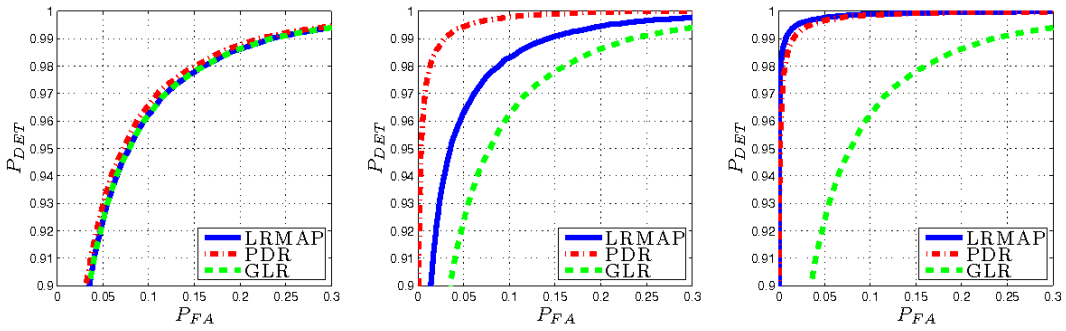



Figure 3.5: LRMAP (in blue, continuous line) and PDR (in red, dash-dotted line) ROC curves for 3 different values of $\eta$ and for the same parameter vector $\boldsymbol{\theta}$ (mean over $5 \cdot 10^4$ realizations) of $N = 400$ components (see text). For $\eta = 0.05$ (on the left), the two tests tend to coincide; for $\eta = 1$ (center), PDR outperforms the LRMAP while for $\eta = 3.3$ (on the right) it is the opposite. The ROC curve of the GLR (independent of $\eta$) is plotted in dashed green.

the ROC curves of the LRMAP, the PDR and the GLR tests for three different values of the $\eta$ parameter, respectively $\eta = 0.05$ (left figure), $\eta = 1$ (center) and $\eta = 3.3$ (right figure). In these figures, the ROCs are shown for the same $\boldsymbol{\theta}$ vector as in Figure 3.4. Since $\eta$ is fixed in each figure, different values of $\mathrm{P_{FA}}$ are obtained by varying $\gamma$. The left figure, which corresponds to $\eta = 0.05$, shows that when $\eta \to 0$, the LRMAP and PDR tests are equivalent and

tend to the GLR. This interesting case of equivalence of the tests was discussed in further detail above. Conversely, for two other values of $\eta$ (center and right figures), the relative performances of the LRMAP and the PDR are inverted (while remaining above that of the GLR). This is due to the strong but different parameters dependency of the two tests that prevents to define (easily) general conditions for which, given a set of parameters, one test is better adapted than the other. Since classical tools based on analytical distributions and Monte Carlo (MC) simulations somewhat fail to address the question of how the two tests compare, the next chapter tackles the problem differently, by a geometrical analysis. This allows a structural study of the detection regions of the tests.

## 3.6 Conclusion

We have introduced two new approaches to sparse signals detection: the LRMAP and the PDR detection tests. The two tests have been defined for a simple vector model and have been deeply analyzed. Taking as references the two classical GLR and the BF tests, already introduced in chapter 1, a comparison has been held in the case of a scalar parameter first and then in the vector case. In the former situation, interesting properties on the tests statistic functions have been derived. For example, we have shown that for a scalar parameter $\theta$ the tests always behave like a GLR test. In the latter situation of a parameter vector $\boldsymbol{\theta}$, we have focused our attention on the dependency of the LRMAP and the PDR tests on the two parameters $\eta$ and $\gamma$. We have seen that due to this dependency, the analytical computation of the tests' $\mathrm{P_{FA}}$ becomes intractable even for the simple vector model considered in this chapter. For this reason we have taken into account some limit cases: fixing one of the two parameters allows in fact to easily implement the tests, and to univocally set the desired value of $\mathrm{P_{FA}}$. We have also shown that while always outperforming the GLR and the BF tests, the parameter's values strongly influence the relative behavior of the LRMAP and PDR tests. With the aim to go in further detail with this question and to specify the conditions under which one test behaves better than the other for sparse signal detection, next section tackles the problem differently, suggesting a geometrical analysis for the structural study of the detection surfaces of the tests.

Chapter 4

# Geometrical study and comparison of the PDR and LRMAP tests

**Contents**


**4.1 Introduction** . . . . . **75**
**4.2 Study of the detection surfaces for $N = 2$** . . . . . **76**
4.2.1 Shape of the detection surfaces . . . . . 76
4.2.2 Special cases of equivalence . . . . . 78
**4.3 General N-dimensional case** . . . . . **79**
**4.4 Comparison of the tests performances** . . . . . **81**
**4.5 Conclusion** . . . . . **81**


## 4.1 Introduction

This chapter aims to point out the structural differences between the two tests, the PDR and the LRMAP tests, previously proposed. This is important to specify the conditions for which one test is better adapted than the other towards sparse signals detection goals. Since classical tools based on an analytical comparison of the test statistics have limited power in this case, as seen in section 3.5, we propose an original study focused on the shapes of the PDR and LRMAP detection surfaces. As we shall see, this study suggests that the PDR tends to better detect signals that are close to the axes.

The detection surfaces, that we respectively call $\mathcal{S}_{\text{PDR}}$ and $\mathcal{S}_{\text{LRMAP}}$, represent the decision boundaries between the two hypotheses of model (3.1), $\mathcal{H}_0$ and $\mathcal{H}_1$, and are defined as

$$\mathcal{S} = \{\mathbf{u} \in \mathbb{R}^N \mid T(\mathbf{u}) = \gamma\}, \tag{4.1}$$

where $\mathbf{u}$ are $N$ component vectors and $T(\mathbf{u})$ represents the test statistics of the LRMAP and the PDR tests. Equation (4.1) means that a vector in the volume delimited by $\mathcal{S}$ will not imply detection, while a vector outside will. Because of the symmetry in regard to components permutation in the vector $\mathbf{u}$, we consider, without loss of generality, that $u_1 > u_2 > \cdots > u_N > 0$.

The chapter is organized as follows: as a warmup, the analysis of the PDR and LRMAP detection surfaces is first presented in the simple case of $N = 2$ component vectors $\mathbf{u}$. As an illustration, the two special cases of equivalence of subsection 3.5.2 are then considered for $N = 2$ and $N = 3$, before turning to the study of the general $N-$dimensional case.

## 4.2 Study of the detection surfaces for $N = 2$

### 4.2.1 Shape of the detection surfaces

As seen in (3.5) and (3.14), for both tests, detection is claimed every time their test statistic is above the $\gamma$ threshold (in order to differentiate the PDR and the LRMAP test thresholds, hereafter we will respectively refer to $\gamma_P$ and $\gamma_L$). As previously mentioned in section 3.2 and section 3.3 , for $N = 2$ this may happen in two different cases:

1. the case in which $u_1 > \eta$ and $u_2 \leq \eta$ (or similarly, permuting the $u_1$ and $u_2$ components);

2. the case in which both components $u_1$ and $u_2$ are above the $\eta$ parameter.

Otherwise (i.e., if both components $u_i \leq \eta$), $T(\mathbf{u}) = 0$. $\mathcal{S}_{\text{PDR}}$ and $\mathcal{S}_{\text{LRMAP}}$ are obtained by the computation of $T(\mathbf{u}) = \gamma$ in the two detection cases mentioned above. For the PDR, the computation of $T_{\text{PDR}}(\mathbf{u}) = \gamma_P$, when only $u_1$ is above the $\eta$ threshold, leads to

$$u_1 = \sqrt{\gamma_P} + \eta. \tag{4.2}$$

On the other hand, when both $u_1$ and $u_2$ are above $\eta$, we have

$$(u_1 - \eta)^2 + (u_2 - \eta)^2 = \gamma_P, \tag{4.3}$$

which corresponds, considering $u_1 > u_2$, to an arc of a circle centered in $(\eta, \eta)$ and with radius $r_{\text{PDR}} = \sqrt{\gamma_P}$.

Fig.4.1(a) shows (in red) the $\mathcal{S}_{\text{PDR}}$ surface plotted in the first quadrant of the $(u_1, u_2)$ plane (the surfaces on the other quadrants are obtained by symmetry with respect to both axes). The region in which $u_1 > u_2$, with $u_1 > \eta$, is highlighted in grey. In this region, the $\mathcal{S}_{\text{PDR}}$ part obtained by the combination of (4.2) and (4.3) is plotted in bold red.

Similarly, $\mathcal{S}_{\text{LRMAP}}$ is obtained from the combination of the first detection condition

$$u_1 = \sqrt{\gamma_L + \eta^2}, \tag{4.4}$$

computed from $T_{\text{LRMAP}}$ when only $u_1$ is above $\eta$, and the second condition

$$u_1^2 + u_2^2 = \gamma_L + 2\eta^2, \tag{4.5}$$

which corresponds, for $u_1 > u_2$, to an arc of a circle centered on the origin of the axes and with radius $r_{\text{LRMAP}} = \sqrt{\gamma_L + 2\eta^2}$. Fig.4.1(b) shows (in blue) the $\mathcal{S}_{\text{LRMAP}}$ surface in the first quadrant of the $(u_1, u_2)$ plane. Considering $u_1 > u_2$ with $u_1 > \eta$ (grey region), the combination of (4.4) and (4.5) leads to the $\mathcal{S}_{\text{LRMAP}}$ part highlighted in bold blue.

Note that the two expressions in (4.2) and (4.4) represent, in the $u_1 > u_2$ case (and similarly, when $u_2 > u_1$), the effective detection thresholds when only one component in the $\mathbf{u}$ vector is above $\eta$ and $\gamma_P, \gamma_L \neq 0$. We can thus refer to these quantities as effective *per-component* thresholds.

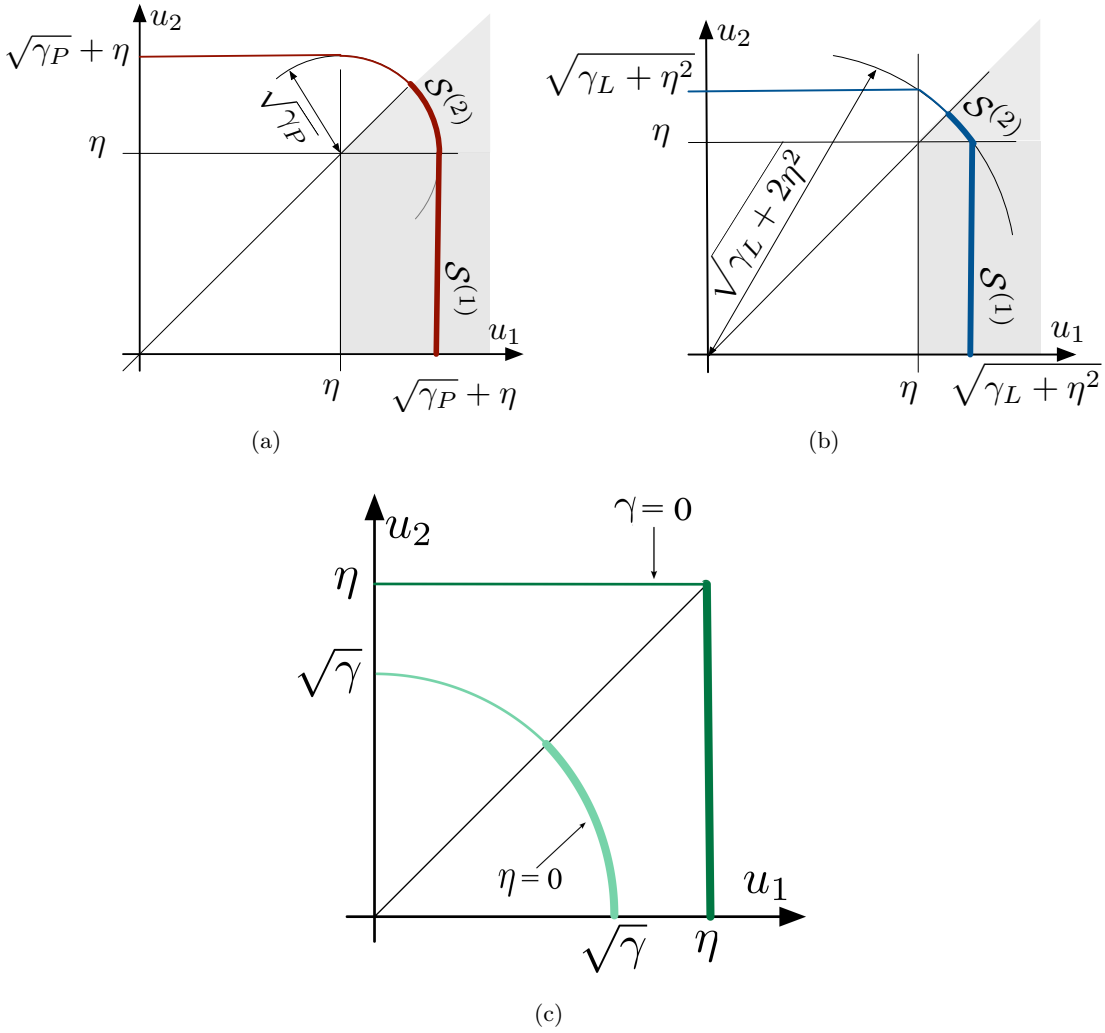



Figure 4.1: Fig.4.1(a) and Fig.4.1(b) show the $\mathcal{S}_{\text{PDR}}$ (in red) and $\mathcal{S}_{\text{LRMAP}}$ (in blue) detection surfaces for an $N = 2$ component vector $\mathbf{u}$. The regions in which $u_1 > u_2$, with $u_1 > \eta$, are reported in grey. In these regions, the detection surfaces, obtained from the combination of (4.2) and (4.3) for the PDR and of (4.4) and (4.5) for the LRMAP, are plotted in bold (see text). Fig.4.1(c) illustrates the detection surfaces, for the two tests indistinctly, in the two special cases of equivalence described in subsection 4.2.2: $\eta = 0$ (light green); $\gamma = 0$ (dark green).

### 4.2.2 Special cases of equivalence

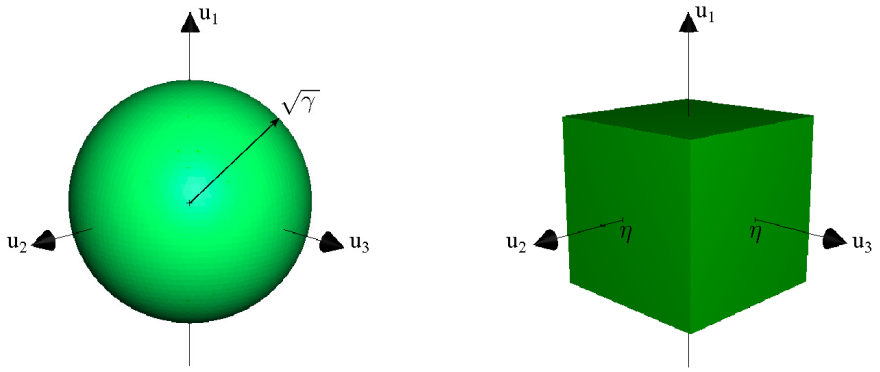
The shapes of the two surfaces are strongly parameter dependent. Despite this fact, as seen in subsection 3.5.2, for particular settings of the $\eta$ and $\gamma$ parameters the two detection tests can be equivalent. This happens in the following cases, easily recognizable for $N = 2$ (see Fig.4.1(c)):

1. For $\eta \neq 0$ and $\gamma_P = \gamma_L = 0$, the two detection surfaces are identical squares (dark green);

2. For $\eta = 0$ and $\gamma_P = \gamma_L \neq 0$, the two detection surfaces are identical circles (light green).

In case 1 , the $\mathcal{S}_{\text{PDR}}$ and $\mathcal{S}_{\text{LRMAP}}$ surfaces are thus characterized only by the detection edges respectively computed in (4.2) and (4.4), with $\gamma_P = \gamma_L = 0$. Those conditions correspond to have only one component above the $\eta$ threshold and lead to two identical square-shaped detection surfaces of edge $\eta$. In the second case of equivalence ($\eta = 0$), $\mathcal{S}_{\text{PDR}}$ and $\mathcal{S}_{\text{LRMAP}}$ derive from (4.3) and (4.5). The surfaces represent two identical circles, centered in the origin of the axes and with radius $\sqrt{\gamma_P} = \sqrt{\gamma_L} = \sqrt{\gamma}$. Fig.4.1(c) illustrates the detection surfaces obtained for the two tests indistinctly, plotted in the first quadrant of the $(u_1, u_2)$ plane for the two cases of equivalence discussed above.

The same cases of equivalence can be considered in the $N = 3$ dimensional case. Here, when $\eta \neq 0$ and $\gamma_P = \gamma_L = 0$, $\mathcal{S}_{\text{PDR}}$ and $\mathcal{S}_{\text{LRMAP}}$ represent two identical cubes, while two identical spheres are obtained for $\eta = 0$ and $\gamma_P = \gamma_L \neq 0$. The detection surfaces obtained for $N = 3$ are shown in Fig.4.2.




(a) (b)

Figure 4.2: 3D detection surfaces in the two different cases of equivalence of the tests. Fig.4.2(a) shows the case of $\eta = 0$ and $\gamma_L = \gamma_P \neq 0$ while Fig.4.2(b) the case of $\eta \neq 0$ and $\gamma_L = \gamma_P = 0$.

This study illustrates geometrically the results obtained in subsection 3.5.2 where, when $\eta \to 0$, the MAP-based tests are the same and coincide with the GLR's one. In this case, detection is claimed outside a sphere and is thus isotropic. Sparsity is not favored since sparse vectors tend to be located close to the axes. On the contrary, when $\gamma \to 0$, sparsity

only on one component (1−sparsity) of the $\boldsymbol{\theta}$ vector is promoted. Indeed, this effect is true but to a lesser extent for signals that are not 1− but $n$−sparse, with $n < N$.

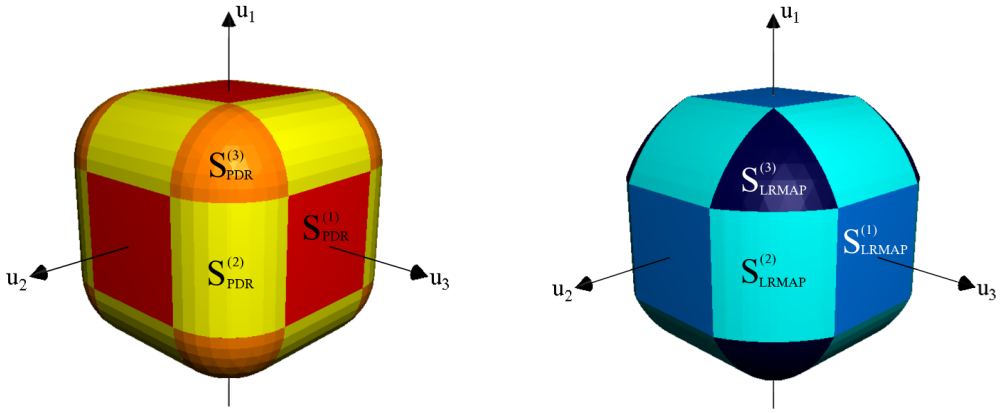
## 4.3 General N-dimensional case

The LRMAP and the PDR tests are now considered in the general case of $N$ component vectors. In this case, we are interested in characterizing the regions of intersection that may exist between $\mathcal{S}_{\text{PDR}}$ and $\mathcal{S}_{\text{LRMAP}}$. Those regions of intersection are of particular interest as they will allow to compare the tests at a same $P_{FA}$.

We first decompose the surfaces $\mathcal{S}$ as a union of subsurfaces $\mathcal{S} = \bigcup_{n=1}^{N} \mathcal{S}^{(n)}$, with the $\mathcal{S}^{(n)}$ defined as

$$\mathcal{S}^{(n)} = \{\mathbf{u} \mid T(\mathbf{u}) = \gamma, u_i > \eta, i = 1, \ldots, n \text{ and } u_i \leq \eta, i > n\}\,. \tag{4.6}$$

Definition (4.6) means that each $\mathbf{u} \in \mathcal{S}^{(n)}$ has the first $n$ components above the $\eta$ threshold and the last $N-n$ components less than or equal to $\eta$. Fig.4.3 shows the $\mathcal{S}_{\text{PDR}}$ and $\mathcal{S}_{\text{LRMAP}}$




(a) (b)

Figure 4.3: 3D $\mathcal{S}_{\text{PDR}}$ and $\mathcal{S}_{\text{LRMAP}}$ detection surfaces. 3D $\mathcal{S}_{\text{PDR}}$, obtained for $\eta = 1$ and $\gamma_P = 1.35$. 3D $\mathcal{S}_{\text{LRMAP}}$, obtained for $\eta = 1$ and $\gamma_L = 4.2$. For both detection surfaces we distinguish three different types of subsurfaces: $\mathcal{S}^{(1)}_{\text{PDR}}$ (in red), $\mathcal{S}^{(2)}_{\text{PDR}}$ (in yellow) and $\mathcal{S}^{(3)}_{\text{PDR}}$ (in orange) in Fig.4.3(a), and $\mathcal{S}^{(1)}_{\text{LRMAP}}$ (in light blue), $\mathcal{S}^{(2)}_{\text{LRMAP}}$ (in cyan) and $\mathcal{S}^{(3)}_{\text{LRMAP}}$ (in dark blue) in Fig.4.3(b).

shapes for $N = 3$, with, in this example, $\eta = 1$, $\gamma_P = 1.35$ and $\gamma_L = 4.2$. For both detection surfaces we can distinguish three different subsurfaces: $\mathcal{S}^{(1)}$ (in which only one component of $\mathbf{u}$ is above the $\eta$ threshold), $\mathcal{S}^{(2)}$ (two components above $\eta$) and $\mathcal{S}^{(3)}$ (three components above $\eta$).

The following proposition reports the conditions that $\gamma_P$ and $\gamma_L$ must satisfy for $\mathcal{S}_{\text{PDR}}$ and $\mathcal{S}_{\text{LRMAP}}$ to intersect.

**Proposition 4.3.1.** *The intersection regions between the $\mathcal{S}^{(n)}_{PDR}$ and $\mathcal{S}^{(n)}_{LRMAP}$ subsurfaces are specified considering the cases of having $n = 1$ or $n \geq 2$ components over the $\eta$ threshold. For each case, the following conditions on the thresholds of the tests hold:*

*1)* $n = 1$*: if* $\gamma_P = \gamma_L = 0$, $\mathcal{S}^{(1)}_{PDR}$ *and* $\mathcal{S}^{(1)}_{LRMAP}$ *coincide. In any other case, no intersection between these two subsurfaces occurs. This means that, depending on the thresholds values, one subsurface is above the other. In particular, if*

$$\gamma_L < \gamma_P + 2\eta\sqrt{\gamma_P}, \tag{4.7}$$

*we have* $\mathcal{S}^{(1)}_{PDR} > \mathcal{S}^{(1)}_{LRMAP}$*. On the other hand, if*

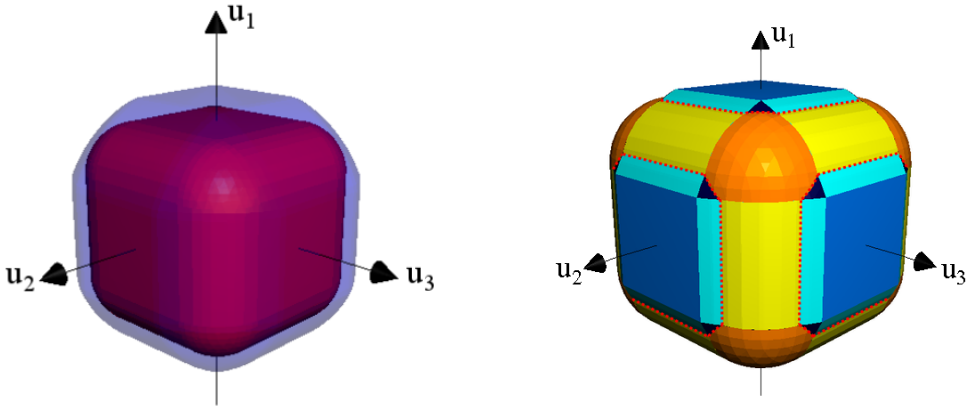
$$\gamma_L > \gamma_P + 2\eta\sqrt{\gamma_P}, \tag{4.8}$$

*we have* $\mathcal{S}^{(1)}_{LRMAP} > \mathcal{S}^{(1)}_{PDR}$*.*

*2)* $n \geq 2$*: intersections between* $\mathcal{S}^{(n)}_{PDR}$ *and* $\mathcal{S}^{(n)}_{LRMAP}$ *occur if*

$$\gamma_P + \eta\sqrt{2n\gamma_P} < \gamma_L < \gamma_P + 2\eta\sqrt{n\gamma_P} \tag{4.9}$$

*(see Proof in Appendix A.4).*




(a) (b)

Figure 4.4: $\mathcal{S}_{\text{PDR}}$ and $\mathcal{S}_{\text{LRMAP}}$ 3D detection surfaces in the $(u_1, u_2, u_3)$ plane. Fig.4.4(a) shows an example of no intersection between the two surfaces, obtained setting $\eta = 1$, $\gamma_P = 1$ and $\gamma_L = 5$. Here, $\mathcal{S}_{\text{PDR}}$ (in red) is included in $\mathcal{S}_{\text{LRMAP}}$ (in blue). On the other hand, Fig.4.4(b) illustrates a case of intersections between $\mathcal{S}_{\text{PDR}}$ and $\mathcal{S}_{\text{LRMAP}}$ (same color code as in Fig.4.3), obtained for $\eta = 1$, $\gamma_P = 1.35$ and $\gamma_L = 4.2$. In particular, intersections (highlighted by red dotted lines) occur for $n = 2$ and $n = 3$ (in which cases $\mathcal{S}^{(n)}_{\text{PDR}} > \mathcal{S}^{(n)}_{\text{LRMAP}}$) , while $\mathcal{S}^{(1)}_{\text{PDR}} < \mathcal{S}^{(1)}_{\text{LRMAP}}$ for $n = 1$.

From the results obtained above we see that if for example condition (4.8) holds, and the right inequality of (4.9) is not satisfied, we fall in the particular case for which one detection surface is included into the other and no intersection occurs. In this case, given two detection surfaces, say for example $\mathcal{S}_1$ and $\mathcal{S}_2$, we state that $\mathcal{S}_1$ is above (below) $\mathcal{S}_2$

if $\forall \mathbf{u} \in \mathcal{S}_1$ and $\forall \mathbf{v} \in \mathcal{S}_2$, we have that $||\mathbf{u}||_2^2 > ||\mathbf{v}||_2^2$ (or, alternatively, $||\mathbf{u}||_2^2 < ||\mathbf{v}||_2^2$). This case is illustrated in Fig.4.4(a), which shows the $\mathcal{S}_{\text{PDR}}$ (in red) and $\mathcal{S}_{\text{LRMAP}}$ (in blue) detection surfaces, computed for $\eta = 1$, $\gamma_P = 1$ and $\gamma_L = 5$. According to (4.8), we have that $\mathcal{S}_{\text{PDR}} \subset \mathcal{S}_{\text{LRMAP}}$, and consequently $P_{FA_{\text{PDR}}} > P_{FA_{\text{LRMAP}}}$ (on the contrary, when (4.8) and the left inequality in (4.9) are not verified, $\mathcal{S}_{\text{LRMAP}} \subset \mathcal{S}_{\text{PDR}}$ and $P_{FA_{\text{PDR}}} < P_{FA_{\text{LRMAP}}}$).

On the other hand, if for instance (4.8) is satisfied and we choose $\gamma_L$ according to (4.9), then $\mathcal{S}^{(1)}_{\text{PDR}} < \mathcal{S}^{(1)}_{\text{LRMAP}}$, with possibility of intersection in at least one of the $n \geq 2$ domains. This is shown in Fig.4.4(b), where the detection surfaces of the tests are obtained for $\eta = 1$, $\gamma_P = 1.35$ and $\gamma_L = 4.2$. Here intersections, which are highlighted by red dotted lines, occur in the $n = 2$ and $n = 3$ domains, with $\mathcal{S}^{(n)}_{\text{PDR}} > \mathcal{S}^{(n)}_{\text{LRMAP}}$. For $n = 1$, as condition (4.8) is verified, we have that $\mathcal{S}^{(1)}_{\text{PDR}} < \mathcal{S}^{(1)}_{\text{LRMAP}}$. This means that, while the LRMAP better detects off-axis (i.e. when two or more components of $\mathbf{u}$ are above the $\eta$ threshold), the PDR test tends to be better than the LRMAP on the axes directions, favoring the detection of highly sparse signal vectors.

## 4.4 Comparison of the tests performances

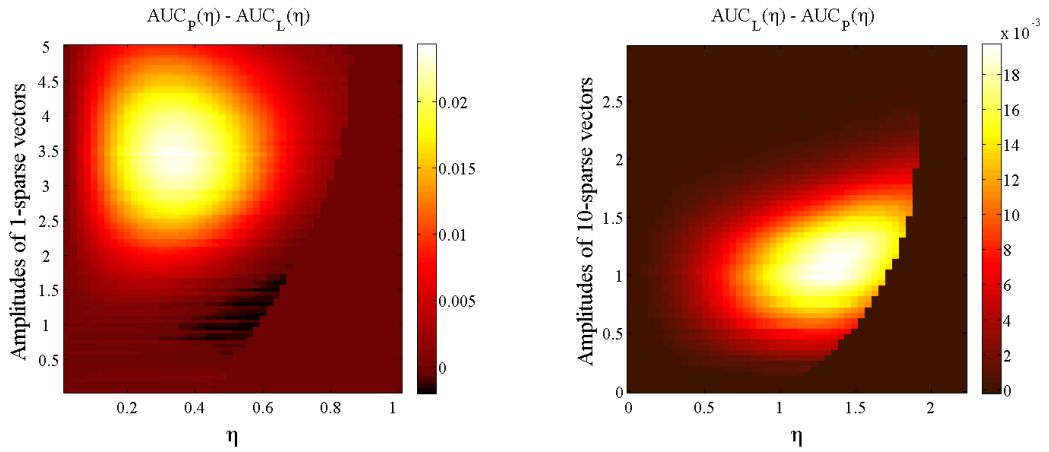
The relevance of this interpretation is investigated by the simulations of Fig.4.5, which compare the tests power in terms of area under the curve (AUC) values, calculated as a function of different $\eta$ thresholds. The AUC is the area under the ROC curves of the tests, and takes consistent values between $0.5 < \text{AUC} \leq 1$ [Barrett 1998]. The AUC is a measure of the global (over all $P_{FA}$) power of a test: the larger the AUC, the more powerful the test.

Fig.4.5(a) compares the tests in the case of different 1−sparse vectors of dimension $N = 20$, in which the nonzero component takes increasing values in $[0.01, 5]$. For each value of $\eta$ (linearly varying in $[0.01, 2.2]$), the difference between the PDR AUC ($\text{AUC}_P$) and the LRMAP's one ($\text{AUC}_L$) is computed. Despite a few cases in which $\text{AUC}_L$ is above $\text{AUC}_P$ (darkest regions in the figure, corresponding to very small negative values of the $\text{AUC}_P - \text{AUC}_L$ difference, the minimum of which is equal to $-3 \cdot 10^{-4}$), the PDR test outperforms the LRMAP. This confirms the interpretation drawn from the geometrical detection surfaces, that the PDR test tends to better detect on the axes directions with respect to the LRMAP. On the contrary, when less sparse vectors are considered, this is the opposite. Fig.4.5(b) shows the results obtained considering $N = 20$ component vectors, in which $n = 10$ components are different from zero and of the same values, linearly varying in the $[0.01, 3]$ interval. This time, the LRMAP test outperforms the PDR.

## 4.5 Conclusion

Our conclusions of this comparative study are as follows. Both tests are equivalent and adapted to 1−sparse signals for $\gamma = 0$. For $\eta \neq 0$ and $\gamma \neq 0$, the PDR tends to better detect highly sparse signals while the LRMAP is more adapted to less sparse signals. There may be marginal counterexamples to this general rule (see Fig.4.5(a)).

A geometrical explanation to this behaviour is that at the same probability of false alarm, the shape of the PDR detection surface is closer to a cube than the detection surface of the LRMAP. Bonferroni/*Max* test [Arias-Castro 2010] (see also (5.24)), whose detection




(a) (b)

Figure 4.5: PDR and LRMAP AUCs difference as a function of $\eta$ threshold values, linearly increasing in the $[0.01, 2.2]$ interval. Fig.4.5(a) shows the results obtained computing the difference $\mathrm{AUC}_P(\eta) - \mathrm{AUC}_L(\eta)$ on $N = 20$ component vectors (considering $5{\cdot}10^4$ realizations for each vector), with only one nonzero value, linearly varying in the $[0.01, 5]$ interval. On the other hand, Fig.4.5(b) shows results obtained computing the difference $\mathrm{AUC}_L(\eta) - \mathrm{AUC}_P(\eta)$ on $N = 20$ component vectors of parameters ($5{\cdot}10^4$ realizations for each vector), with $n = 10$ nonzero values, in $[0.01, 3]$. The lower-right part of both figures corresponds to regions of non-detectability of signals (AUC $= 0.5$). For highly sparse vectors (1−sparse, here) the PDR test tends to be better than the LRMAP, since $\mathrm{AUC}_P(\eta) - \mathrm{AUC}_L(\eta)$ is positive in the vast majority of cases, while for lesser sparse vectors the situation is reversed.

surface is a cube, is very efficient in detecting $k$-sparse signals for $k = 1$, but its power decreases for $k > 1$. Hence, for data sets in line with Fig.4.5(b) (that is, composed by $k > 1$ nonzero signal components that have comparable magnitudes), the PDR presents lower performances than the LRMAP. For such signals, the too sharp PDR detection surface (cube-like) is less adapted than the LRMAP surface, whose softer edges account for the possibility of lower sparsity.

We end by noting that another possible way to compare the performances of the two tests could be to evaluate, for a same value of the $\mathrm{P_{FA}}$, the corresponding values of the $\mathrm{P_{DET}} = Pr(T(\boldsymbol{u}, \eta) > \gamma \mid \mathcal{H}_1)$, in function of both the $\eta$ and $\gamma$ parameters. Since the analytical computation of the $\mathrm{P_{FA}}$ and $\mathrm{P_{DET}}$ of the tests with respect to both the ($\eta$,$\gamma$) parameters shows up to be very difficult even in the case of the simple model (3.1), we suggest to perform this analysis by geometrically constructing and numerically computing minimizing-maximizing approximating functions aimed at bounding the $\mathrm{P_{FA}}$ and $\mathrm{P_{DET}}$ of the tests. The bases for such a problem, which is still under investigation, are presented in Appendix B.5.

Subpart B

# Refined sparsity-based detection strategies

Chapter 5

# Spectral model based strategies

Contents


**5.1 Introduction . . . . . 85**
**5.2 1D (spectral) model . . . . . 86**
5.2.1 ML estimate of $\boldsymbol{\theta}$ and $T_{\mathrm{GLR}}$ for a dictionary-based model . . . . . 87
5.2.2 MAP estimate of $\boldsymbol{\theta}$ for a dictionary-based model . . . . . 87
5.2.3 $T_{\mathrm{PDR}}$ and $T_{\mathrm{LRMAP}}$ test statistics for a dictionary-based model . . . . . 88
5.2.4 Computation of $\mathrm{P_{FA}}$ and $\mathrm{P_{DET}}$ at $\gamma = 0$: the PDR/LRMAP test . . . 89
**5.3 The *Max* test as a constrained GLR . . . . . 90**
**5.4 Comparison between the $\mathrm{GLR}_{\mathbf{1s}}^{(\mathbf{1D})}$ test and the HC . . . . . 91**
**5.5 Conclusion . . . . . 95**


## 5.1 Introduction

To be closer to realistic scenarios involving data acquisition systems and data sparse in dictionaries, the model considered in this chapter accounts explicitly for appropriate known redundant dictionaries $\mathbf{R}$ and for the spread of information caused by the instrument on the signal explanatory dimension (which could be for example a temporal, frequency or energy dimension). To be in agreement with the application domain considered later in this work (see Part III), we refer to this dimension as $\lambda$ (wavelength). We call *Wavelength Spread Function* (WSF) the mathematical function that describes the instrument acquisition system's response to an infinitesimally thin line. Note that the WSF may not always be translation invariant (we refer to subsection 7.2.4 for an instance where the WSF may vary over the spectral bands of the considered signals).

We study and compare the detection tests of chapter 3 reformulated and adapted to the case of a dictionary-based spectral model. In particular, we consider in section 5.2 the case of the PDR and LRMAP detection tests based on a data model that takes advantage from the use of a one-dimensional (spectral) redundant dictionary. Similarly to subsection 3.5.2, we show that when setting the $\gamma$ threshold to 0 the PDR and LRMAP approaches resume to the same test, and are equivalent to the *Max* testing procedure considered in [Arias-Castro 2010]. This setting also allows to univocally compute the maximal $\mathrm{P_{FA}}$ as a function of the sole $\eta$ parameter. In section 5.3, an alternative formulation of the *Max* test is proposed in terms of a constrained GLR test. A comparison of the detection performances of the constrained GLR test and the Higher Criticism, a second-level significance detection procedure introduced in [Tukey 1976, Donoho 2004] (cf. subsection 2.4.2), is finally broached in section 5.4.

## 5.2 1D (spectral) model

We consider here the tests introduced in chapter 3 for the following simple one-dimensional (1D) spectral model:

$$\begin{cases} \mathcal{H}_0 : \mathbf{s} = \boldsymbol{\epsilon}, & \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \\ \mathcal{H}_1 : \mathbf{s} = \mathbf{H}\mathbf{R}\boldsymbol{\alpha} + \boldsymbol{\epsilon} \end{cases} \tag{5.1}$$

where $\mathbf{s} \in \mathbb{R}^{\Lambda}$ is a vector of noisy observations, $\mathbf{R}$ is a $(\Lambda \times L)$ appropriate known redundant dictionary, which allows to efficiently (sparsely) model $\mathbf{s}$ (cf. subsection 7.3.1) and $\boldsymbol{\alpha}$ is a sparse $L-$vector of unknown parameters. Following the same choices made in chapter 3 for model (3.1), the covariance matrix $\boldsymbol{\Sigma}$ is considered diagonal with $\boldsymbol{\Sigma} = \text{diag}\{\sigma_1^2, \ldots, \sigma_\Lambda^2\}$. In this model, the spectral variations due to the spreading effects of the data acquisition system are taken into account through the use of its matrix form $\mathbf{H}$. In particular, $\mathbf{H}$ represents a $(\Lambda \times \Lambda)$ sparse matrix, whose non-zero elements placed around the diagonal contain the WSF coefficients over all the $\Lambda$ spectral bands. Considering the whitening matrix $\boldsymbol{\Sigma}^{-1/2}$, the model above can be re-written as

$$\begin{cases} \mathcal{H}_0 : \boldsymbol{\Sigma}^{-1/2}\mathbf{s} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\epsilon} = \mathbf{w}, & \mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \\ \mathcal{H}_1 : \boldsymbol{\Sigma}^{-1/2}\mathbf{s} = \boldsymbol{\Sigma}^{-1/2}\mathbf{H}\mathbf{R}\boldsymbol{\alpha} + \mathbf{w}, \end{cases} \tag{5.2}$$

where $\boldsymbol{\Sigma}^{-1/2}\mathbf{H}\mathbf{R} = \mathbf{D}_{\boldsymbol{\Sigma}\mathbf{H}}$ appears as an equivalent dictionary. By noting

$$\begin{aligned} \mathbf{z} &= \boldsymbol{\Sigma}^{-1/2}\mathbf{s} \\ \mathbf{D} &= \mathbf{D}_{\boldsymbol{\Sigma}\mathbf{H}}\mathbf{N}_{\mathbf{D}_{\boldsymbol{\Sigma}\mathbf{H}}}^{-1} \\ \text{and} \quad \boldsymbol{\theta} &= \mathbf{N}_{\mathbf{D}_{\boldsymbol{\Sigma}\mathbf{H}}}\boldsymbol{\alpha}, \end{aligned} \tag{5.3}$$

with $\mathbf{D} = [\mathbf{d}_1 \ldots \mathbf{d}_L]$ a possibly redundant normalized ($||\mathbf{d}_i||_2^2 = 1, \forall i = 1, \ldots, L$) dictionary of size $(\Lambda \times L)$ with $L \geq \Lambda$ atoms, $\mathbf{N}_{\mathbf{D}_{\boldsymbol{\Sigma}\mathbf{H}}}$ the diagonal matrix composed of the norms of the columns of $\mathbf{D}_{\boldsymbol{\Sigma}\mathbf{H}}$ and $\boldsymbol{\theta}$ a column vector of $L$ components assumed to be sparse, model (5.2) becomes

$$\begin{cases} \mathcal{H}_0 : \mathbf{z} = \mathbf{w}, & \mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \\ \mathcal{H}_1 : \mathbf{z} = \mathbf{D}\boldsymbol{\theta} + \mathbf{w}. \end{cases} \tag{5.4}$$

Model (5.4) acts in the weighted (by $\boldsymbol{\Sigma}^{-1/2}$) data domain. This normalization simplifies the setting and interpretation of the detection tests discussed in chapter 3 because all atoms are treated equally, and allows to implement the tests for any target $\text{P}_{\text{FA}}$ while including the specificities of the data through the redundant dictionary $\mathbf{R}$. Note that with $L = \Lambda$ and $\mathbf{D} = \mathbf{I}$, (5.4) reverts back to the model (3.1) studied in chapter 3. Also, $L = \Lambda$ and $\mathbf{D} = \boldsymbol{\Sigma}^{-\frac{1}{2}}$ applies to the case of sparse additive signals against a non-diagonal Gaussian background with covariance $\boldsymbol{\Sigma}$.

In the following, we first compute the ML (subsection 5.2.1) and MAP (subsection 5.2.2) estimate expressions for model (5.4). The computations of the PDR and LRMAP test statistics are then derived in subsection 5.2.3. A particular case of equivalence of the PDR and the LRMAP tests is considered in subsection 5.2.4. This equivalence leads to a unique approach, the PDR/LRMAP test, allowing to implement the tests using the sole $\eta$ threshold.

### 5.2.1 ML estimate of $\boldsymbol{\theta}$ and $T_{\mathbf{GLR}}$ for a dictionary-based model

As previously seen in Sec.1.3.4, the ML estimate of an unknown vector of parameters $\boldsymbol{\theta}$ writes

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{\mathrm{ML}} &= \arg\max_{\boldsymbol{\theta}} p\left(\mathbf{z} \mid \boldsymbol{\theta}\right) \\ &= \arg\max_{\boldsymbol{\theta}} \ln p\left(\mathbf{z} \mid \boldsymbol{\theta}\right). \end{aligned} \tag{5.5}$$

Hence, according to the linear model (5.4)

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{\mathrm{ML}} &= \arg\max_{\boldsymbol{\theta}} -\frac{1}{2}\left(\mathbf{z}-\mathbf{D}\boldsymbol{\theta}\right)^t\left(\mathbf{z}-\mathbf{D}\boldsymbol{\theta}\right) \\ &= \arg\min_{\boldsymbol{\theta}} \frac{1}{2}\left\|\mathbf{z}-\mathbf{D}\boldsymbol{\theta}\right\|_2^2. \end{aligned} \tag{5.6}$$

The ML estimator for $\boldsymbol{\theta}$ obviously minimizes the square error $\left\|\mathbf{z}-\mathbf{D}\boldsymbol{\theta}\right\|_2^2$. From (5.6), the solution $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ that minimizes the least norm computed for a dictionary $\mathbf{D}$ $(N \times L)$ with $N \ll L$ through the right inverse $\mathbf{D}^t(\mathbf{D}\mathbf{D}^t)^{-1}$ is

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \mathbf{D}^t(\mathbf{D}\mathbf{D}^t)^{-1}\mathbf{z}, \tag{5.7}$$

which leads to

$$\mathbf{D}\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \mathbf{z}. \tag{5.8}$$

It is easy to see that, similarly to the definition of $T_{\mathrm{GLR}}$ given in subsection 1.3.2, we have

$$T_{\mathrm{GLR}}(\mathbf{z}) = \frac{\max_{\boldsymbol{\theta}} p\left(\mathbf{z} \mid \mathbf{D}\boldsymbol{\theta}\right)}{p\left(\mathbf{z} \mid \mathbf{0}\right)} = \left\|\mathbf{z}\right\|_2^2. \tag{5.9}$$

This shows that even for model (5.4) the GLR test behaves as an energy detector, without taking any advantage from the use of the redundant dictionary.

We will keep considering the GLR test for the purpose of comparison (see section 7.3).

### 5.2.2 MAP estimate of $\boldsymbol{\theta}$ for a dictionary-based model

Following the notation of subsection 2.3.3, the MAP estimate of an unknown vector of parameters $\boldsymbol{\theta}$ writes:

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{\mathrm{MAP}} &= \arg\max_{\boldsymbol{\theta}} \; p\left(\boldsymbol{\theta} \mid \mathbf{z}\right) \\ &= \arg\max_{\boldsymbol{\theta}} \; p\left(\mathbf{z} \mid \boldsymbol{\theta}\right)\pi\left(\boldsymbol{\theta}\right) \\ &= \arg\min_{\boldsymbol{\theta}} \; -\ln\left[p\left(\mathbf{z} \mid \boldsymbol{\theta}\right)\pi\left(\boldsymbol{\theta}\right)\right]. \end{aligned} \tag{5.10}$$

For model (5.4) and a Laplacian prior of the same form as seen in (2.47), definition (5.10) becomes

$$\hat{\boldsymbol{\theta}}_{\mathrm{MAP}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{2}\left\|\mathbf{z}-\mathbf{D}\boldsymbol{\theta}\right\|_2^2 + \eta\left\|\boldsymbol{\theta}\right\|_1, \tag{5.11}$$

where in this case the hyper-parameter $\eta$ equals $\frac{1}{\lambda} > 0$ (thanks to the normalization in (5.4) $\eta$ is the same for all the vector's components, see section 3.5). Equation (5.11) coincides with the solution of the Basis Pursuit DeNoising (BPDN) introduced in [Tibshirani 1996, Chen 2001]. In contrast with subsection 2.3.3, $\hat{\boldsymbol{\theta}}_{\text{MAP}}$ cannot be obtained by direct soft-thresholding here. A necessary and sufficient condition for $\hat{\boldsymbol{\theta}}_{\text{MAP}}$ to be the global minimum solution of (5.11) is that there exists a vector $\mathbf{v}$ in the sub-differential of $\|\boldsymbol{\theta}\|_1$ which verifies [Fuchs 2004]

$$\mathbf{D}^t(\mathbf{D}\hat{\boldsymbol{\theta}}_{\text{MAP}} - \mathbf{z}) + \eta\mathbf{v} = 0, \tag{5.12}$$

and which satisfies

$$\begin{cases} v_i = \text{sgn}(\hat{\theta}_{MAP_i}) & \text{if} \quad \hat{\theta}_{MAP_i} \neq 0 \\ v_i \leq 1 & \text{if} \quad \hat{\theta}_{MAP_i} = 0. \end{cases} \tag{5.13}$$

As a consequence, for sufficiently large values of $\eta$, $\hat{\boldsymbol{\theta}}_{\text{MAP}}$ is identically zero. The first non-zero component of $\hat{\boldsymbol{\theta}}_{\text{MAP}}$ appears when $\eta$ falls below the quantity $\max_i(|\mathbf{d}_i{}^t\mathbf{z}|)$. It is easy to see that, if $\eta > \max_i(|\mathbf{d}_i{}^t\mathbf{z}|)$, taking $\hat{\boldsymbol{\theta}}_{\text{MAP}} = \mathbf{0}$ in (5.12) automatically yields a $\mathbf{v}$ which satisfies (5.13) and thus the solution in zero.

### 5.2.3 $T_{\text{PDR}}$ and $T_{\text{LRMAP}}$ test statistics for a dictionary-based model

According to the analysis described above, we compute the expressions of the $T_{\text{LRMAP}}$ and $T_{\text{PDR}}$ test statistics.

For model (5.4), the definition of the $T_{\text{LRMAP}}$ test statistic seen in (3.2) gives

$$\begin{aligned} T_{\text{LRMAP}}(\mathbf{z}) &= -\frac{1}{2}\left[\left(\mathbf{z} - \mathbf{D}\hat{\boldsymbol{\theta}}_{\text{MAP}}\right)^t\left(\mathbf{z} - \mathbf{D}\hat{\boldsymbol{\theta}}_{\text{MAP}}\right) - \mathbf{z}^t\mathbf{z}\right] \\ &= \mathbf{z}^t\mathbf{D}\hat{\boldsymbol{\theta}}_{\text{MAP}} - \frac{1}{2}\hat{\boldsymbol{\theta}}^t_{\text{MAP}}\mathbf{D}^t\mathbf{D}\hat{\boldsymbol{\theta}}_{\text{MAP}}. \end{aligned} \tag{5.14}$$

Multiplying (5.12) by $\hat{\boldsymbol{\theta}}^t_{\text{MAP}}$ leads to

$$\hat{\boldsymbol{\theta}}^t_{\text{MAP}}\mathbf{D}^t(\mathbf{D}\hat{\boldsymbol{\theta}}_{\text{MAP}} - \mathbf{z}) + \eta\,\hat{\boldsymbol{\theta}}^t_{\text{MAP}}\,\text{sgn}(\hat{\boldsymbol{\theta}}_{\text{MAP}}) = 0. \tag{5.15}$$

Noting

$$\mathbf{D}\hat{\boldsymbol{\theta}}_{\text{MAP}} = \hat{\mathbf{z}}_{\text{MAP}},$$

and

$$\hat{\boldsymbol{\theta}}^t_{\text{MAP}}\,\text{sgn}(\hat{\boldsymbol{\theta}}_{\text{MAP}}) = \left|\left|\hat{\boldsymbol{\theta}}_{\text{MAP}}\right|\right|_1,$$

where $\left|\left|\hat{\boldsymbol{\theta}}_{\text{MAP}}\right|\right|_1$ represents the $l_1-$norm of the $\hat{\boldsymbol{\theta}}_{\text{MAP}}$ vector, we have

$$\hat{\mathbf{z}}^t_{\text{MAP}}(\hat{\mathbf{z}}_{\text{MAP}} - \mathbf{z}) + \eta\left|\left|\hat{\boldsymbol{\theta}}_{\text{MAP}}\right|\right|_1 = 0. \tag{5.16}$$

Adding the quantity $\frac{1}{2}\hat{\mathbf{z}}^t_{\text{MAP}}\hat{\mathbf{z}}_{\text{MAP}}$ to both terms in (5.16) yields

$$\hat{\mathbf{z}}^t_{\text{MAP}}\mathbf{z} - \frac{1}{2}\hat{\mathbf{z}}^t_{\text{MAP}}\hat{\mathbf{z}}_{\text{MAP}} = \frac{1}{2}\hat{\mathbf{z}}^t_{\text{MAP}}\hat{\mathbf{z}}_{\text{MAP}} + \eta\left|\left|\hat{\boldsymbol{\theta}}_{\text{MAP}}\right|\right|_1. \tag{5.17}$$

Comparing (5.17) to the $T_{\text{LRMAP}}(\mathbf{z})$ expression in (5.14), we finally obtain

$$T_{\text{LRMAP}}(\mathbf{z}) = \frac{1}{2}||\hat{\mathbf{z}}_{\text{MAP}}||_2^2 + \eta \left\|\hat{\boldsymbol{\theta}}_{\text{MAP}}\right\|_1, \tag{5.18}$$

which is the expression of the $T_{\text{LRMAP}}$ test statistic for model (5.4).

For the computation of the $T_{\text{PDR}}$ test statistic we follow a similar reasoning. According to model (5.4), the definition of the test statistic $T_{\text{PDR}}$ seen in (3.8), for a Laplacian prior of the form (2.47), gives

$$T_{\text{PDR}}(\mathbf{z}) \quad = \quad \hat{\boldsymbol{\theta}}_{\text{MAP}}^t \left[\mathbf{D}^t\mathbf{z} - \frac{1}{2}\mathbf{D}^t\mathbf{D}\hat{\boldsymbol{\theta}}_{\text{MAP}} - \eta\,\text{sgn}(\hat{\boldsymbol{\theta}}_{\text{MAP}})\right], \tag{5.19}$$

which leads, noting $\mathbf{D}\hat{\boldsymbol{\theta}}_{\text{MAP}} = \hat{\mathbf{z}}_{\text{MAP}}$, to

$$T_{\text{PDR}}(\mathbf{z}) = \hat{\mathbf{z}}_{\text{MAP}}^t\mathbf{z} - \frac{1}{2}\hat{\mathbf{z}}_{\text{MAP}}^t\hat{\mathbf{z}}_{\text{MAP}} - \eta \left\|\hat{\boldsymbol{\theta}}_{\text{MAP}}\right\|_1. \tag{5.20}$$

From (5.17) we have that

$$\frac{1}{2}\hat{\mathbf{z}}_{\text{MAP}}^t\hat{\mathbf{z}}_{\text{MAP}} = \hat{\mathbf{z}}_{\text{MAP}}^t\mathbf{z} - \frac{1}{2}\hat{\mathbf{z}}_{\text{MAP}}^t\hat{\mathbf{z}}_{\text{MAP}} - \eta \left|\left|\hat{\boldsymbol{\theta}}_{\text{MAP}}\right|\right|_1. \tag{5.21}$$

Replacing (5.21) in (5.20), we finally obtain

$$T_{\text{PDR}}(\mathbf{z}) = \frac{1}{2}\,||\hat{\mathbf{z}}_{\text{MAP}}||_2^2, \tag{5.22}$$

which is the expression of the $T_{\text{PDR}}$ test statistic for model (5.4).

Note the similarity of expression (5.22) and (5.18) with that of the unconstrained GLR test defined in (5.9). Basically, while the GLR test acts as a global energy detector (5.9), the PDR and LRMAP tests concentrate only on the energy in the subspace defined by the MAP estimate through $||\hat{\mathbf{z}}_{\text{MAP}}||_2^2$.

From equations (5.18) and (5.22) we also see that one of the main characteristics of the two tests statistics is that both are strictly positive, or null if and only if $\hat{\boldsymbol{\theta}}_{\text{MAP}} = \mathbf{0}$ (that is, if $\max_i(|\boldsymbol{d}_i{}^t\mathbf{z}|) \leq \eta$). Moreover, they both depend only on the term $\left|\left|\hat{\boldsymbol{\theta}}_{\text{MAP}}\right|\right|_1$.

### 5.2.4 Computation of $\mathbf{P_{FA}}$ and $\mathbf{P_{DET}}$ at $\gamma = 0$: the PDR/LRMAP test

Similarly to subsection 3.5.2, we address the problem of computing the maximal probability of false alarm $\text{P}_{\text{FA}_0}$, considering both tests at $\gamma = 0$. This assumption allows us to bypass the tedious question of the double parameter dependency of the tests (on the $\gamma$ and $\eta$ thresholds), and to define the $\text{P}_{\text{FA}}$ in terms of the sole $\eta$ parameter. As previously mentioned, for both the LRMAP and PDR tests at least one non-zero component appears in $\hat{\boldsymbol{\theta}}_{\text{MAP}}$ (and thus in $\hat{\mathbf{z}}_{\text{MAP}}$) if the quantity $\max_i(|\mathbf{d}_i{}^t\mathbf{z}|)$ falls above $\eta$.

For a fixed value of $\eta$, we have

$$\begin{aligned} \text{P}_{\text{FA}_0} &= Pr(T > 0 \mid \mathcal{H}_0) \\ &= Pr(\max_i(\left|\boldsymbol{d}_i{}^t\mathbf{z}\right|) > \eta \mid \mathcal{H}_0), \end{aligned} \tag{5.23}$$

where $T$ represents either the LRMAP or the PDR test statistic. For this reason, the two tests are strictly equivalent at $\gamma = 0$. We refer to this unique test as the PDR/LRMAP test. Note that, since the dictionary $\mathbf{D}$ is redundant, the components of the $L-$vector $\mathbf{D}^t\mathbf{z}$ are not independent, and finding $\mathrm{P}_{\mathrm{FA}_0}$ analytically is thus a difficult problem (see [Fuchs 2010]). We can nevertheless resort to MC simulations to obtain an accurate correspondence between $\eta$ and $\mathrm{P}_{\mathrm{FA}_0}$ (as illustrated in section 7.3). Note finally that for $\gamma = 0$, it is not necessary to solve (5.11) to implement these detection tests, as they amount to reject the null hypothesis if

$$\max_i(\left|\mathbf{d}_i{}^t\mathbf{z}\right|) > \eta. \tag{5.24}$$

This test is not new in the literature and corresponds to the test used in [Fuchs 2010]. This test is also called the *Max* test in [Arias-Castro 2010] and resembles a generalized matched filter where $L$ filters $\mathbf{d}_1, \ldots, \mathbf{d}_L$ are tested. While less effective for the detection of low sparsity level signals, the *Max* test was shown to be optimal with respect to the Bayes risk[9] in the case of high sparsity levels, where it exhibits a power that is comparable to that of the HC detector of [Donoho 2004]. The comparison of those two tests when applied to the entries of $\mathbf{D}^t\mathbf{z}$ [Arias-Castro 2010] is reported in section 5.4.

We note for now that the strategy of choosing $\gamma = 0$ leads to focus on highly sparse signals in which the few non-zero components are only those above the $\eta$ threshold. This is justified in our application framework (see Part III), where the SNR is very low, so that for most interesting signals under $\mathcal{H}_1$ essentially one or very few spectral features should be detectable.

## 5.3 The *Max* test as a constrained GLR

As previously mentioned, an appropriate detector for very sparse data is the *Max* test, introduced in [Arias-Castro 2010] as a Bonferroni-type testing approach. For this test, we propose a simple alternative formulation as a constrained GLR test. Let us consider the following constrained model:

$$\mathcal{H}_1 : \mathbf{s} = \mathbf{HR}\boldsymbol{\alpha} + \boldsymbol{\epsilon}, \quad ||\boldsymbol{\alpha}||_0 = 1, \tag{5.25}$$

which similarly to (5.1)-(5.4), in the weighted domain leads to

$$\mathcal{H}_1 : \mathbf{z} = \mathbf{D}\boldsymbol{\theta} + \mathbf{w}, \quad ||\boldsymbol{\theta}||_0 = 1. \tag{5.26}$$

In this new model, the signal component of the whitened data vector $\mathbf{z}$ is obtained as one of the columns of the redundant dictionary $\mathbf{D}$, selected and weighted by the sole non-zero component of $\boldsymbol{\theta}$. We thus have

$$\mathcal{H}_1 \ : \ \mathbf{z} = \mathbf{d}_i\theta_i + \mathbf{w}, \tag{5.27}$$

with $i$ and $\theta_i$ unknown.

According to model (5.27), we define the constrained one-sparse GLR test as:

$$\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})} \ : \ \frac{\max\limits_{j,\ \theta_j} p\left(\mathbf{z} \mid \mathbf{d}_j, \theta_j\right)}{p\left(\mathbf{z} \mid \mathbf{0}\right)} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma_{1\mathrm{s}}, \tag{5.28}$$

[9] Defined in Appendix A.2

which is equal to

$$\text{GLR}_{1s}^{(1D)} \; : \; \frac{\max\limits_{j,\,\theta_j} \exp\left(-\frac{1}{2}\left\|\mathbf{z}-\mathbf{d}_j\theta_j\right\|_2^2\right)}{\exp\left(-\frac{1}{2}\left\|\mathbf{z}\right\|_2^2\right)} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma_{1s}. \tag{5.29}$$

The (1D) apex refers here to the one-dimensionality of the considered model. This notation will help to distinguish this test from the three-dimensional (3D) model-based detectors introduced later (see chapter 6). The computation of the $T_{\text{GLR}_{1s}}^{(1D)}$ test statistic gives

$$T_{\text{GLR}_{1s}}^{(1D)}(\mathbf{z}) = \min_{j,\,\theta_j} -\mathbf{z}^T\mathbf{d}_j\theta_j + \frac{1}{2}\theta_j^2\mathbf{d}_j^T\mathbf{d}_j. \tag{5.30}$$

For $\mathbf{d}_j$ fixed, the ML estimate of $\theta_j$ is

$$\hat{\theta}_{j_{\text{ML}}} = \mathbf{d}_j^T\mathbf{z}. \tag{5.31}$$

Replacing this expression in (5.30) we have

$$T_{\text{GLR}_{1s}}^{(1D)}(\mathbf{z}) = \min_{j} -\mathbf{z}^T\mathbf{d}_j\hat{\theta}_{j_{\text{ML}}} + \frac{1}{2}\hat{\theta}_{j_{\text{ML}}}^2\mathbf{d}_j^T\mathbf{d}_j \tag{5.32}$$

$$= \min_{j} -\frac{1}{2}\left(\mathbf{d}_j^T\mathbf{z}\right)^2 = \max_{j} \frac{1}{2}\left\|\mathbf{d}_j^T\mathbf{z}\right\|_2^2. \tag{5.33}$$

This is equivalent to maximise

$$T_{\text{GLR}_{1s}}^{(1D)}(\mathbf{z}) = \max_{j} \left|\mathbf{d}_j^T\mathbf{z}\right|. \tag{5.34}$$

The GLR test for model (5.27) finally writes

$$\text{GLR}_{1s}^{(1D)} \; : \; T_{\text{GLR}_{1s}}^{(1D)}(\mathbf{z}) \underset{H_0}{\overset{H_1}{\gtrless}} \xi_{1s}, \tag{5.35}$$

where $\xi_{1s} = 2\ln\gamma_{1s}$. Thus

$$T_{\text{GLR}_{1s}}^{(1D)}(\mathbf{z}) = \max_{j} \left|\mathbf{d}_j^T\mathbf{z}\right| = T_{Max}(\mathbf{z}), \tag{5.36}$$

as in (5.24). Summarizing, the GLR test for (5.26) is equivalent to the *Max* test of [Arias-Castro 2010] applied to $\left|\mathbf{D}^t\mathbf{z}\right|$ and is also the particular case of the PDR and LRMAP tests at $\gamma = 0$.

For a simple notation, we hereafter refer to this test as the $\text{GLR}_{1s}^{(1D)}$. This notation will allow in chapter 6 to distinguish with the $\text{GLR}_{1s}^{(3D)}$, a GLR test applied to a more complete 3D (spatio-spectral) model.

## 5.4 Comparison between the $\text{GLR}_{1s}^{(1D)}$ test and the Higher Criticism

This section proposes a numerical comparison of the $\text{GLR}_{1s}^{(1D)}$ test (5.35) and the HC detector of [Donoho 2004] (cf. subsection 2.4.2), here applied to the entries of $\left|\mathbf{D}^t\mathbf{z}\right|$, so that $p_i =$

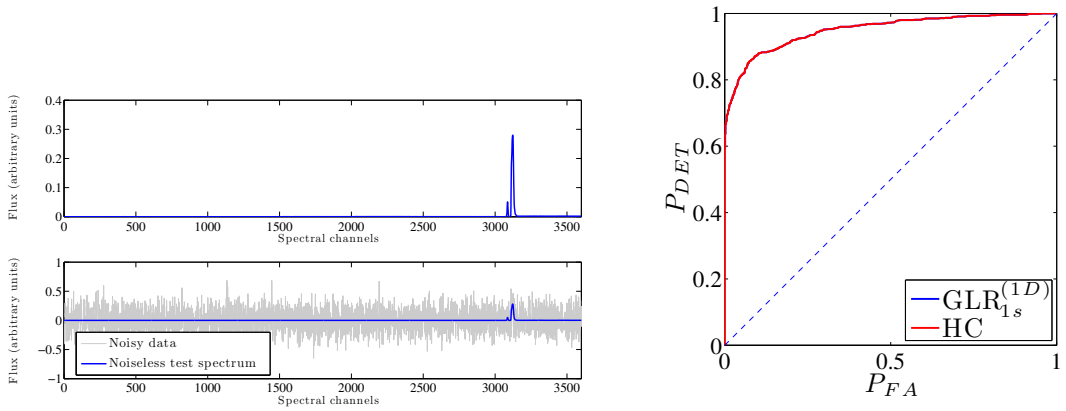
$Pr\left\{|\mathcal{N}(0,1)| > \left|\mathbf{D}^t\mathbf{z}\right|_i\right\}$ here. In particular, we are interested in showing the corresponding powers of the two tests for the detection of weak sparse signals buried in white noise.

We compare the HC detector and the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ test for the three signals (which are galactic spectra) of Figure 5.1. Here, the left figures show the noiseless test spectra (in blue) and the corresponding noisy data (in gray). We note that the considered signals have very different spectral characteristics: the first one (Figure 5.1(a)) essentially shows a strong emission line around spectral channel 3100; the second one (Figure 5.1(b)) is mainly characterized by an oscillatory behavior; the third one (Figure 5.1(c)) mainly combines a continuum component with an emission line around spectral channel 2200.

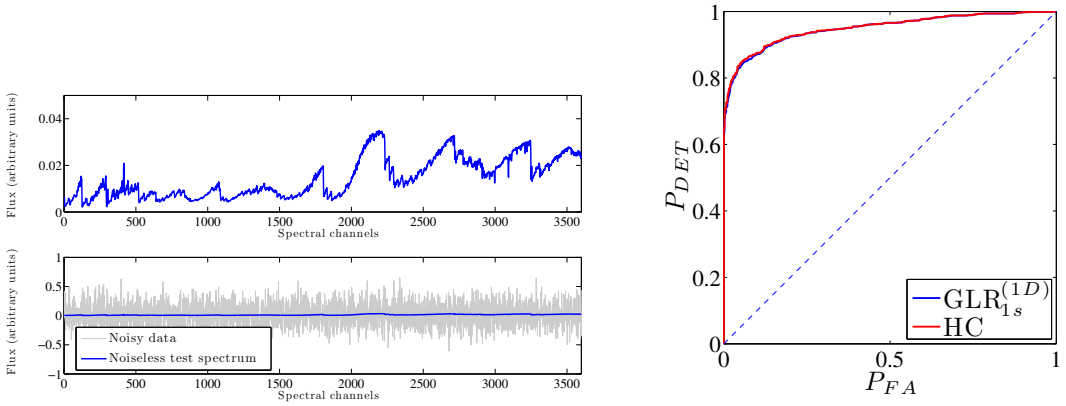
According to the detection regime we are interested in, which reflects the detection conditions we need to face in the application domain of this work, the spectra are buried in a strong white Gaussian noise (SNR of about $-21$dB in these simulations). In this context, the signals can all be considered as weak and sparse possibly in some appropriate dictionary $\mathbf{D}$, the construction of which will be explained in subsection 7.3.1. The plots in the right column of Figure 5.1 compare the ROCs (run over $10^3$ realizations) of the HC test applied to the entries of $\left|\mathbf{D}^t\mathbf{z}\right|$ (in red) and of the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ test (in blue). Clearly, the HC and the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ test have very close detection power. This shows the effectiveness of the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ detector for very sparse signals with very low amplitude.

This equivalence holds only for very sparse signals: in the case of less sparse signals, the HC detector shows to be more powerful than the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ test. This is illustrated by the simulations reported in Figure 5.2, in which we have run the HC and the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ tests on the two following spectra: the first one (Figure 5.2(a), left) obtained as the sum of the noiseless spectra in Figure 5.1(b) and Figure 5.1(c); the second one (Figure 5.2(b), left) obtained as the sum of the three noiseless spectra in Figure 5.1(a), Figure 5.1(b) and Figure 5.1(c). Obviously, these two spectra are less sparse than those of Figure 5.1.

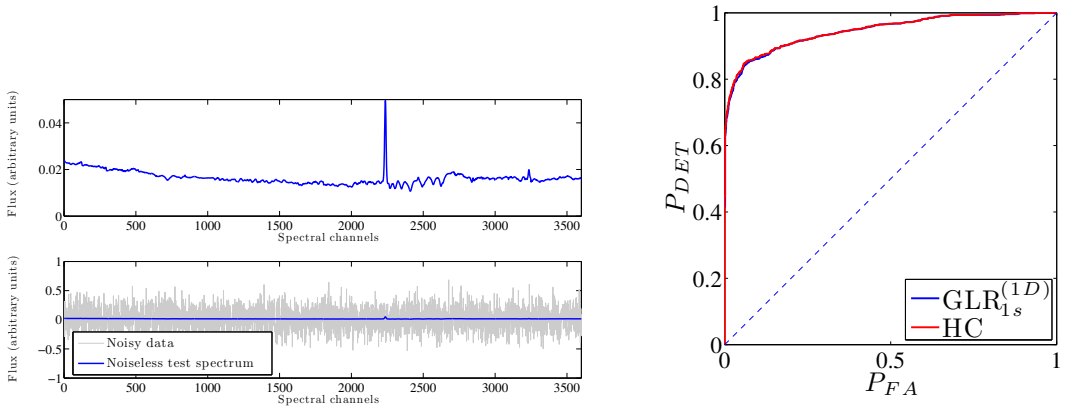
For both noisy signals of Figure 5.2, the SNR is of about $-21$dB. The corresponding ROC curves for the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ (in blue) and the HC test (in red), obtained over $2 \cdot 10^3$ realizations, are reported in Figure 5.2(a) (right) and Figure 5.2(b) (right). From these curves, we clearly note that for less sparse signals the HC detector outperforms the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ test.




(a)

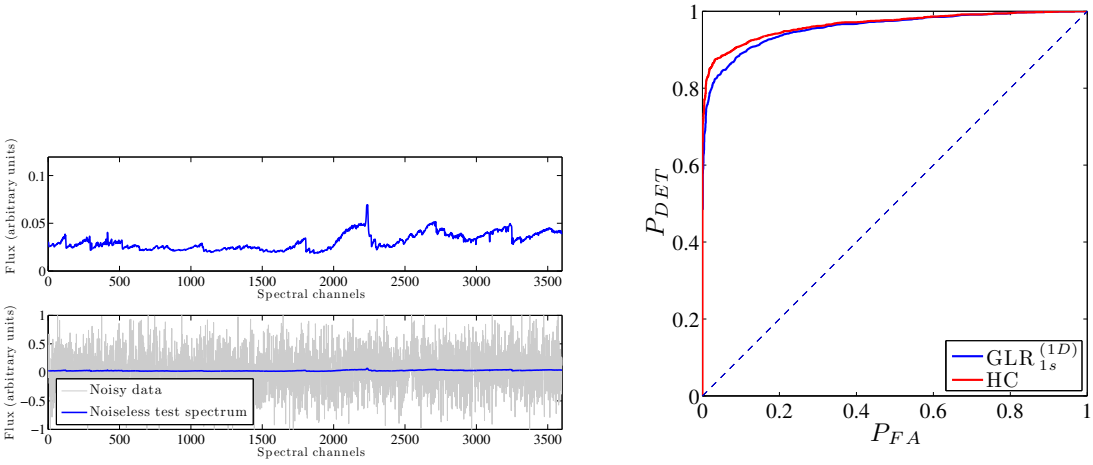



(b)

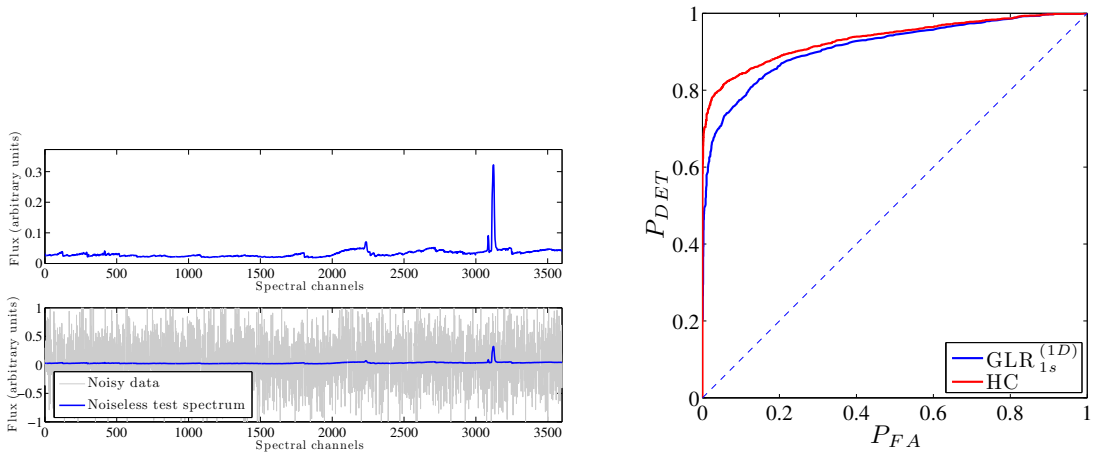



(c)

Figure 5.1: Numerical comparisons of the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ versus the HC for signals that can be considered very sparse accounting for noise power: the left figures show the noiseless test spectra (in blue) and the corresponding noisy data (in gray); the right figures compare the ROCs (run over $10^3$ realizations) of the HC detector (in red) and of the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ test (in blue). Clearly, the HC and the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ tests have very close detection power.




(a)




(b)

Figure 5.2: Numerical comparisons of the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ versus the HC detector in the case of less sparse signals (with respect to the spectra shown in Figure 5.1). The first noiseless test spectrum in Figure 5.2(a) (right) is obtained as the sum of the noiseless spectra in Figure 5.1(b) and Figure 5.1(c); the second noiseless test spectrum (Figure 5.2(b), left) is obtained as the sum of the noiseless spectra in Figure 5.1(a), Figure 5.1(b) and Figure 5.1(c). For both spectra, the SNR is of about $-21$dB. The corresponding ROC curves for the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ and the HC tests, obtained over $2 \cdot 10^3$ realizations, are reported in Figure 5.2(a) (right) and Figure 5.2(b) (right). We note that for less sparse signals the HC detector outperforms the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$.

## 5.5 Conclusion

The PDR and LRMAP detection tests firstly introduced in chapter 3 have been here studied in the more general context of a dictionary-based model accounting for instrument model (through matrix $\mathbf{H}$) and signal structure (through $\mathbf{R}$). When setting the $\gamma$ test threshold to zero, the PDR and the LRMAP tests are equivalent. Thanks to this setting the $P_{FA}$ can be uniquely fixed, allowing to implement the tests for a false alarm rate that is function of the sole $\eta$ parameter. For model (5.4) and $\gamma = 0$ the PDR/LRMAP test rejects the null hypothesis if $\max_i(|\mathbf{d}_i{}^t\mathbf{z}|) > \eta$. This corresponds to the *Max* test introduced in [Arias-Castro 2010], which was shown to be optimal with respect to the Bayes risk in the case of high sparsity levels.

The *Max* test can also be derived as a one-dimensional constrained GLR test and we note this test as $\mathrm{GLR}_{1s}^{(1D)}$. This alternative notation will be later contrasted with a GLR on a 3D spatio-spectral model (chapter 6).

The $\mathrm{GLR}_{1s}^{(1D)}$ test has been compared to the HC detector, a second-level significance technique introduced in [Tukey 1976, Donoho 2004]. The comparison on different types of weak and sparse spectra buried in white Gaussian noise have numerically shown the comparable detection power of the two methods in the case of highly sparse signals.

The detection tests developed in this chapter accounted only for data's spectral information. When considering multi-dimensional data sets however, as it is the case in chapter 7, the use of the sole spectral information may not be enough to reach satisfactory detection results for extremely faint signals. In a multi-dimensional data set, spatial dependencies between neighbor elements usually exist. This fact can be exploited to increase the detection performances of the proposed tests. We investigate this situation in the next chapter.

Chapter 6

# Spatio - Spectral model based strategies

**Contents**


**6.1 Introduction . . . . . 97**
**6.2 First attempts to account for spatial dependencies . . . . . 98**
6.2.1 A simple toy-case with equal spectra and same noise statistics . . . . 98
6.2.2 Refined detection model: proportional neighbor spectra and spatially variable noise statistics. . . . . 102
6.2.3 Multiple-round detection strategy . . . . . 106
**6.3 A (spatio-spectral) 3D model-based strategy . . . . . 109**
6.3.1 3D data Model and $\mathrm{GLR}_{1\mathrm{s}}^{(3\mathrm{D})}$ detection test . . . . . 110
6.3.2 Design strategies for the choice of the dictionary . . . . . 111
**6.4 Conclusion . . . . . 112**


## 6.1 Introduction

We tackle now the problem of designing efficient detection strategies for the identification of weak, sparse signals in highly noisy *three-dimensional* (3D) data. 3D data sets (or data cubes) usually combine two spatial directions $x$ and $y$ (e.g. image or video frame dimensions) with an additional direction $\lambda$ (e.g. temporal, spectral or energy dimension). Such data most often suffer from information leakage caused by the data acquisition system, which may be different and variable in the three dimensions. This effect is usually described by a 3D Point Spread Function (PSF), a function that mathematically specifies any instrument acquisition system's response to a point source in the $x$, $y$ and $\lambda$ dimensions. In this chapter, spatial (resp. spectral) will refer to the $x$, $y$ (resp. $\lambda$) dimensions, so that the proposed methods can be generically applied regardless of the physical nature of $x$, $y$ and $\lambda$. Accordingly, we separate the PSF into a Wavelength Spread Function (WSF, already defined in the previous chapter) and a Spatial Spread Function (SSF) that, similarly to the WSF in the $\lambda$ direction, depicts the acquisition system's response to a point source in the spatial dimensions.

We consider 3D data sets as a collection of spectra (or pixel-vectors), each spectrum covering the whole set of wavelengths in the $\lambda$ dimension at specific $(x, y)$ spatial coordinates. In this case, two spatially contiguous spectra are likely to share some spectral information, either because the sources we look at are resolved (i.e. spread over several pixels), or because

of the instrument' SSF. Thus, it is likely that the detection of very faint spectra can be improved by using the spectral information detected in contiguous brighter pixels. Next sections describe the detection models and the detection tests designed to implement such ideas. With respect to chapter 5, improved detection performances are reached in section 6.2 through the introduction of a multiple-round detection approach based on a simple model of spatial dependencies.

A more accurate 3D (spatio-spectral) model-based detection test is then considered in section 6.3. In order to correctly take into account the spatio-spectral blur of information caused by the instrument's PSF, both the WSF and the SSF effects are modeled. The detection techniques based on this new model will take advantage from dedicated dictionaries composed of 3D elementary atoms.

## 6.2 First attempts to account for spatial dependencies

### 6.2.1 A simple toy-case with equal spectra and same noise statistics

We denote $\mathbf{z}_b$ and $\mathbf{z}_f$ two contiguous, respectively bright and faint spectra in a given data cube. The problem we address in this section is that of detecting unknown salient spectral features in the faint spectrum $\mathbf{z}_f$ by exploiting the spatial (and spectral) proximity with the brighter one ($\mathbf{z}_b$), which we know contains consistent information about $\mathbf{z}_f$.

#### 6.2.1.1 The model

For both signals we refer to spectral model (5.2), which in the weighted domain and under the $\mathcal{H}_1$ hypothesis writes

$$\begin{aligned} \mathbf{z} &= \boldsymbol{\Sigma}^{-1/2}\mathbf{s} \\ &= \boldsymbol{\Sigma}^{-1/2}\mathbf{H}\mathbf{R}\boldsymbol{\alpha} + \mathbf{w}, \end{aligned} \tag{6.1}$$

where, as seen in section 5.2, $\boldsymbol{\Sigma}^{-1/2}$ is the whitening matrix, $\mathbf{H}$ is the matrix form of the instrument's WSF, $\mathbf{R}$ is the redundant dictionary, $\boldsymbol{\alpha}$ the sparse vector of coefficients and $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ a white Gaussian noise vector. This model can be re-written as

$$\mathbf{z} = \mathbf{D}\boldsymbol{\theta} + \mathbf{w}, \tag{6.2}$$

where $\mathbf{D}$ denotes a known normalized redundant dictionary and $\boldsymbol{\theta}$ is an unknown vector of weighting parameters. We thus obtain

$$\begin{aligned} \mathbf{z}_f &= \mathbf{D}\boldsymbol{\theta}_f + \mathbf{w} \\ &= \mathbf{x}_f + \mathbf{w}, \end{aligned} \tag{6.3}$$

for the faint spectrum, and

$$\begin{aligned} \mathbf{z}_b &= \mathbf{D}\boldsymbol{\theta}_b + \mathbf{w} \\ &= \mathbf{x}_b + \mathbf{w} \end{aligned} \tag{6.4}$$

for the bright one.

In the first (crude) approximation considered in this section we assume that the whitening matrix $\mathbf{\Sigma}^{-1/2}$ involved in $\mathbf{D}$ is the same for the two spectra. Therefore, the normalized redundant dictionary $\mathbf{D} = \mathbf{\Sigma}^{-1/2}\mathbf{H}\mathbf{R}\mathbf{N}^{-1}_{\mathbf{D}_{\mathbf{\Sigma}\mathbf{H}}}$ (cf. (5.3)) is unique for both $\mathbf{z}_f$ and $\mathbf{z}_b$.

Since two spatially contiguous spectra are likely to share some spectral information, we assume, again crudely, that the two unknown noiseless spectra $\mathbf{x}_b$ and $\mathbf{x}_f$ are the same:

$$\text{Simple spatial dependency: } \mathbf{x}_b = \mathbf{x}_f. \tag{6.5}$$

In this context, a meaningful estimate of the noiseless faint spectrum $\mathbf{x}_f$ can be computed as

$$\hat{\mathbf{x}}_f = \mathbf{D}\hat{\boldsymbol{\theta}}(\mathbf{x}_b) = \hat{\mathbf{x}}_b, \tag{6.6}$$

where $\hat{\boldsymbol{\theta}}(\mathbf{x}_b)$ is the sparse parameter vector estimate based on the informative content of the bright spectrum $\mathbf{x}_b$. $\hat{\boldsymbol{\theta}}(\mathbf{x}_b)$ can for instance be obtained as the solution of the BPDN problem defined in (5.11) for $\mathbf{x} = \mathbf{x}_b$ (efficient strategies are discussed in [Bourguignon 2011]).

A fast and well known alternative to compute $\hat{\boldsymbol{\theta}}(\mathbf{x}_b)$ is to use a greedy algorithm such as for example the Matching Pursuit (MP), firstly introduced in [Mallat 1993]. The MP algorithm seeks to approach a given signal with a sparse decomposition of atoms selected from a redundant dictionary $\mathbf{D}$, initially set. Thus, at each iteration, the column of the dictionary that is most correlated with the residual is scaled and subtracted from the data to obtain a new residual, and the sparse vector of coefficients $\hat{\boldsymbol{\theta}}$ is updated. The iterations stop when the maximal correlation falls below a fixed significance threshold.

Hence, in the first approximation considered here, we obtain for the faint spectrum the following simple model

$$\mathcal{H}_1 : \mathbf{z}_f = \hat{\mathbf{x}}_f + \mathbf{w}, \tag{6.7}$$

where $\hat{\mathbf{x}}_f = \mathbf{D}\hat{\boldsymbol{\theta}}(\mathbf{x}_b)$.

#### 6.2.1.2 The LR-MP test

Injecting the estimate $\hat{\mathbf{x}}_f$ in the Likelihood Ratio (1.39), we define a matched filter-like detection test for $\mathbf{z}_f$, which we call the LR-MP (Likelihood Ratio with Matching Pursuit estimate) test

$$\text{LR-MP} : \frac{p(\mathbf{z}_f|\hat{\mathbf{x}}_f)}{p(\mathbf{z}_f|\mathbf{0})} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma. \tag{6.8}$$

The computation of the LR-MP test statistic gives

$$T_{\text{LR-MP}}(\hat{\mathbf{x}}_f) = \mathbf{z}_f^t\hat{\mathbf{x}}_f \; - \; \frac{1}{2}\hat{\mathbf{x}}_f^t\hat{\mathbf{x}}_f \underset{H_0}{\overset{H_1}{\gtrless}} \gamma', \tag{6.9}$$

where $\gamma' = \ln\gamma$.

The $\mathrm{P_{FA}}$ corresponding to the LR-MP test is computed as

$$\begin{aligned}
\mathrm{P_{FA}}(\hat{\mathbf{x}}_f) &= Pr\left(T_{\text{LR-MP}} > \gamma' \mid \mathcal{H}_0\right) \\
&= 1 - Pr\left(T_{\text{LR-MP}} < \gamma' \mid \mathcal{H}_0\right) \\
&= 1 - Pr\left(\mathbf{z}_f^t \hat{\mathbf{x}}_f \; - \; \frac{1}{2}\hat{\mathbf{x}}_f^t \hat{\mathbf{x}}_f < \gamma' \mid \mathcal{H}_0\right) \\
&= 1 - Pr\left(\frac{\mathbf{z}_f^t \hat{\mathbf{x}}_f}{||\hat{\mathbf{x}}_f||_2} < \frac{\gamma' + \frac{1}{2}\hat{\mathbf{x}}_f^t \hat{\mathbf{x}}_f}{||\hat{\mathbf{x}}_f||_2} \mid \mathcal{H}_0\right) \\
&= 1 - \Phi\left(\frac{\gamma' + \frac{1}{2}\hat{\mathbf{x}}_f^t \hat{\mathbf{x}}_f}{||\hat{\mathbf{x}}_f||_2}\right),
\end{aligned} \tag{6.10}$$

where $\Phi(\cdot)$ is the CDF of a standard normal distribution defined in (1.25). Note that since the $\mathrm{P_{FA}}$ (6.10) depends on the faint spectrum estimate $\hat{\mathbf{x}}_f$ and on the noise realization, for a given test threshold $\gamma$ slightly variable values of the $\mathrm{P_{FA}}$ will be obtained from one pixel to another. This will be numerically illustrated in Figure 6.4(a).

#### 6.2.1.3 Upper limits

In the purpose of establishing upper limits, we compare the performances of the proposed detection strategy with those of two Oracle tests (cf. (3.17)):

- **Oracle 1**: This Oracle has knowledge only of the faint spectrum (Oracle 1, for short), and of the support $\mathcal{I}_f$ of $\boldsymbol{\theta}_f$ (which is the same as that of $\boldsymbol{\theta}_b$), but not of the amplitudes $\theta_i$, $i \in \mathcal{I}_f$. Those are here estimated by least squares using the faint spectrum $\mathbf{z}_f$. This leads to the estimate

$$\hat{\boldsymbol{\theta}}_{\mathrm{Or_1}} = \arg\min_{\boldsymbol{\theta}} ||\mathbf{z}_f - \mathbf{D}\boldsymbol{\theta}_f||_2^2. \tag{6.11}$$

  Denoting by $\mathbf{D}_{\mathcal{I}_f}$ the restriction of the dictionary $\mathbf{D}$ to the columns corresponding to $\mathcal{I}_f$, the solution of (6.11) gives

$$\hat{\boldsymbol{\theta}}_{\mathrm{Or_1}} = (\mathbf{D}_{\mathcal{I}_f}^t \mathbf{D}_{\mathcal{I}_f})^{-1} \mathbf{D}_{\mathcal{I}_f}^t \mathbf{z}_f. \tag{6.12}$$

  According to this, the faint spectrum estimate is equal to

$$\hat{\mathbf{x}}_{\mathrm{Or_1}} = \mathbf{D}_{\mathcal{I}_f} \hat{\boldsymbol{\theta}}_{\mathrm{Or_1}}. \tag{6.13}$$

  For the following observation model

$$\begin{cases} \mathcal{H}_0 : \mathbf{z}_f = \mathbf{w}, \quad \mathbf{w} \sim \mathcal{N}(0, \mathbf{I}) \\ \mathcal{H}_1 : \mathbf{z}_f = \hat{\mathbf{x}}_{\mathrm{Or_1}} + \mathbf{w}, \end{cases} \tag{6.14}$$

  the Oracle 1 test is thus a GLR test with $\hat{\mathbf{x}}_{\mathrm{Or_1}}$ estimate under $\mathcal{H}_1$ and yields

$$T_{\mathrm{Or_1}} = \mathbf{z}_f^t \hat{\mathbf{x}}_{\mathrm{Or_1}} - \frac{1}{2}\hat{\mathbf{x}}_{\mathrm{Or_1}}^t \hat{\mathbf{x}}_{\mathrm{Or_1}} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma. \tag{6.15}$$

  The Oracle 1 test is indeed an upper limit to any realizable detection test which considers the detection of the faint spectrum *independently from others*.

- **Oracle 2**: As another upper bound for the detection limit of (6.9) we consider the case of an Oracle test that estimates the non-zero amplitudes in the parameter vector $\boldsymbol{\theta}_f$ using a neighbor bright spectrum $\mathbf{z}_b$, instead of the faint $\mathbf{z}_f$ (Oracle 2, for short). Since both parameters $\boldsymbol{\theta}_f$ and $\boldsymbol{\theta}_b$ are similar, the resulting estimate will still be meaningful, up to a scaling factor for the faint spectrum. This leads to a parameter vector estimate

$$\hat{\boldsymbol{\theta}}_{\mathrm{Or}_2} = (\mathbf{D}_{\mathcal{I}_f}^t \mathbf{D}_{\mathcal{I}_f})^{-1} \mathbf{D}_{\mathcal{I}_f}^t \mathbf{z}_b, \tag{6.16}$$

and to a faint spectrum estimate

$$\hat{\mathbf{x}}_{\mathrm{Or}_2} = \mathbf{D}_{\mathcal{I}_f} \hat{\boldsymbol{\theta}}_{\mathrm{Or}_2}. \tag{6.17}$$

Similarly to the Oracle 1 for model (6.14), the Oracle 2 test is a GLR with $\hat{\mathbf{x}}_{\mathrm{Or}_2}$ estimate under $\mathcal{H}_1$ and yields

$$T_{\mathrm{Or}_2} = \mathbf{z}_f^t \hat{\mathbf{x}}_{\mathrm{Or}_2} - \frac{1}{2} \hat{\mathbf{x}}_{\mathrm{Or}_2}^t \hat{\mathbf{x}}_{\mathrm{Or}_2} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma. \tag{6.18}$$

Since the bright spectrum is relatively less noisy than the faint one, we expect the Oracle 2 to perform better than the Oracle 1.

#### 6.2.1.4 Numerical results

The relevance of the LR-MP approach (6.9) presented in this section is illustrated taking a synthetic bright spectrum

$$\mathbf{z}_b = \mathbf{D}\boldsymbol{\theta}_b + \mathbf{w}, \tag{6.19}$$

where $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\boldsymbol{\theta}_b$ is a sparse parameter vector with only 4 non-zero components, which "activate" and weight the following four atoms in the dictionary $\mathbf{D}$: one Heaviside step, one smooth global oscillation, plus two emission lines with Gaussian shapes. The resulting noiseless bright spectrum $\mathbf{x}_b$ is illustrated in blue in Figure 6.1(a). The neighbor faint spectrum is taken with parameters

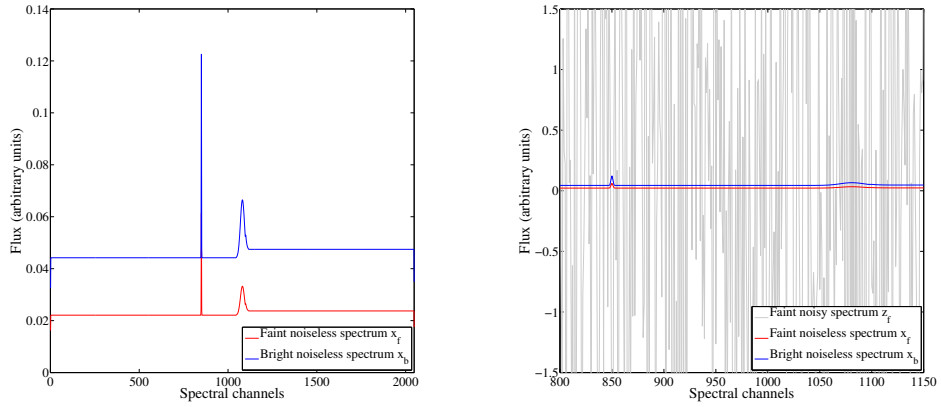
$$\boldsymbol{\theta}_f = 0.5\, \boldsymbol{\theta}_b, \tag{6.20}$$

leading to

$$\mathbf{x}_f = 0.5\, \mathbf{x}_b. \tag{6.21}$$

The factor 0.5 is introduced in order to make model (6.7) only approximately accurate and the data slightly more realistic, as in practice neighbor spectra will not have the same parameters (see subsection 6.2.2). The faint noiseless simulated spectrum is illustrated in red in Figure 6.1(a). The noisy data corresponding to the bright spectrum are shown in gray Figure 6.1(b). The SNR of the two convolved spectra is equal to $-20.9$ dB for the bright signal and of $-26.9$ dB for the faint signal, respectively. As clearly visible from Figure 6.1(b), if compared to the strong noise level, the signals' significant content seems to be almost null.

Numerical results are shown in Figure 6.2. Simulations were run considering the sparse faint and bright spectra of Figure 6.1. The ROCs obtained when considering contiguous spectra as independent ("Detection round on independent spectra") are compared for the




(a) (b)

Figure 6.1: 6.1(a) The bright (in blue) and the faint (in red) noiseless simulated spectra for detection exploiting spatial proximity and spectral similarities. 6.1(b) Bright and the faint noiseless simulated spectra and corresponding noisy data (in gray).

unconstrained GLR ((5.9), green crosses), the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ ((5.35), blue stars) and the Oracle 1 ((6.15), red stars) tests. As expected, the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ outperforms the unconstrained GLR while the Oracle 1 is superior to both the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ and the GLR.

We then compare the performances of the tests exploiting the spectral dependencies between neighbor spectra (6.21). Here, the LR-MP detection test (6.9) is implemented using a faint noiseless spectrum MP estimate $\hat{\mathbf{x}}_f$ based on $\mathbf{x}_b$. The resulting ROC is plotted in cyan stars. The ROC of the Oracle 2 test (6.18) is reported as a reference (red dashed-dotted line). The improvement of the LR-MP test with respect to the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ test is significant. Interestingly, the corresponding ROC is even better than the ROC of the Oracle 1 test: in the Oracle 1, the knowledge of the good support of $\boldsymbol{\theta}_f$ is not enough to compensate the strong estimation noise level of the faint signal's components with respect to the less noisy bright one used by the LR-MP test. This shows that accounting for spatial dependencies is indeed useful, even when using a simplistic model such as the one considered here.

### 6.2.2 Refined detection model: proportional neighbor spectra and spatially variable noise statistics.

In practice, the relationships between spatial neighbors spectra in a data cube are not as simple as that considered in the toy-case of the previous section (6.5). This section aims at setting a more realistic detection model accounting for neighbors spectra with noise statistics that are different because they depend on the spatial coordinates, and proportional noiseless spectral contents depending on an unknown scaling factor.

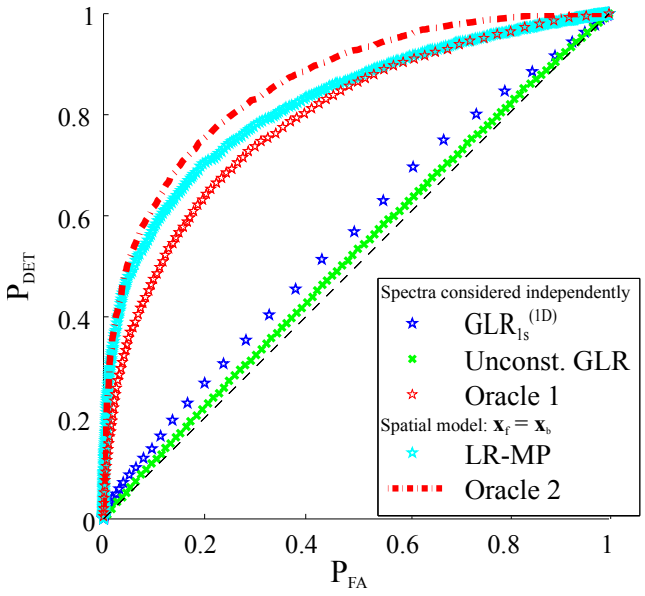



Figure 6.2: "Spectra considered independently": compared ROC curves for the unconstrained GLR ((5.9), green crosses), the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ ((5.35), blue stars) and the Oracle 1 ((6.15), red stars) tests, by considering contiguous spectra as independent. "Spatial model $\mathbf{x}_f = \mathbf{x}_b$": ROC curves for the LR-MP ((6.9), cyan stars) and the Oracle 2 ((6.18), red dashed-dotted line) tests by exploiting spatial dependencies.

#### 6.2.2.1 The model

Let us consider again $\mathbf{z}_b$ and $\mathbf{z}_f$ two whitened contiguous, respectively bright and faint spectra (with $\mathbf{x}_b$ and $\mathbf{x}_f \neq \mathbf{0}$) such that

$$\begin{aligned} \mathbf{z}_b &= \mathbf{x}_b + \mathbf{w} = \mathbf{\Sigma}_b^{-1/2}\mathbf{H}\mathbf{R}\boldsymbol{\alpha}_b + \mathbf{w} \\ \mathbf{z}_f &= \mathbf{x}_f + \mathbf{w} = \mathbf{\Sigma}_f^{-1/2}\mathbf{H}\mathbf{R}\boldsymbol{\alpha}_f + \mathbf{w}, \end{aligned} \tag{6.22}$$

where $\mathbf{\Sigma}_b$ (resp. $\mathbf{\Sigma}_f$) is the covariance matrix of the noise on the bright (resp. faint) spectrum, $\mathbf{H}$ represents the matrix form of the WSF (see section 5.2), and $\mathbf{w} \sim \mathcal{N}(0, \mathbf{I})$. For clarity of exposition we denote hereafter

$$\mathbf{R}\boldsymbol{\alpha}_f = \mathbf{x'}_f \quad \text{and} \quad \mathbf{R}\boldsymbol{\alpha}_b = \mathbf{x'}_b. \tag{6.23}$$

A model which is more realistic than (6.5) defining the common spectral content that the two neighbor signals share can be

$$\mathbf{x'}_f = \beta \mathbf{x'}_b, \tag{6.24}$$

with $\beta$ expected to be $< 1$, but we do not impose this constraint. According to this, in the weighted data domain the model for the faint spectrum can be re-written as

$$\mathbf{z}_f = \beta \mathbf{\Sigma}_f^{-\frac{1}{2}} \mathbf{H}\mathbf{x'}_b + \mathbf{w}. \tag{6.25}$$

For such a spectrum, the most powerful detection test is the LR test (see subsection 1.3.1), which correlates the spectrum $\mathbf{z}_f$ to the noiseless signal component $\beta\boldsymbol{\Sigma}_f^{-\frac{1}{2}}\mathbf{H}\mathbf{x'}_b$.

Since in practice the contiguous noiseless, unconvolved bright spectrum $\mathbf{x}'_b$ and the $\beta$ factor are unknown, they must be estimated. First, we propose to estimate the noiseless bright spectrum $\mathbf{x}'_b$ by a MP algorithm which, as seen before, is a fast and efficient procedure. This is important if we keep in mind implementations on large data sets. As explained in section 5.2, in the weighted data domain the bright spectrum can be written as

$$\begin{aligned}\mathbf{z}_b &= \boldsymbol{\Sigma}_b^{-\frac{1}{2}}\mathbf{H}\mathbf{x'}_b + \mathbf{w} \\ &= \mathbf{D}_b\boldsymbol{\theta}_b + \mathbf{w}.\end{aligned} \tag{6.26}$$

Since $\mathbf{D}_b$ is a normalized dictionary, the implementation of the MP algorithm is standard. At each iteration, the column of $\mathbf{D}_b$ that is most correlated with the residual is scaled and subtracted from the data to obtain a new residual, and the vector of coefficients $\hat{\boldsymbol{\alpha}}_b$ is updated. The final estimate $\hat{\boldsymbol{\alpha}}_b$ is used to obtain $\hat{\mathbf{x}}'_b$ by

$$\hat{\mathbf{x}}'_b = \mathbf{R}\hat{\boldsymbol{\alpha}}_b. \tag{6.27}$$

Second, the factor $\beta$ is estimated by least square:

$$\begin{aligned}\hat{\beta} &= \arg\min_{\beta} \frac{1}{2}||\mathbf{z}_f - \beta\boldsymbol{\Sigma}_f^{-\frac{1}{2}}\mathbf{H}\hat{\mathbf{x}}'_b||_2^2 \\ &= \arg\min_{\beta} \frac{1}{2}\left(\mathbf{z}_f - \beta\boldsymbol{\Sigma}_f^{-\frac{1}{2}}\mathbf{H}\hat{\mathbf{x}}'_b\right)^t\left(\mathbf{z}_f - \beta\boldsymbol{\Sigma}_f^{-\frac{1}{2}}\mathbf{H}\hat{\mathbf{x}}'_b\right) \\ &= \arg\min_{\beta} \frac{1}{2}\mathbf{z}_f^t\mathbf{z}_f - \mathbf{z}_f^t\beta\boldsymbol{\Sigma}_f^{-\frac{1}{2}}\mathbf{H}\hat{\mathbf{x}}'_b + \frac{1}{2}\beta^2||\boldsymbol{\Sigma}_f^{-\frac{1}{2}}\mathbf{H}\hat{\mathbf{x}}'_b||_2^2.\end{aligned} \tag{6.28}$$

Deriving (6.28) with respect to $\beta$ and putting the obtained expression equal to 0 we finally find

$$\hat{\beta} = \frac{\mathbf{z}_f^t\boldsymbol{\Sigma}_f^{-\frac{1}{2}}\mathbf{H}\hat{\mathbf{x}}'_b}{||\boldsymbol{\Sigma}_f^{-\frac{1}{2}}\mathbf{H}\hat{\mathbf{x}}'_b||_2^2}. \tag{6.29}$$

The model (6.25) for the faint spectrum thus becomes

$$\begin{aligned}\mathbf{z}_f &= \hat{\beta}\boldsymbol{\Sigma}_f^{-\frac{1}{2}}\mathbf{H}\hat{\mathbf{x}}'_b + \mathbf{w} \\ &= \hat{\mathbf{x}}^*_f + \mathbf{w},\end{aligned} \tag{6.30}$$

in which we have set $\hat{\mathbf{x}}^*_f = \hat{\beta}\boldsymbol{\Sigma}_f^{-\frac{1}{2}}\mathbf{H}\hat{\mathbf{x}}'_b$. An example illustrating the relevance of model (6.30) is shown in Figure 6.3. Here, the bright (in green) and faint (in blue) considered spectra are drawn from a simulated but highly realistic data set, and hence do not exactly follow (6.30). Nevertheless, a meaningful estimate $\hat{\mathbf{x}}^*_f$ for the faint spectrum is obtained (see subsubsection 6.2.2.3).

#### 6.2.2.2 The LR-MP$\beta$ detection test

In accordance to model (6.30), we define the following matched-filter like detection test

$$\text{LR-MP}\beta: \quad \frac{p(\mathbf{z}_f; \hat{\mathbf{x}}^*_f)}{p(\mathbf{z}_f; \mathbf{0})} \underset{H_0}{\overset{H_1}{\gtrless}} \xi, \tag{6.31}$$

where MP refers to Matching Pursuit and $\beta$ to the unknown amplitude. The computation of the LR-MP$\beta$ test statistic explicitly writes

$$\begin{aligned} T_{\text{LR-MP}\beta} &= \frac{p(\mathbf{z}_f; \hat{\mathbf{x}}_f^*)}{p(\mathbf{z}_f; \mathbf{0})} \\ &= \frac{\exp\left(-\frac{1}{2}||\mathbf{z}_f - \hat{\beta}\boldsymbol{\Sigma}_f^{-\frac{1}{2}}\mathbf{H}\hat{\mathbf{x}}_b'||_2^2\right)}{\exp\left(-\frac{1}{2}||\mathbf{z}_f||_2^2\right)}. \end{aligned} \tag{6.32}$$

Taking the logarithm of (6.32) we have

$$T_{\text{LR-MP}\beta} = \mathbf{z}_f^t \hat{\beta}\boldsymbol{\Sigma}_f^{-\frac{1}{2}}\mathbf{H}\hat{\mathbf{x}}_b' - \frac{1}{2}\hat{\beta}^2||\boldsymbol{\Sigma}_f^{-\frac{1}{2}}\mathbf{H}\hat{\mathbf{x}}_b'||_2^2. \tag{6.33}$$

Replacing in (6.33) the expression of $\hat{\beta}$ computed in (6.29) and taking the square root of both terms in the LR-MP$\beta$ test we finally find that

$$T_{\text{LR-MP}\beta} = \frac{|\mathbf{z}_f^t \boldsymbol{\Sigma}_f^{-\frac{1}{2}}\mathbf{H}\hat{\mathbf{x}}_b'|}{||\boldsymbol{\Sigma}_f^{-\frac{1}{2}}\mathbf{H}\hat{\mathbf{x}}_b'||_2} \underset{H_0}{\overset{H_1}{\gtrless}} \xi', \tag{6.34}$$

with $\xi' = \sqrt{2\ln\xi}$. We recall that this model improves on the one proposed in (subsection 6.2.1), in which the model simply considered $\beta = 1$ and $\boldsymbol{\Sigma}_f = \boldsymbol{\Sigma}_b$.

The $\text{P}_{\text{FA}}$ of the LR-MP$\beta$ test is

$$\begin{aligned} \text{P}_{\text{FA}}(\hat{\mathbf{x}}_b') &= Pr\left(T_{\text{LR-MP}\beta} > \xi \mid \mathcal{H}_0\right) \\ &= 2[1 - \Phi(\xi)], \end{aligned} \tag{6.35}$$

because $T_{\text{LR-MP}\beta} \sim \mathcal{N}(0, \mathbf{I})$ under $\mathcal{H}_0$ (cf. (1.32) and (1.34) with $\theta_0 = 0$ and $\sigma = 1$). The $\text{P}_{\text{FA}}$ (6.35) here depends on the estimate of the considered bright spectrum ($\hat{\mathbf{x}}_b'$) and on the noisy realizations of the faint ($\mathbf{x}_f$) and of the bright ($\mathbf{x}_b$) spectra. As already mentioned in subsubsection 6.2.1.2, setting a particular test threshold $\xi$ thus leads to slightly variable $\text{P}_{\text{FA}}$ values from one test spectrum to another. This effect is illustrated in the numerical simulation of Figure 6.4(a).

#### 6.2.2.3 Numerical results

An example illustrating the relevance of model (6.30) is shown in Figure 6.3. This figure reports a bright noiseless spectrum (in green) and a faint contiguous noiseless spectrum (in blue). These spectra are drawn from the simulated 3D data set considered in section 8.2 (their precise positions are marked in Figure 8.1(a) with green and blue crosses, respectively). Although these spectra are not strictly proportional, they share common features (e.g. the emission line around spectral band interval $1500 - 1550$). In grey are plotted the noisy data corresponding to the faint spectrum. Clearly, the emission line around this interval is below the noise level. In magenta is plotted the estimated noiseless faint spectrum $\hat{\mathbf{x}}_f^*$, built taking advantage from the MP denoised version of the bright spectrum. The emission line was captured by the MP algorithm in the bright data spectrum. This estimate is indeed meaningful for the faint spectrum.

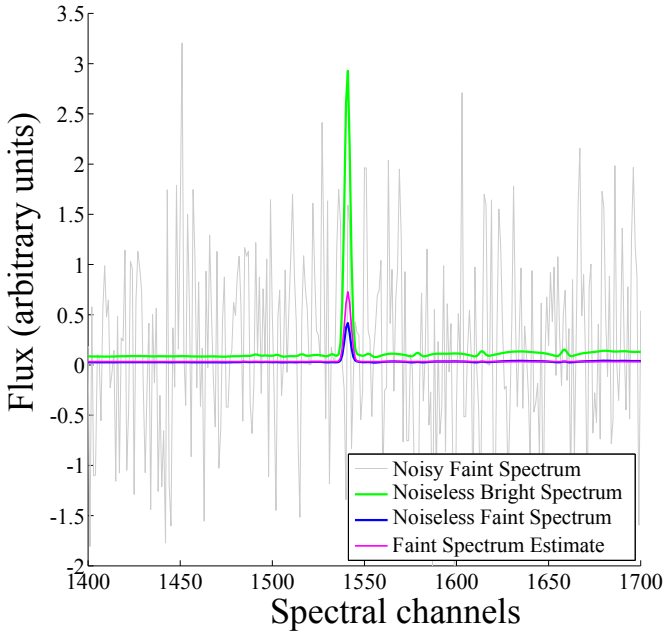



Figure 6.3: Faint spectrum estimate using model (6.30). The plot reports the bright noiseless spectrum (in green) and a faint contiguous noiseless spectrum (in blue). In grey are plotted the noisy data corresponding to the faint spectrum. The estimated noiseless faint spectrum $\hat{\mathbf{x}}_f^*$ is plotted in magenta.

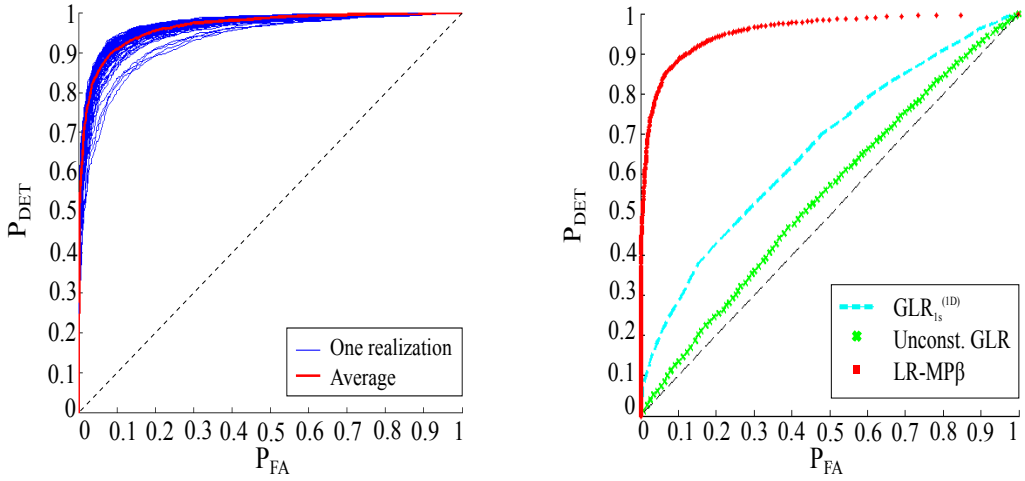
The performances of the test LR-MP$\beta$ are shown in Figure 6.4 for the same bright and faint spectra of Figure 6.3. When implementing the test with the faint spectrum estimate $\hat{\mathbf{x}}_f^* = \hat{\beta}\boldsymbol{\Sigma}_f^{-\frac{1}{2}}\mathbf{H}\hat{\mathbf{x}}_b'$, we get one particular ROC curve. Figure 6.4(a) plots hundreds of such ROC curves, for different noisy realizations of the bright spectrum, each realization leading to a different estimate $\hat{\mathbf{x}}_f^*$. The resulting average ROC is shown in red. Figure 6.4(b) compares the detection performances of the LR-MP$\beta$ (6.34) to those of the unconstrained GLR test (5.9), and of the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ test (5.35). The LR-MP$\beta$ (in red) shows better performances than the unconstrained GLR (in green) and the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ (in cyan), taking advantages from spatial dependencies existing between the two considered spectra.

### 6.2.3 Multiple-round detection strategy and analysis of the corresponding $\mathrm{P_{FA}}$ and $\mathrm{P_{DET}}$

#### 6.2.3.1 Multiple-round detection strategy

Aiming to improve detection performances on spectra that are too faint to be directly detected with the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ approach (5.35), but that actually do contain salient information because contiguous to brighter informative ones, we propose the following multiple-round detection strategy for the whole data cube, illustrated in Figure 6.5.

After performing detection with the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ test for a target $\mathrm{P}_{\mathrm{FA}_0}(\eta)$ (5.23), we are




(a) (b)

Figure 6.4: Detection performances of the LR-MP$\beta$ test: 6.4(a) LR-MP$\beta$ ROC curves obtained for 100 experiments. 6.4(b) Compared ROC curves of the LR-MP$\beta$ (in red), the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ (in cyan) and the unconstrained GLR (in green) tests for the faint spectrum shown in blue in Figure 6.3.

left with a set $\mathbf{\Gamma}$ of spectra flagged as "detected" (in green, in Figure 6.5(a)), and the complementary set $\bar{\mathbf{\Gamma}}$ of spectra in which no feature was found at a significance level set by $\mathrm{P}_{\mathrm{FA}_0}$ (in blue, in Figure 6.5(a)). Among all the spectra in $\bar{\mathbf{\Gamma}}$, we focus on those which are contiguous to detected ones (in brown, in Figure 6.5(b)). We take those spectra to perform a *second round* of detection with the LR-MP$\beta$ test: the spectra for which detection was found with the LR-MP$\beta$ test are shown in red in Figure 6.5(c). These spectra are aggregated to the set of "detected" spectra $\mathbf{\Gamma}$ (in green, in Figure 6.5(d), which are the union of the green and red pixels in Figure 6.5(c)). New contour pixels are then computed (brown pixels in Figure 6.5(d)), on which the LR-MP$\beta$ test can be performed again until no adjacent detection is found.

Detection results of the multiple-round detection approach will be reported in section 8.2 on astrophysical 3D data sets.

#### 6.2.3.2 Analysis of the $\mathrm{P}_{\mathrm{FA}}$ and $\mathrm{P}_{\mathrm{DET}}$ for more than one detection round

We address the problem of the analytical computation of the $\mathrm{P}_{\mathrm{FA}}$ and the $\mathrm{P}_{\mathrm{DET}}$ expressions when more that one detection round is performed. In fact, when cascading two (or more) detection tests, the overall false alarm and detection rates increase with respect to the case where only one test is performed.

For one faint spectrum $\mathbf{z}_f$ in $\bar{\mathbf{\Gamma}}$ contiguous to one bright spectrum $\mathbf{z}_b$ in in $\mathbf{\Gamma}$, a *two-step* testing procedure of the previous section cascades two tests: the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ and the LR-MP$\beta$. For the first test, the $\mathrm{P}_{\mathrm{FA}_0}$ depends only on the threshold $\eta$, while the probability of detection $\mathrm{P}_{\mathrm{DET}0} = Pr(\max_i(|\boldsymbol{d}_i{}^t\mathbf{z}|) > \eta \mid \mathcal{H}_1)$ depends on $\eta$ and on the considered spectrum $\mathbf{z}_f$. For the LR-MP$\beta$, the $\mathrm{P}_{\mathrm{FA}}$ depends on the fixed threshold $\xi$, while the $\mathrm{P}_{\mathrm{DET}}$

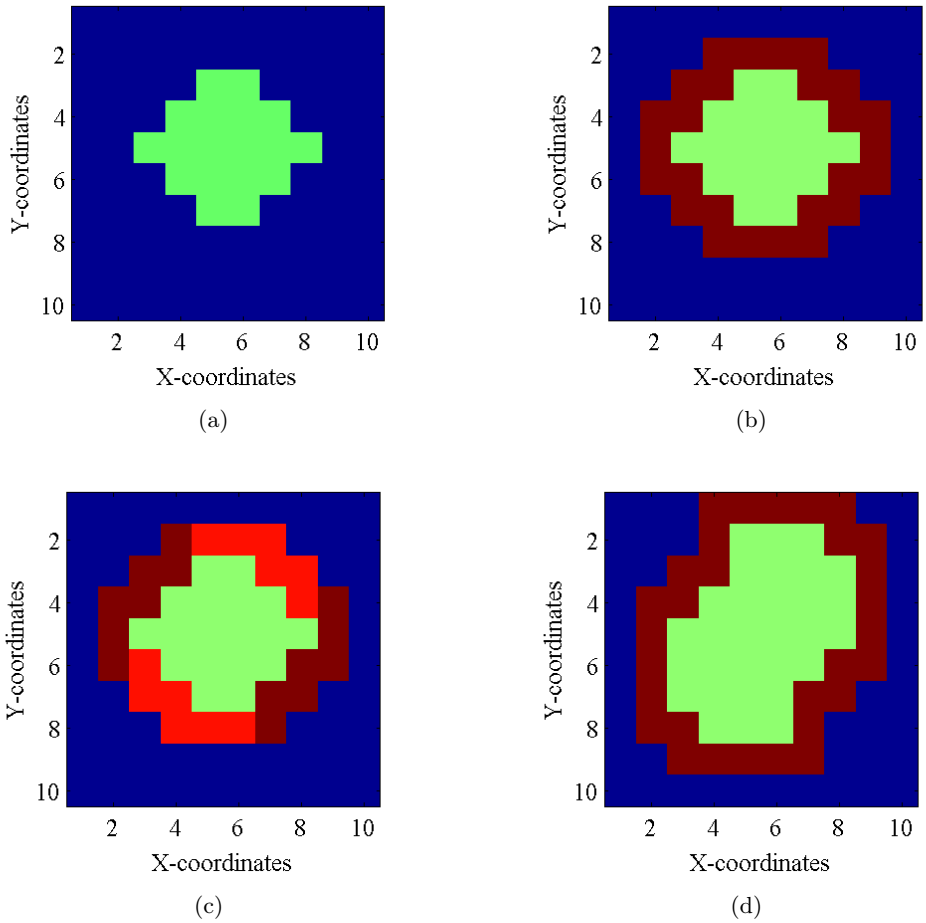



Figure 6.5: An illustration of the multiple-round detection strategy.

depends on $\xi$, $\mathbf{z}_f$ and on the bright neighbor $\mathbf{z}_b$. For those two consecutive tests performed on $\mathbf{z}_f$, a FA may occur:

i) when the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ makes a FA on $\mathbf{z}_f$, which happens with probability $\mathrm{P}_{\mathrm{FA}_0}(\eta)$;

ii) when there is:

- no FA by $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ on $\mathbf{z}_f$ (probability $(1-\mathrm{P}_{\mathrm{FA}_0}(\eta))$),
- and detection of the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ test on the spectrum $\mathbf{z}_b$ (probability $\mathrm{P}_{\mathrm{DET}0}(\mathbf{z}_f,\eta)$),
- and FA of LR-MP$\beta$ with $\mathbf{z}_b$ (probability $\mathrm{P}_{\mathrm{FA}}(\mathbf{z}_b,\xi)$).

The probability of detection of the two tests can be derived similarly, which gives:

$$\begin{cases} {\mathrm{P}_{\mathrm{FA}}}^{2tests}(\mathbf{z}_f) = \mathrm{P}_{\mathrm{FA}_0}(\eta) + (1-\mathrm{P}_{\mathrm{FA}_0}(\eta))\mathrm{P}_{\mathrm{DET}0}(\mathbf{z}_b,\eta)\mathrm{P}_{\mathrm{FA}}(\mathbf{z}_b,\xi) \\ {\mathrm{P}_{\mathrm{DET}}}^{2tests}(\mathbf{z}_f) = \mathrm{P}_{\mathrm{DET}0}(\mathbf{z}_f,\eta) + (1-\mathrm{P}_{\mathrm{DET}0}(\mathbf{z}_f,\eta))\mathrm{P}_{\mathrm{DET}0}(\mathbf{z}_b,\eta)\mathrm{P}_{\mathrm{DET}}(\mathbf{z}_b,\xi), \end{cases} \tag{6.36}$$

where the characteristics of the first stage are obtained by (3.48) and (3.49) and the characteristics of the second stage are given by the red (average) curve in Figure 6.4(a).

The expression of the $\mathrm{P_{FA}}$ and $\mathrm{P_{DET}}$ for a multiple-round detection method can be finally generalized to the case in which, given a faint spectrum $\mathbf{z}_f$, the bright spectrum $\mathbf{z}_b$ is chosen as one among the 8 neighbors of $\mathbf{z}_f$. The expressions in (6.36) thus become

$$\begin{cases} \mathrm{P_{FA}}^{2tests}(\mathbf{z}_f) = \mathrm{P_{FA_0}}(\eta) + (1 - \mathrm{P_{FA_0}}(\eta)) \times \frac{1}{8} \sum_{i=1}^{8} \mathrm{P_{DET0}}(\mathbf{z}_{b_i}, \eta) \mathrm{P_{FA}}(\mathbf{z}_{b_i}, \xi) \\ \mathrm{P_{DET}}^{2tests}(\mathbf{z}_f) = \mathrm{P_{DET0}}(\mathbf{z}_f, \eta) + (1 - \mathrm{P_{DET0}}(\mathbf{z}_f, \eta)) \times \frac{1}{8} \sum_{i=1}^{8} \mathrm{P_{DET0}}(\mathbf{z}_{b_i}, \eta) \mathrm{P_{DET}}(\mathbf{z}_{b_i}, \xi). \end{cases} \tag{6.37}$$

Results obtained in terms of FA rate, when performing the cascade of two tests, will be shown in section 8.2, in the context of astrophysical 3D data sets.

A first encouraging attempt to account for the spatio-spectral dependencies of neighbors spectra in a 3D data set was presented in this section. Nevertheless, performing such multiple-round detection approach is complicated, especially with regard to the analytical computation of the overall $\mathrm{P_{FA}}$ and $\mathrm{P_{DET}}$. We thus introduce in the next section a detection strategy based on a new data model, in which both the effects of the WSF and of the SSF are explicitly taken into account and which is also simple to control and implement. Based on this model, a detection strategy taking advantage from dedicated dictionaries composed of 3D elementary atoms is proposed.

## 6.3 A (spatio-spectral) 3D model-based strategy

The data model (5.1) and (5.25) account only for the effects of the instrument's WSF, through the use of the matrix $\mathbf{H}$. In particular, the spectral content of each 1D atom $\mathbf{r}_i$, $i = 1, \ldots, L$ of the $\mathbf{R}$ dictionary is assumed to be spread by $\mathbf{H}$ in the $\lambda$ direction. However, we further assume that at every single wavelength $\lambda$, the instrument's SSF also dilutes spatially the information over a number of $(N_x \times N_y)$ adjacent columns. With the aim of further improving the detection performances of our tests, this section introduces a 3D spatio-spectral data model that also accounts for the effects of the instrument' SSF.

To our knowledge, few related detection approaches exist. Among these, [Starck 2009] proposes a detection/reconstruction method for low flux sources using a wavelet-based spatio-spectral dictionary, along with a multi-scale variance stabilization transform. In contrast to our approach, the dictionary of [Starck 2009] is generic, and does not account for the instrumental spreading functions. We also note the use of (one-dimensional) specific spectral dictionaries in [Chan 2012, Ramos 2012]. The first paper considers the case of Poisson noise with a regularization term favouring group sparsity, and the second paper a dictionary of Gaussian lines in conjunction with a Bayesian version of the support-vector machine-learning technique [Tipping 2012].

In the following, the 3D model together with the corresponding expression of the constrained GLR test are derived. We also discuss the choice made on the strategy used to design the considered dictionary. Numerical results, which confirm the superiority of the 3D approach with respect to the previous ones, will be shown in section 8.3.

### 6.3.1 3D data Model and $\mathrm{GLR}_{1\mathbf{s}}^{(3\mathbf{D})}$ detection test

The spread of information in the three dimensions can be represented by the composition of the signal feature with both the instrument's WSF and SSF. For the simplicity of exposition we consider that the SSF is spectrally invariant, but, as specified below, this is not necessary. In a given data cube, to each spectral column $\mathbf{s}(x, y)$, with $x = 1, \ldots, X$ and $y = 1, \ldots, Y$, is associated a data sub-cube of dimension $(N_x \times N_y \times \Lambda)$ where the energy is smeared because of the SSF. We explicit the data model in vector form. All the $(N_x \times N_y)$ $\Lambda$-dimensional spectra in each such data sub-cube are stacked one on top of the other to form a $N_x N_y \Lambda$ column vector $\mathbf{s}_V$. This leads to the following model:

$$\mathcal{H}_1 : \mathbf{s}_V(x, y) = \mathbf{F}(x, y)\mathbf{H}\mathbf{R}\boldsymbol{\alpha} + \boldsymbol{\epsilon}_V, \tag{6.38}$$

where $\mathbf{s}_V$ and $\boldsymbol{\epsilon}_V \in \mathbb{R}^{N_x N_y \Lambda}$, and $\boldsymbol{\epsilon}_V \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_V(x, y))$ is a Gaussian noise vector characterized by a $(N_x N_y \Lambda \times N_x N_y \Lambda)$ diagonal covariance matrix $\boldsymbol{\Sigma}_V$, in which the diagonal is the vectorized form of the associated sub-cube containing the noise variances of each voxel. Here, the matrix $\mathbf{H}$, the dictionary $\mathbf{R}$, and the coefficient vector $\boldsymbol{\alpha}$ are the same as in model (5.1) and (5.25), while $\mathbf{F}(x, y)$ represents the $(N_x N_y \Lambda \times \Lambda)$ matrix composition by the SSF. If we denote by

$$\boldsymbol{f} = \begin{bmatrix} f_{11} & \cdots & f_{1N_y} \\ f_{21} & \cdots & f_{2N_y} \\ \vdots & & \vdots \\ f_{N_x 1} & \cdots & f_{N_x N_y} \end{bmatrix} \tag{6.39}$$

the array of the SSF coefficients at position $(x, y)$ (that we first assume invariant in the $\lambda$ dimension, for simplicity), and by

$$\boldsymbol{\phi} = \mathrm{vec}\boldsymbol{f} = [f_{11} \ldots f_{N_x 1}, \ \ldots \ , f_{1N_y} \ldots f_{N_x N_y}]^t \tag{6.40}$$

its vector form, then

$$\begin{aligned} \mathbf{F} &= [f_{11}\boldsymbol{I}_\Lambda \ \ldots \ f_{N_x N_y}\boldsymbol{I}_\Lambda]^t \\ &= \boldsymbol{\phi} \ \otimes \ \boldsymbol{I}_\Lambda, \end{aligned} \tag{6.41}$$

where $\otimes$ denotes the Kronecker product and $\boldsymbol{I}_\Lambda$ the $\Lambda$-dimensional Identity matrix.

Note that when the SSF is not spectrally invariant, $\mathbf{F}$ should be replaced by

$$\mathbf{F} = [\mathrm{diag}(\boldsymbol{f}_{11}), \ldots, \mathrm{diag}(\boldsymbol{f}_{N_x N_y})]^t, \tag{6.42}$$

where

$$\boldsymbol{f}_{ij} = [f_{ij}(\lambda_1), \ldots, f_{ij}(\lambda_\Lambda)], \tag{6.43}$$

with $i = 1, \ldots, N_x$, $j = 1, \ldots, N_y$.

According to (6.38), in the weighted data domain we obtain, for each spectral column $\mathbf{s}(x, y)$, the following model for the associated vectorized data cube $\mathbf{s}_V$:

$$\mathcal{H}_1 : \boldsymbol{\Sigma}_V^{-\frac{1}{2}}\mathbf{s}_V = \boldsymbol{\Sigma}_V^{-\frac{1}{2}}\mathbf{F}\mathbf{H}\mathbf{R}\boldsymbol{\alpha} + \mathbf{w}_V, \quad \|\boldsymbol{\alpha}\|_0 = 1, \tag{6.44}$$

where $\boldsymbol{w}_V \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}_{N_x N_y \Lambda}\right)$ and the dependence on $(x, y)$ of $\boldsymbol{\Sigma}_V$ and $\mathbf{F}$ is made implicit for the clarity of exposition. Similarly to section 5.2, we refer to $\boldsymbol{\Sigma}_V^{-\frac{1}{2}}\mathbf{FHR} = \mathbf{D}_{\boldsymbol{\Sigma}\mathbf{FH}}$ as an equivalent dictionary. Noting

$$\begin{aligned} \boldsymbol{z}_V &= \boldsymbol{\Sigma}_V^{-\frac{1}{2}}\mathbf{s}_V, \\ \mathbf{D}_V &= \boldsymbol{D}_{\boldsymbol{\Sigma}\mathbf{FH}}\mathbf{N}_{\mathbf{D}_{\boldsymbol{\Sigma}\mathbf{FH}}}^{-1} \\ \text{and} \quad \boldsymbol{\theta} &= \mathbf{N}_{\mathbf{D}_{\boldsymbol{\Sigma}\mathbf{FH}}}\boldsymbol{\alpha}, \end{aligned} \tag{6.45}$$

with $\mathbf{N}_{\mathbf{D}_{\boldsymbol{\Sigma}\mathbf{FH}}}$ the diagonal matrix composed of the norms of the columns of $\mathbf{D}_{\boldsymbol{\Sigma}\mathbf{FH}}$, (6.44) leads to the model

$$\mathcal{H}_1 \ : \ \boldsymbol{z}_V = \mathbf{D}_V\boldsymbol{\theta} + \mathbf{w}_V, \quad \|\boldsymbol{\theta}\|_0 = 1, \tag{6.46}$$

which corresponds to the expression of model (5.26), with the crucial difference that this time both the WSF and SSF are taken into account in the normalized dictionary $\mathbf{D}_V$.

We now derive the expression of the constrained 3D GLR detection test for model (6.46). Similarly to what said in section 5.3 for the 1D spectral model (5.25), the GLR test for (6.46) yields

$$\mathrm{GLR}_{1\mathrm{s}}^{(3\mathrm{D})} \ : \ \frac{\max\limits_{j,\ \theta_j} p\left(\mathbf{z}_V \mid \mathbf{d}_{V_j}, \theta_j\right)}{p\left(\mathbf{z}_V \mid \mathbf{0}\right)} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma_{1\mathrm{s}}. \tag{6.47}$$

Following the same steps as those described in (5.29) - (5.36), we obtain

$$\mathrm{GLR}_{1\mathrm{s}}^{(3\mathrm{D})} \ : \ T_{\mathrm{GLR}_{1\mathrm{s}}}^{(3\mathrm{D})}(\mathbf{z}_V) \underset{H_0}{\overset{H_1}{\gtrless}} \xi_{1\mathrm{s}}, \tag{6.48}$$

where $\xi_{1\mathrm{s}} = 2\ln\gamma_{1\mathrm{s}}$. The expression of the $\mathrm{GLR}_{1\mathrm{s}}^{(3\mathrm{D})}$ test statistic gives

$$T_{\mathrm{GLR}_{1\mathrm{s}}}^{(3\mathrm{D})}(\mathbf{z}_V) = \max_j \left|\boldsymbol{d}_{V_j}^T\mathbf{z}_V\right|. \tag{6.49}$$

Again, the $\mathrm{GLR}_{1\mathrm{s}}^{(3\mathrm{D})}$ is equivalent to the *Max* test of [Arias-Castro 2010] applied to the vector form of the 3D atoms $\boldsymbol{d}_{V_j}$ in $\mathbf{D}_V$ and of the data $\mathbf{z}_V$. Note that we do not attempt to first deconvolve the data, and then detect. The reason is that in such a two-step approach, the dependence of deconvolution on a regularization term would influence the data distribution. In addition, deconvolution somewhat performs a detection, but without rigorously defining the realized detection test (nor controlling its size and power). Thus, detection would be performed twice. In contrast, our approach is one-shot: the sparsity based test essentially performs information re-concentration to improve detection (quite in the spirit of matched filtering principles), while keeping the statistical description of the data accurate.

The threshold $\xi_{1\mathrm{s}}$ controlling the $\mathrm{P_{FA}}$ is computed numerically in function of the $\xi_{1\mathrm{s}}$ threshold. An example of the empirical computation of the $\mathrm{P_{FA}}$ as a function of the corresponding test threshold will be given in subsection 7.3.3. Detection results in the context of astrophysical 3D data sets will be shown in section 8.3.

### 6.3.2 Design strategies for the choice of the dictionary

An important point is represented by the choice of the spectral dictionary $\mathbf{R}$. As discussed in subsection 2.2.2, there are several ways to choose $\mathbf{R}$. Of course, this dictionary can be

taken as a generic one (for instance wavelets). However, if some atoms of such dictionary are known not to well represent the data to be detected, they should be removed, as they will essentially increase the $P_{FA}$ but only marginally increase the $P_{DET}$. In addition, when a representative data set of the signals under $\mathcal{H}_1$ is available, it may be more efficient to design a specific dictionary. In such cases, $\mathbf{R}$ is often simply taken as the mean or the first singular vector obtained by a SVD of the data set [Manolakis 2009], or it can be optimized using now classical sparse dictionary techniques, as for instance the K-SVD approach of [Aharon 2006] (see Appendix B.3). Another set of approaches called minimax seeks a dictionary of reduced dimension, which maximizes the worst probability of detection under $\mathcal{H}_1$ [Lehmann 2005].

We note an important issue in the case where a spectral library is available. Assume this (possibly very large) library $\boldsymbol{\mathcal{L}}$ is so accurate that the signal under $\mathcal{H}_1$ can be considered to belong to $\boldsymbol{\mathcal{L}}$. In this case, it may be not only computationally prohibitive, but also suboptimal (w.r.t. detection power) to implement a test of the form (6.48) with $\mathbf{R} = \boldsymbol{\mathcal{L}}$ (*i.e* the concatenation of all possible alternatives). This is because the $P_{FA}$ may increase wildly with the number of alternatives [Suleiman 2013]. Hence, dictionaries with reduced dimensions such as those cited above are very useful. The design of the dictionary also critically influences the success of a detection strategy.

As a final remark, we comment on the one-sparse hypothesis made in (5.25) and (6.44). If the signals under $\mathcal{H}_1$ are sparsely decomposed in $\mathbf{R}$ using a few atoms of rapidly decreasing amplitudes, then only the largest feature will be detectable in highly noisy data. On the other hand, if $\mathbf{R}$ is sufficiently well designed/large and thus contains (up to an amplitude factor) an atom equal to the signal under $\mathcal{H}_1$, we can also consider that the signals are one-sparse in $\mathbf{R}$. Hence, the $1-$sparse hypothesis is justified either because of high noise level or because it is indeed true.

A comparison between the 1D (5.29) and the 3D (6.48) one-sparse GLR tests using the K-SVD [Aharon 2006] and the minimax dictionary learning approaches of [Suleiman 2013] will be reported in section 8.3, in the context of astrophysical hyperspectral data.

## 6.4 Conclusion

With the aim to improve detection with respect to strategies based only on spectral models for 3D data sets, we have investigated different approaches exploiting both the spatial and the spectral dependencies between neighbors spectra in 3D data cubes, through the 1D and 2D PSF, and a sparse spectral model using an appropriate dictionary.

First, several strategies assuming eventually that neighbors spectra are proportional have been derived.

Second, we have introduced a new detection method that allies dictionary learning techniques, sparse models and spread functions to concentrate information prior to detection. This method, based on a data model that takes into account both the WSF and SSF instrument's effects, is currently generalized to the case where sources are spatially structured (instead of punctual).

In the next part of the manuscript, the dictionary model-based detection strategies introduced in chapter 5 and chapter 6 are applied to signal detection in highly noisy astrophysical hyperspectral data.

# Part III

# Detection of Faint Sources in Astrophysical Hyperspectral Data

Chapter 7

# Detection results using spectral model based strategies

Contents


**7.1 Introduction . . . . . 115**
**7.2 Hyperspectral Imaging and its application to Astrophysics . . . . 116**
7.2.1 Spectroscopy in Astronomy . . . . . 116
7.2.2 Hyperspectral Imaging . . . . . 118
7.2.3 An instance of Hyperspectral Imaging in Astrophysics: The MUSE instrument . . . . . 119
7.2.4 MUSE instrument's Point Spread Function . . . . . 120
**7.3 Sparsity-constrained vs unconstrained GLR test . . . . . 121**
7.3.1 The design of the dictionary $\mathbf{R}$ . . . . . 122
7.3.2 Comparison of the detection performances on a single test spectrum . 123
7.3.3 Comparison of the detection performances on hyperspectral data cubes 126
7.3.4 Analysis of the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ test detection results in terms of $p$-value . . . 129
**7.4 Conclusion . . . . . 131**


## 7.1 Introduction

Thanks to their rich informative content, multi-dimensional data sets are exploited in an increasing number of modern applications. For instance, 3D hyperspectral images can be acquired by spectro-imagers combined with microscopes, satellites or even space probes and be then used successfully in different domains such as bio-medicine [Vo-Dinh 2004], defense [Manolakis 2003] and astrophysics [Starck 2009]. Astrophysics is the field we are interested in the applicative part of our work.

In this chapter, a brief overview on Hyperspectral Imaging and some of its applications to Astrophysics is firstly presented in section 7.2. The ESO's MUSE (Multi Unit Spectroscopic Explorer) instrument and its characteristics are introduced in subsection 7.2.3 and 7.2.4. Conceived to be installed at the Very Large Telescope in Chili in 2013, MUSE is an integral field spectrograph that will provide a huge quantity of hyperspectral data. As the instrument is still under construction, all the numerical simulations presented hereafter are performed on simulated, but very realistic, hyperspectral data cubes coming from the consortium in charge of MUSE's assembly.

We then illustrate the detection results obtained by applying the sparsity-constrained $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ (5.35) and the unconstrained GLR (5.9) detection strategies to MUSE data in section 7.3. After a discussion concerning the choice of the redundant dictionary $\mathbf{R}$ on which the considered data model (5.2) is based (subsection 7.3.1), results on astrophysical data are shown in the case of a single test spectrum (subsection 7.3.2) and in the case of MUSE data cubes (subsection 7.3.3).

## 7.2 Hyperspectral Imaging and its application to Astrophysics

### 7.2.1 Spectroscopy in Astronomy

Astronomers will unfortunately never be able to bring a star in their laboratory and study it. Astronomical discoveries have thus been and continue to be essentially made through the study of the light emitted by the stars. While this study can be made considering the light as a whole, it is also possible to divide it into its component parts to better understand the nature of the observed emitting celestial objects.

Sir Isaac Newton was the first who in the 17th century scientifically proved that the rainbow colors coming out of a prism hit by a light beam were not due to impurities in the glass but to the light itself [Newton 1672]. Newton adopted the Latin word *spectrum* ("*the means to see*") to describe the whole set of these colors. Newton's beliefs were based on the corpuscular theory, according to which light was made of a stream of particles traveling in a narrow beam. This was in contrast with the "Huygens' principle", by the dutch physicist Christiaan Huygens, who in his *"Traité de la lumière"* [Huygens 1690] had theorized light was made of moving waves propagating through the free space at the speed of light. It was only in the 19th century with its *wave theory of light* that Thomas Young [Young 1802] made a significant contribution in the particle versus wave theory debate, in favor of Huygens' wave theory. Nowadays this duality persists, representing still a fundamental property of light.

In the meanwhile, another well known scientist named Sir Frederick William Herschel, improved on Newton's results measuring the temperature of each color of the light: blue and green were cooler, while red was warmer. Herschel also (accidentally) discovered the presence of a warmer form of light beyond the red light [Herschel 1800], opening the door to the study of types of electromagnetic radiation invisible to the human eye. That region became known as *infrared* radiation. Today we know that visible light is only a small part of the set of all possible frequencies of electromagnetic radiation. In particular, the visible spectrum covers all the wavelengths (the distance a wave travels during one oscillation) between 0.4 $\mu$m (ultraviolet limit) and 0.8 $\mu$m (infrared limit).

In astronomy, what it is called spectrum of a source is the whole set of radiations that the observed object emits. The field of physics that allows to experimentally study the spectrum of a physical phenomenon is called *spectroscopy*. The rise of spectroscopy has begun in the 19th century. The development of the first spectroscope is attributable to Joseph von Fraunhofer who in 1819, thanks to this tool, was able to identify absorption lines in the solar spectrum [Brand 1995] then named as *Fraunhofer lines*. A spectroscope allows to separate the various components of a beam of light along wavelengths. If combined

with a device for measuring the intensity of light at the various wavelengths, it is called a spectrometer. The sampling of the spectrum together with the range of covered wavelengths depend on the optical devices of the spectrograph.

The spectrum of an astronomical object contains many key informations about its chemical and physical composition. This knowledge also provides informations on the past history of the object and on its evolution. Depending on its composition, a spectrum is generally characterized by a specific combination of continuum, absorption and/or emission spectral lines. The more chemical elements form the observed object, the more its spectrum is complicated, presenting the spectral signature of each one of the composing elements. An example of a star and of a a galaxy spectrum is illustrated in Figure 7.1 in blue and red, respectively.

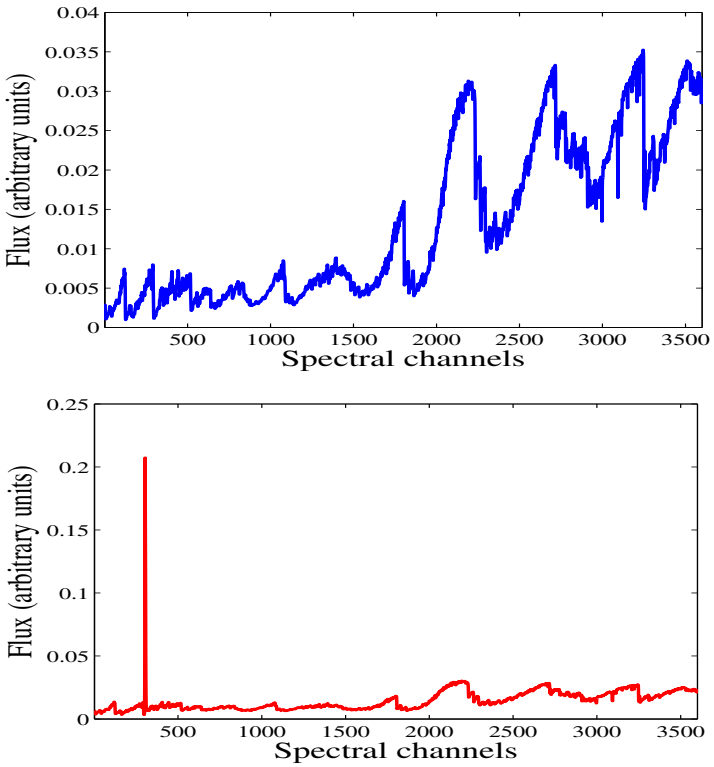



Figure 7.1: Example of a star (in blue) and galaxy (in red) spectrum.

Obviously, studying only the spectral characteristics of an object leads to neglect the spatial information. Over the last decades, scientists and engineers have thus focused their interest on how to extend the information provided by spectroscopic analysis at all the points of a given spectral scene, such as for example a telescope's field of view. These efforts have given rise to what is known today as *Imaging Spectroscopy*, *Hyperspectral Imaging* or *Integral Field Spectroscopy* in astronomy: a set of multi-band techniques allowing to simultaneously acquire both the spectral and the spatial types of information. A brief survey is presented in the next section.

### 7.2.2 Hyperspectral Imaging

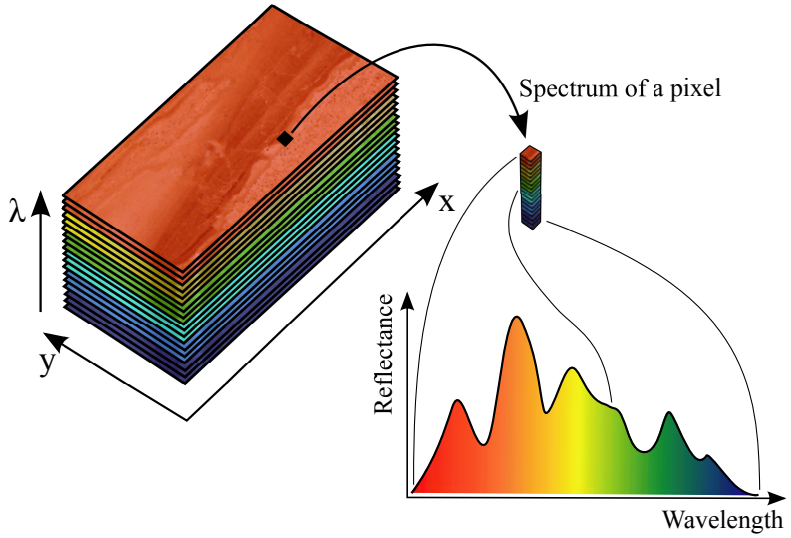



Figure 7.2: Example of hyperspectral image and spectrum associated to a pixel

Hyperspectral Imaging (HSI) combines the power of digital imaging with that of spectroscopy. In fact, HSI allows to represent a same observed scene over several narrow spectral bands, in different ranges of wavelengths (visible, infrared, and more). To each spatial position in the image thus corresponds a well-defined spectral information. Data obtained through HSI are arranged in cubes (data cubes) that are characterized by two spatial dimensions $(x, y)$ and a spectral dimension $\lambda$. An example of hyperspectral image is illustrated in Figure 7.2. The prefix *hyper-* was coined referring to the large number of wavelengths at which spectra in hyperspectral data cubes are sampled. The literature in fact differentiates between:

- *Multi*-spectral imagers, that allow data collection in a dozen of spectral bands;

- *Hyper*-spectral imagers, that allow data collection in several hundreds or even thousands of spectral bands.

Multi-dimensional data sets often possess intrinsic characteristics that lead to processing difficulties. First, the large data size: for example, hyperspectral images made up of hundreds spectral bands require increasing storage capacities and fast processing algorithms. Second, the poor SNR: splitting the information over three dimensions and increasing the sampling resolution often leads to relatively stronger noise levels in the voxel elements. In addition, the features of interest may be intrinsically weak and (spatially) sparse: an example is provided by the astrophysical data considered in this chapter. In such cases, it may be extremely difficult to *detect* the presence of information, because it is both rare and weak. This is the situation considered all along this chapter.

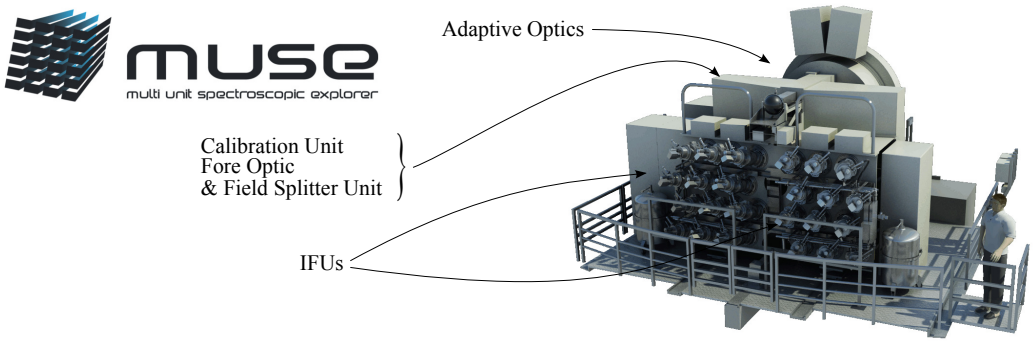



Figure 7.3: A 3D illustration of the MUSE instrument. Comparison of the instrument's dimension with respect to the human being on the right of the Figure. (Original picture from ESO's website)

### 7.2.3 An instance of Hyperspectral Imaging in Astrophysics: The MUSE instrument

The Multi Unit Spectroscopic Explorer (MUSE) is an integral field spectrograph in development for one of the four 8 meters telescopes of the European Southern Observatory's Very Large Telescope (VLT). The MUSE instrument will allow to obtain the spectroscopic equivalent of the *Hubble Deep Fields* [Williams 1996]. Moreover, the main scientific goals of this instrument will concern: the study of stellar population in nearby galaxies, the study of supermassive black holes, planets and small bodies, and the study of the formation and evolution of galaxies. In the latter case, the MUSE instrument will enable to study the progenitor of normal galaxies, which are also the farthest galaxies from the Earth.

One of the major challenges of the MUSE instrument concerns the detection and characterization of such distant galaxies. Those galaxies, which are more than ten billion light years far away from Earth, appear as spatially localized within one or a few pixels (because of the instrument's spatial Point Spread Function, as explained in the next section) and present spectra characterized by a particular, very faint and narrow line, known as Lyman-alpha line (Ly-$\alpha$). Examples of Ly-$\alpha$ *objects* and corresponding emitting lines are given in section 8.3 (see Figure 8.3 and 8.8(a)). As it will be emphasized in section 8.3, the huge distance together with the low luminosity of the objects and the disruptive effects of the atmosphere, make the detection of those sources very difficult.

The MUSE instrument combines the power and the accuracy of 24 Integral Field Units (IFU). These 3D spectrographs allow to cover a very large field of view and to simultaneously obtain a spectrum for each one of its points (pixels). Figure 7.3 shows how the MUSE instrument will appear once assembled at the VLT.

More specifically, MUSE will deliver hyperspectral data cubes composed of 300×300 spectra sampled at approximately 3600 wavelengths of the visible spectrum, with a spectral resolution of 0.13 nm and in a spectral range going from 0.48 to 1 $\mu$m [Kosmalski 2011]. Most data acquired by the MUSE instrument will be characterized by a very low SNR. In particular, data will be strongly contaminated by the spectral signature of atmospheric molecules. In addition, data will be largely affected by Poisson noise, which is indeed data

dependent. Moreover, the instrumental detection efficiency is variable with wavelength [Bourguignon 2011]. Consequently, the noise level is highly variable from one wavelength to another, and also from one spectrum to another. An example of noise variance involved in one of the MUSE's spectra is given in Figure 7.4.

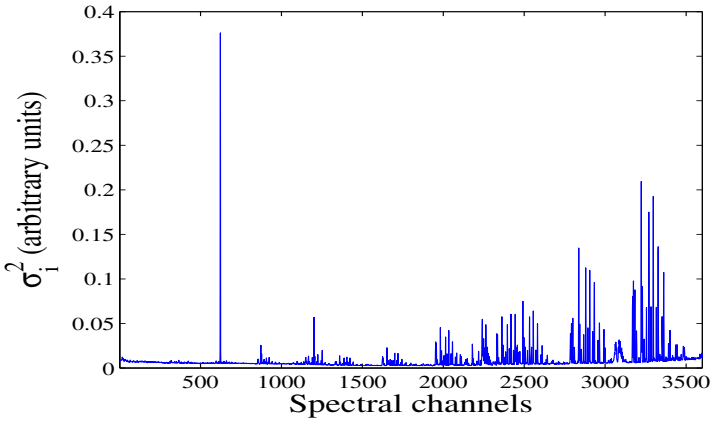



Figure 7.4: Example of noise variances affecting the components of a MUSE spectrum.

### 7.2.4 MUSE instrument's Point Spread Function

MUSE is characterized by a three-dimensional Point Spread Function (PSF), which is variable in $(x, y)$ and $\lambda$. As previously defined in chapter 6, the PSF is the mathematical function that describes the instrument acquisition system's response to a point source. For an astronomical object such as a point star, the effect of the instrument's PSF is visible by a blurring of the information over all the data dimensions. When the instrument can be considered as linear and invariant, the data are modeled by a convolution of the object (celestial body) by a function (kernel) that mathematically describes the blur.

In the scope of this work, MUSE is considered as a linear system. MUSE's PSF can thus be separated into two distinct functions: a Spatial Spread Function (SSF), which covers $(13 \times 13)$ pixels in the spatial domain, and a Wavelength Spread Function (WSF), which spreads over 7 spectral elements. Both the WSF and the SSF of the instrument slightly vary over the $\Lambda = 3600$ wavelengths of MUSE data cubes. All along this chapter and also in the next one we will consider a constant approximation of the WSF and the SSF, obtained computing the mean over all wavelengths. We emphasize however that this is without any loss of generality because all the proposed methods could account for such variations.

As an illustration, MUSE's WSF and SSF are reported in Figure 7.5 and Figure 7.6, respectively. In particular, Figure 7.5(a) and Figure 7.5(b) show the instrument's WSF for the first ($\Lambda_1$) and the last ($\Lambda_{3600}$) wavelengths. Figure 7.5(c) shows the set of 3600 MUSE's WSFs (in blue) and the corresponding mean (approximate) WSF (in red), on which we work. Similarly, Figure 7.6(a) and Figure 7.6(b) show the instrument's SSF for the first wavelength ($\Lambda_1$) and the last one ($\Lambda_{3600}$). The mean SSF is reported in Figure 7.6(c). These approximations (Figures 7.5(c) and 7.6(c)) are not critical given the small variations of the WSF and the SSF in $\Lambda$ and the high level of noise that affects the data.

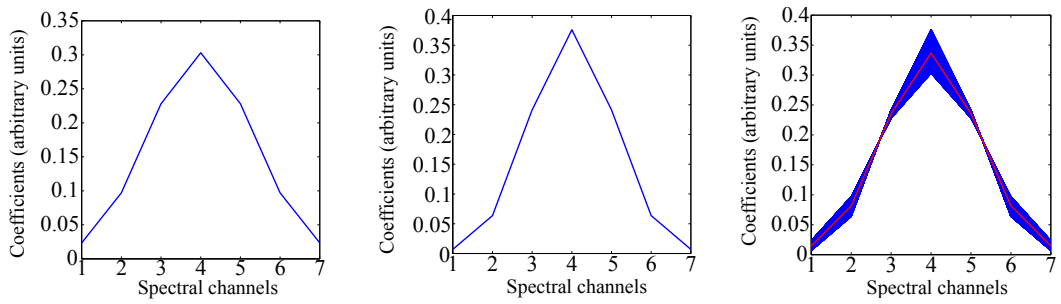



Figure 7.5: MUSE's WSF. 7.5(a) WSF at $\Lambda_1$. 7.5(b) WSF at $\Lambda_{3600}$. 7.5(c) WSFs for all the wavelengths (in blue) and mean (approximate) WSF (in red).

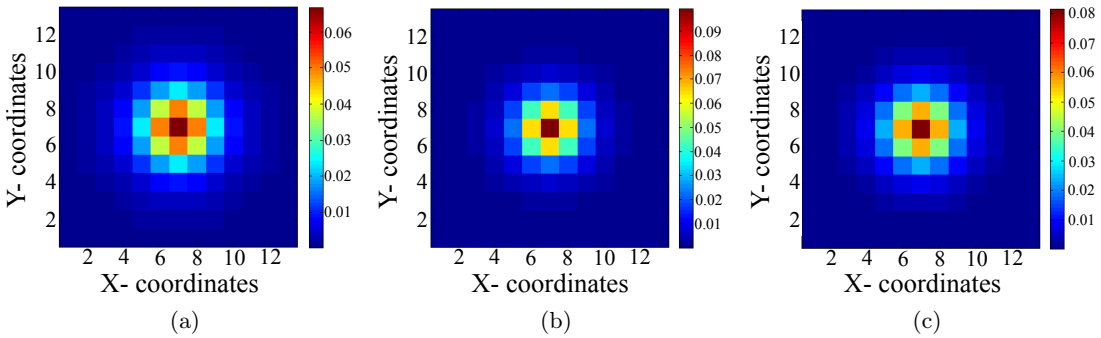



Figure 7.6: MUSE's SSF. 7.6(a) SSF at $\Lambda_1$. 7.6(b) SSF at $\Lambda_{3600}$. 7.6(c) mean (approximate) SSF.

## 7.3 Sparsity-constrained vs unconstrained GLR test: comparison of the detection performances.

We present the detection results obtained on MUSE astrophysical hyperspectral data performing the sparsity-constrained $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ ((5.35), where the (1D) apex helps to distinguish this test from the 3D model-based $\mathrm{GLR}_{1\mathrm{s}}^{(3\mathrm{D})}$ detector considered in the next chapter) and the unconstrained GLR (5.9) approaches introduced in chapter 5. Both strategies are based on the 1D (spectral) whitened data model of (5.2)

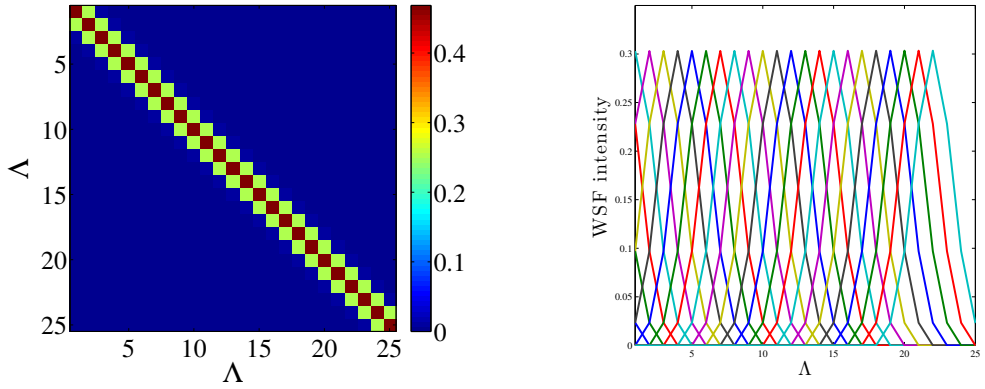
$$\begin{cases} \mathcal{H}_0 : \boldsymbol{\Sigma}^{-1/2}\mathbf{s} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\epsilon} = \mathbf{w}, & \mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \\ \mathcal{H}_1 : \boldsymbol{\Sigma}^{-1/2}\mathbf{s} = \boldsymbol{\Sigma}^{-1/2}\mathbf{H}\mathbf{R}\boldsymbol{\alpha} + \mathbf{w}, \end{cases}$$

where $\mathbf{s} \in \mathbb{R}^{\Lambda}$ is the observation vector, $\mathbf{w} \in \mathbb{R}^{\Lambda}$ is the noise vector, $\boldsymbol{\alpha} \in \mathbb{R}^{L}$ is the unknown coefficient vector, $\boldsymbol{\Sigma}^{-1/2}$ represents the whitening matrix, $\mathbf{R}$ is a $(\Lambda \times L)$ known redundant dictionary (see next section), and the effect of the instrument WSF is taken into account through the use of its matrix form $\mathbf{H}$ so spectra are considered independently here.

As illustrated in Figure 7.7(a), $\mathbf{H}$ represents a $(\Lambda \times \Lambda)$ sparse matrix, whose columns

contain the translated replica of the MUSE's mean WSF (Figure 7.7(b)) over all the $\Lambda$ wavelengths.

A discussion on the design of dictionary $\mathbf{R}$ in the context of MUSE data is presented below.




(a) (b)

Figure 7.7: Matrix form of the Wavelength Spread Function. 7.7(a) Sparse diagonal $\mathbf{H}$ matrix (glimpse over the first 25 spectral bands). 7.7(b) $\mathbf{H}$ columns (mean WSF translated replica) plotted for the first 22 spectral bands.

### 7.3.1 The design of the dictionary R

Restoration of MUSE-like spectra was recently addressed in [Bourguignon 2011], where prior information was incorporated through sparsity constraints. A redundant dictionary $\mathbf{R}$ of elementary spectral features was built in accordance with astrophysical knowledge, so that the sparsely estimated nonzero components can be interpreted as physically meaningful features. More precisely, $\mathbf{R}$ concatenates three sub-dictionaries: $\mathbf{R} = [\mathbf{R}^\ell \, \mathbf{R}^c \, \mathbf{R}^b]$, each of which corresponds to a specific spectral component: a line spectrum, a continuous spectrum and a series of discontinuities. $\mathbf{R}^\ell$ ($\ell$ for lines) is a dictionary of discrete splines with several widths, which are centered along the reconstruction wavelength axis. Eleven width values were used, varying from 1 (delta functions) to 138 points. Delta functions and splines model respectively unresolved and resolved absorption or emission spectral lines. MUSE' spectral resolution equals 0.13 nm so that the maximal width equals $138 \times 0.13 \simeq 18$ nm. $\mathbf{R}^c$ ($c$ for continuum) is a dictionary of continuous spectra composed of sine functions with low frequencies (reduced frequencies vary from $1/N$ to $8/N$, where $N$ is the number of data) and 8 discretized phase shifts. It also includes the continuous component by means of a constant signal. Finally, dictionary $\mathbf{R}^b$ ($b$ for breaks) models a series of breaks in the spectrum, and is composed of step functions which are also centered along the wavelength axis. While the works [Bourguignon 2011] used this dictionary in the framework of spectral restoration, we use it for detection purpose here.

The redundant dictionary $\mathbf{R}$ considered in this chapter contains a total of $L = 27680$ atoms. Examples of such atoms are given in Figure 7.8: Figures 7.8(a), 7.8(b) and 7.8(c)

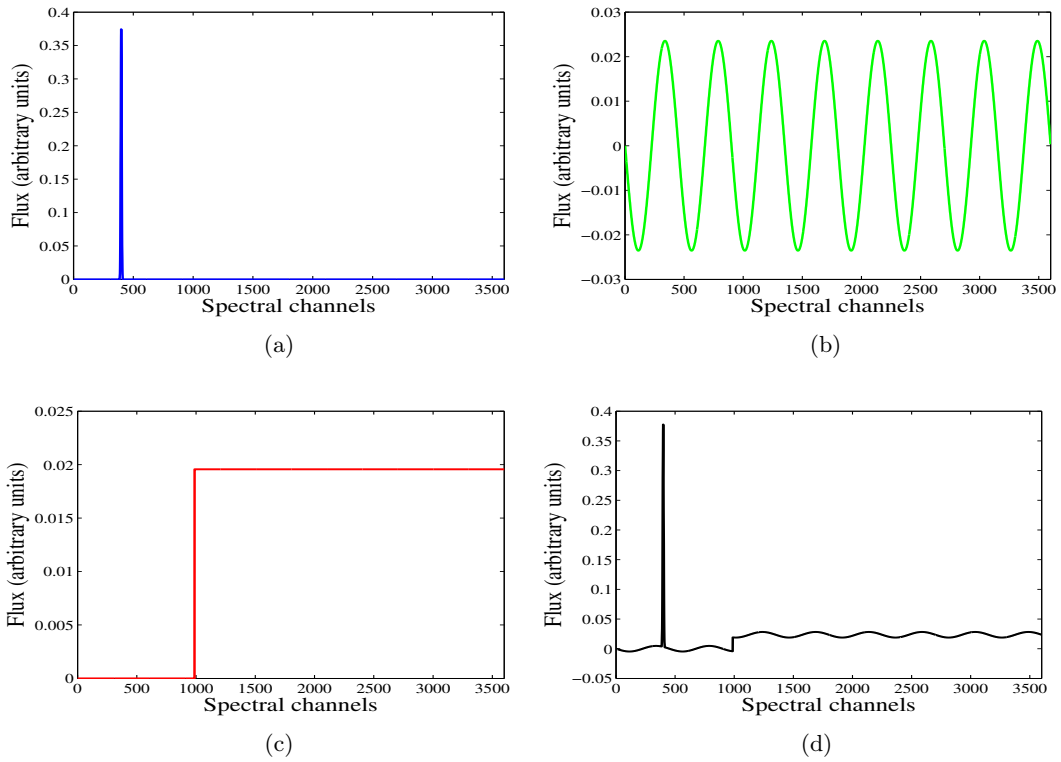



Figure 7.8: Examples of atoms in dictionary $\mathbf{R}$. 7.8(a) Example of spline atom modeling spectral emission lines. 7.8(b) Example of sine function modeling oscillatory behavior of continuous spectra. 7.8(c) Example of step function modeling spectral breaks. 7.8(d) Example of weighted linear combination of dictionary atoms obtained from the three atoms in Figures 7.8(a), 7.8(b) and 7.8(c), representing a real spectrum.

respectively plot an example of spline atom (in blue) modeling an emission spectral line, an example of sine function (in green) modeling a continuous spectrum, and an example of step function modeling a spectral break (in red). An illustration of weighted linear combination obtained from these three dictionary atoms and representing a real spectrum is plotted in black in 7.8(d).

### 7.3.2 Comparison of the detection performances on a single test spectrum

The unconstrained GLR (5.9) and the sparsity-constrained $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ (5.35) tests are compared on one single test spectrum, simulated by the MUSE consortium (Figure 7.9). For the purpose of MC simulations, results will be shown for spectra shorter than those that will be provided by the MUSE instrument ($N = 2048$ instead of $\approx 3600$). To quantify how the noise $\boldsymbol{\epsilon}$ affects a particular convolved spectrum $\boldsymbol{s}_{\mathbf{H}} = \mathbf{HR}\boldsymbol{\alpha}$ in (5.1), we define the SNR

as

$$\mathrm{SNR}(\boldsymbol{s}_{\mathbf{H}}) = 10\log_{10}\frac{||\boldsymbol{s}_{\mathbf{H}}||_2^2}{\mathrm{Tr}\{\boldsymbol{\Sigma}\}}. \tag{7.1}$$

To fix the ideas, a SNR of 0 dB corresponds to a noise that has the same power as the convolved spectrum, while a SNR of $-20$ dB corresponds to a noise power hundred times greater than the power of the convolved spectrum.

The first step concerns the empirical computation of the correspondence between the maximal probability of false alarm $\mathrm{P_{FA_0}}$ and the $\eta$ threshold values (5.23), necessary for the implementation of the sparsity-constrained $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ test. The result, obtained via MC simulation, is shown in Figure 7.9(a). For instance, we obtain that in order to fix $\mathrm{P_{FA}} = 0.01$ we have to choose $\eta = 4.72$. Figure 7.9(b) shows the considered noiseless spectrum (in blue) and the corresponding noisy data, on which we work (in cyan). This spectrum is characterized by a SNR of $-19.3$ dB, which is representative of MUSE spectra (as reported later in Figure 7.10(b)). Figure 7.9(c) compares the ROC curves of the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ (in blue) and the unconstrained GLR (in red) at a fixed SNR of $-19.3$ dB. Since $\gamma$ is fixed, different values of $\mathrm{P_{FA}}$ are obtained by varying $\eta$. Clearly, the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ test is superior to the unconstrained GLR. While the latter test is "blind" to the dictionary model and acts as a simple energy detector, the former test allows to detect more efficiently particular spectral features. For instance, for the considered spectrum, the emission line in Figure 7.9(b) is detected by a wide spline in the normalized redundant dictionary $\mathbf{D} = \boldsymbol{\Sigma}^{-1/2}\mathbf{H}\mathbf{R}\mathbf{N}_{\mathbf{D}_{\boldsymbol{\Sigma}\mathbf{H}}}^{-1}$ (cf. (5.3)), showing the relevance of the 1-sparse GLR model (5.26). Figure 7.9(d) finally compares the tests in terms of probability of detection $\mathrm{P_{DET}}$ versus the SNR, at a fixed of $\mathrm{P_{FA}} = 0.01$. With respect to the unconstrained GLR test, the proposed approach allows substantially better detection for a wide range of SNR ($-10$dB to $-20$dB for this spectrum).

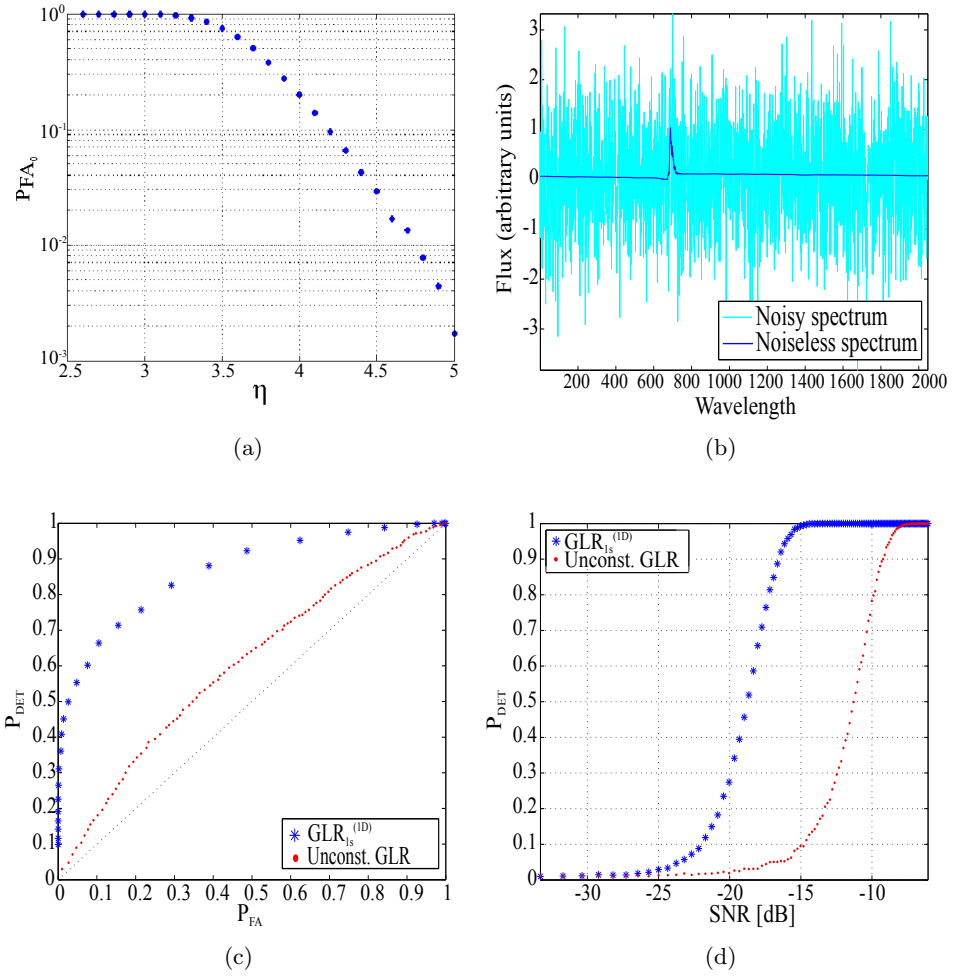



Figure 7.9: Comparison of the $\mathrm{GLR}_{\mathrm{1s}}^{(\mathrm{1D})}$ and GLR tests on a single MUSE spectrum. 7.9(a) Empirical $\mathrm{P}_{\mathrm{FA}_0}(\eta)$. 7.9(b) $\boldsymbol{s}_{\mathbf{H}}$ (blue) and corresponding data (cyan) at SNR $= -19.3$ dB. 7.9(c) ROC of $\mathrm{GLR}_{\mathrm{1s}}^{(\mathrm{1D})}$ vs unconstrained GLR at fixed SNR $= -19.3$ dB. 7.9(d) $\mathrm{P}_{\mathrm{DET}}$ vs SNR at fixed $\mathrm{P}_{\mathrm{FA}} = 0.01$.

### 7.3.3 Comparison of the detection performances on hyperspectral data cubes

After examining the tests on a single MUSE spectrum, we now consider their comparison on a first small MUSE-like hyperspectral data cube of size ($100 \times 100 \times 2048$), which is illustrated in Figure 7.10. Figure 7.10(a) shows the log version of the noisy considered cube (mean over all wavelengths). The SNR map of each spectrum is shown in Figure 7.10(b). This map can be considered as a reference showing where the signal is located under $\mathcal{H}_1$. Most spectra are buried in noise and have SNR below $-20$ dB. The $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ and the unconstrained GLR tests were run on each noisy spectrum of the data cube and because of the variance diversity from one spectrum to another ($\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(x, y)$), the testing procedures use one normalized dictionary $\mathbf{D}_{\boldsymbol{\Sigma}\mathbf{H}} = \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{H}\mathbf{R}$ per spectrum. An example of the noise variances involved in the diagonal of $\boldsymbol{\Sigma}(x, y)$ was shown in Figure 7.4.

Figure 7.10(c) shows in white the location of the spectra which have been detected using the unconstrained GLR test, while Figure 7.10(d) shows the equivalent for the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ test. The $\mathrm{P_{FA}}$ is set to 0.01 for both tests. In the black zones no detection was claimed. At the same $\mathrm{P_{FA}}$, the unconstrained GLR test claimed 448 detections, and the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ 915, which illustrates the better performances (power nearly doubled) reached by the proposed test.

To make a further example, we run the unconstrained GLR and the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ tests on the latest *"dry-run"* version of the simulated MUSE hyperspectral data cube, released by the MUSE consortium at the end of December 2011. The dry-run data cube is ($100 \times 100 \times 3600$) big and contains several objects of different size and type. The reconstructed white light image of the reference noiseless scene is shown in Figure 7.11, where the color indicates intensity. The objects of the scene are numbered from 1 to 17.

Figure 7.12 reports the detection results for the unconstrained GLR and the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ tests on the dry-run. The noisy scene and the SNR map are reported in Figure 7.12(a) and Figure 7.12(b). Similarly to the simulations above both tests were run at a $\mathrm{P_{FA}} = 0.002$, corresponding to a threshold $\eta = 5.2$ for the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ test. The $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ (Figure 7.12(d)) outperforms the unconstrained GLR test (Figure 7.12(c)) with 767 detections versus 450.

Note that despite the satisfactory detection performances shown by the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ on the MUSE hyperspectral data cubes of Figure 7.10(a) and Figure 7.11, the proposed method fails in detecting particularly faint objects. This is the case of objects number 6, 7 and 13 in the dry-run scene of Figure 7.11. An approach to the problem of detecting such extremely faint signals (SNR $\approx -30$ dB) is given by the $\mathrm{GLR}_{1\mathrm{s}}^{(3\mathrm{D})}$ spatio-spectral detection test of (6.48), and will be discussed in section 8.3.

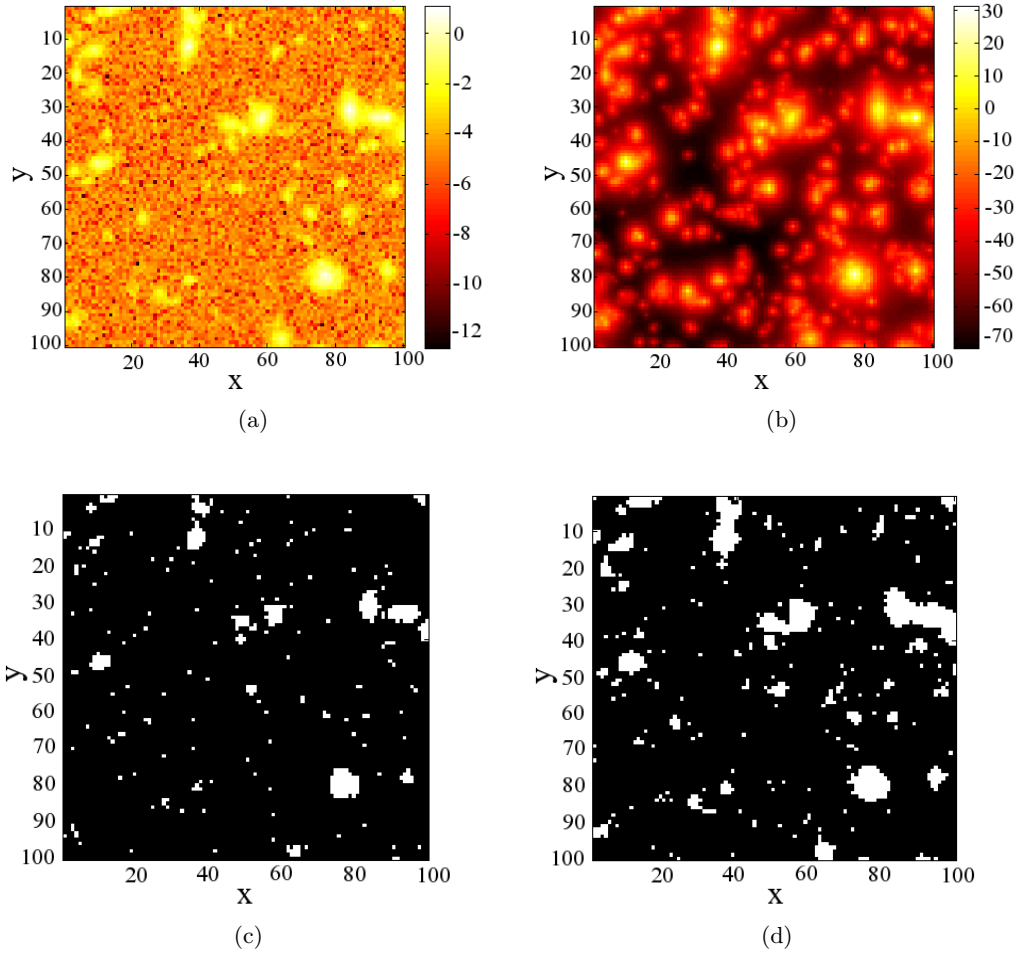



Figure 7.10: Comparison of the $\mathrm{GLR}_{\mathrm{1s}}^{(\mathrm{1D})}$ and unconstrained GLR tests on data provided by MUSE consortium. The $(x, y)$ axes represent the spatial dimensions as seen by MUSE. 7.10(a) Noisy MUSE subcube : Absolute value of the mean over wavelengths (logscale). 7.10(b) SNR of each spectrum. 7.10(c) Detection using the unconstrained GLR. 7.10(d) Detection using the $\mathrm{GLR}_{\mathrm{1s}}^{(\mathrm{1D})}$.

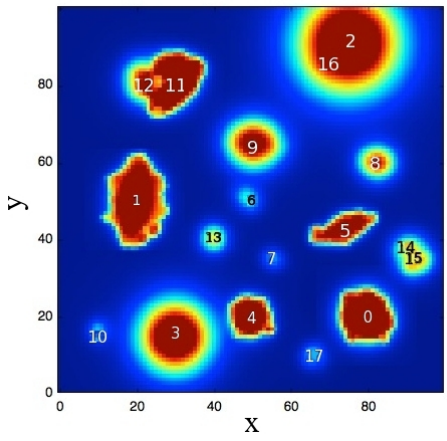



Figure 7.11: Reconstructed white light image of the noiseless dry-run data cube.

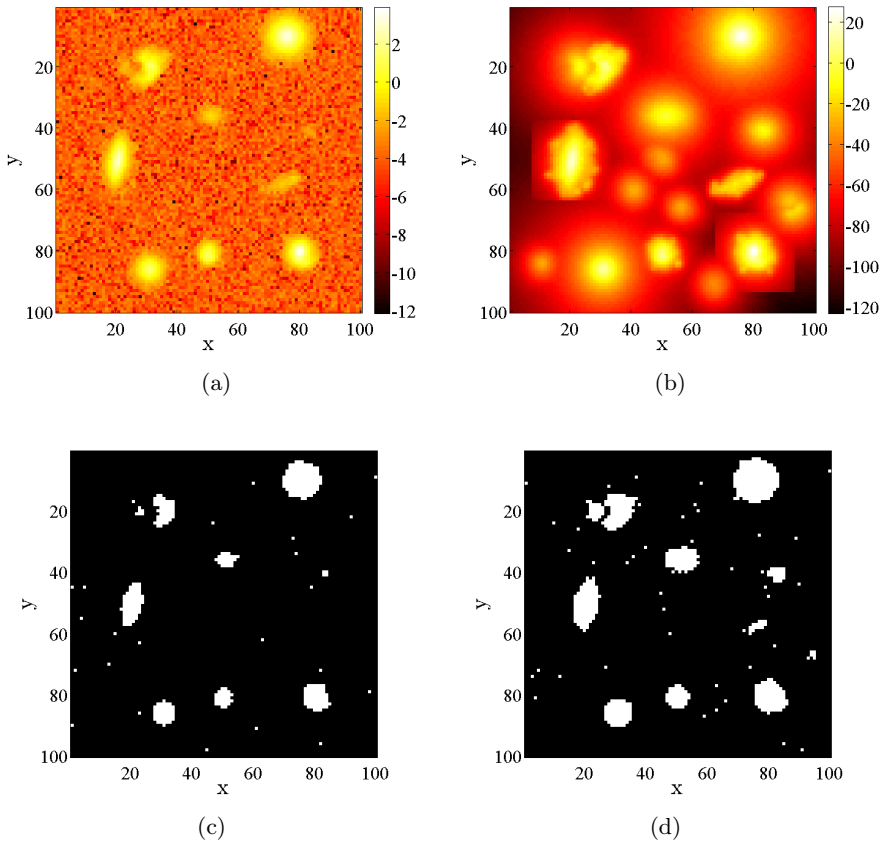



Figure 7.12: Compared detection performances of the $GLR_{1s}^{(1D)}$ and GLR tests on the *dry-run*, for a same $P_{FA} = 0.002$. 7.12(a) Log scaled version of the considered noisy MUSE sub-cube. 7.12(b) Dry-run's SNR map. 7.12(c) Detection results using the unconstrained GLR test (450 detections). 7.12(d) Detection results using the $GLR_{1s}^{(1D)}$ test (767 detections).

### 7.3.4 Analysis of the $\text{GLR}_{\text{1s}}^{(\text{1D})}$ test detection results in terms of $p$-value

To assess post-data reliability of the detection results obtained with the $\text{GLR}_{\text{1s}}^{(\text{1D})}$ test (5.35) in this and in the previous simulation, an analysis in terms of $p$-values (cf. subsection 1.2.4) is proposed. For reasons of clarity, results are illustrated considering the top-left dry-run's data sub-cube of dimensions $(65 \times 65 \times 3600)$. Six objects are present in this data cube (numbers 1, 6, 9, 11, 12 and 13 of Figure 7.11)

Defined as

$$p = Pr(\max_i(|\mathbf{D}_i{}^t\mathbf{x}| \geq t_{obs} \mid \mathcal{H}_0), \tag{7.2}$$

the $p$-value corresponds to the probability of obtaining, under the null hypothesis, a test statistic at least as extreme as the one that was actually observed. In this case, $t_{obs} = \max_i |\mathbf{D}_i{}^t\mathbf{x}_{obs}|$, where $\mathbf{x}_{obs}$ is a considered spectrum in the cube.

The results of such analysis are reported in Figure 7.13. Figure 7.13(a) shows the $\text{GLR}_{\text{1s}}^{(\text{1D})}$ detection map at $\text{P}_{\text{FA}} = 0.002$ for the considered data sub-cube (same detection map as Figure 7.12(d), top-left). Note that, among the six objects present in the scene, only the number 1, 9, 12 and 13 (the brightest ones) are detected, while objects number 6 and 13 are not (this question will be undertaken in section 8.3).

According to (7.2), we first compute the $p$-value $p$ for each spectrum in the considered hyperspectral data sub-cube. The result of such computation is graphically shown in Figure 7.13(b), where at each colored pixel $(x, y)$ in the image corresponds a specific $p$-value rate.

Now, to test the reliability of the detection results obtained with the $\text{GLR}_{\text{1s}}^{(\text{1D})}$ test (detection map of Figure 7.13(a)), we threshold the $p$-value map of Figure 7.13(b) to different decreasing values of the significance level $\alpha \in [5 \cdot 10^{-3}, 2 \cdot 10^{-3}, 10^{-3}, 5 \cdot 10^{-4}]$. Figures 7.13(c) − 7.13(f) illustrate the thresholded maps we obtained. In particular, with reference to colored pixels: Figure 7.13(c) reports $p$-value rates that are below $\alpha = 5 \cdot 10^{-3}$; Figure 7.13(d) shows $p$-value rates below $\alpha = 10^{-3}$; Figure 7.13(e) reports $p$-value rates below $\alpha = 5 \cdot 10^{-3}$; Figure 7.13(f) reports $p$-value rates below $\alpha = 5 \cdot 10^{-4}$.

Interestingly, despite few false alarms, pixels detected with the $\text{GLR}_{\text{1s}}^{(\text{1D})}$ test (detection map in Figure 7.13(a)) present very low values of $p$. This shows that when detection was claimed, the case was actually strongly made against $\mathcal{H}_0$ for objects $1, 9, 11$ and 12. Hence, this shows that the dictionary appropriately captures signal features, while obviously the unconstrained GLR does not.

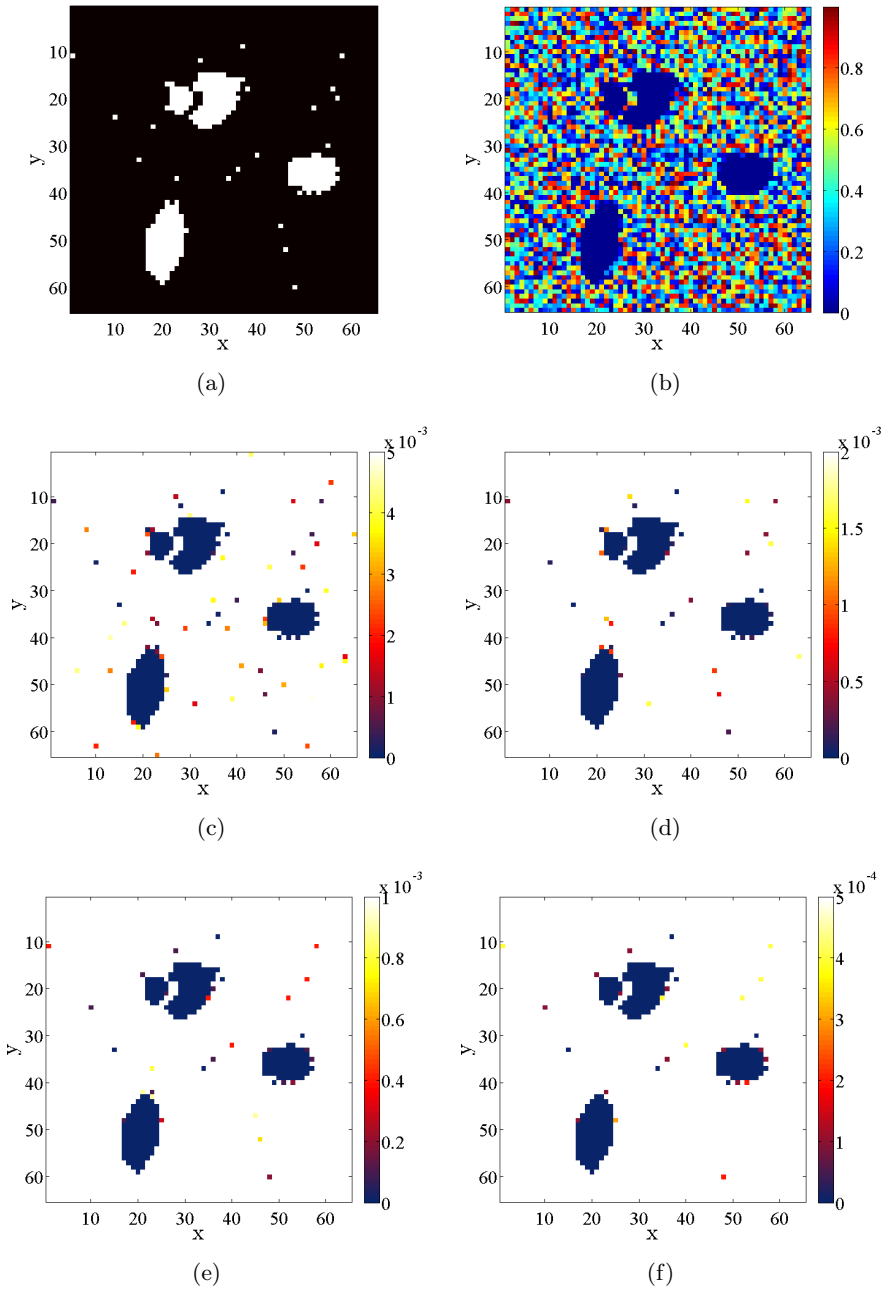



Figure 7.13: Comparison of the $p$-value maps computed for all spectra in a MUSE subcube at different values of the significance level $\alpha$. 7.13(a) Reference: detection map using $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ test with $\mathrm{P_{FA}} = 0.002$. 7.13(b) $p$-value map computed for all the spectra in the subcube. 7.13(c) $p$-value map thresholded at significance level $\alpha = 5{\cdot}10^{-3}$. 7.13(d) $p$-value map thresholded at significance level $\alpha = 2{\cdot}10^{-3}$. 7.13(e) $p$-value map thresholded at significance level $\alpha = 10^{-3}$.7.13(f) $p$-value map thresholded at significance level $\alpha = 5 \cdot 10^{-4}$.

## 7.4 Conclusion

We have presented a brief sketch of historical developments in Hyperspectral Imaging and some of its applications to Astrophysics, from the beginnings of spectroscopy to one of the most recent technological innovations: the ESO's MUSE instrument. Conceived to be installed at the Very Large Telescope in Chili in 2013, the MUSE instrument is an integral field spectrograph that will provide a huge quantity of hyperspectral data.

In the context of MUSE astrophysical hyperspectral data, comparative analysis of the detection performances of the unconstrained GLR and the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ spectral model-based detection tests (cf. chapter 5) has been proposed. In accordance to the spectral characteristics of the considered astrophysical data, we firstly discussed a suitable and specific design for the redundant dictionary $\mathbf{R}$ of model (5.2), on which the tests are based.

Numerical detection results of the GLR and the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ tests have then been shown on a single MUSE spectrum, and also on MUSE hyperspectral data cubes. In both circumstances, the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ approach has shown higher detection power with respect to the unconstrained GLR test, for same fixed values of the $\mathrm{P}_{\mathrm{FA}}$.

To highlight the reliability of the detection results obtained with the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ test an analysis of the results in terms of $p$-value has been proposed.

Despite the superior power of the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ test with respect to the unconstrained GLR on MUSE data sets, a few extremely faint sources in the hyperspectral cube remained undetected with the proposed approach. Improved detection results using the spatio-spectral model based strategies of chapter 6 are reported in the next chapter.

Chapter 8

# Detection results using spatio-spectral model based strategies

**Contents**


**8.1 Introduction . . . . . 133**
**8.2 Multiple-round detection strategy applied to MUSE data . . . . . 133**
**8.3 Detection results of the spatio-spectral model based approach on MUSE data . . . . . 139**
8.3.1 3D (spatio-spectral) dictionary . . . . . 139
8.3.2 Spatial $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ vs spatio-spectral $\mathrm{GLR}_{1\mathrm{s}}^{(3\mathrm{D})}$: Detection results on three faint Ly-$\alpha$ emitters . . . . . 141
**8.4 Conclusion . . . . . 146**


## 8.1 Introduction

We report in this chapter the results obtained applying the spatio-spectral model based detection strategies developed in chapter 6 to MUSE hyperspectral data.

In the first section of this chapter, we focus on the multi-round detection strategy of section 6.2 that accounts for the cascade of the sparsity-constrained GLR with a spectral model, the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ test of (5.35), and the test LR-MP$\beta$ (6.31), which accounts for spatial dependency in a basic way. According to the analytical study held in subsection 6.2.3 on the $\mathrm{P_{FA}}$ and $\mathrm{P_{DET}}$ of this multi-round approach, results in terms of the overall FA rate computed on each pixel vector of the considered hyperspectral cubes are shown.

The second section of this chapter presents the detection results obtained performing the sparsity-constrained GLR with a spatio-spectral model, the $\mathrm{GLR}_{1\mathrm{s}}^{(3\mathrm{D})}$ test of (6.48), introduced in section 6.3. This strategy allies dictionary learning techniques, sparse models and spread functions to concentrate information prior to detection. In particular, we focus on the detection of very faint and highly sparse signals in MUSE data cubes. As mentioned in subsection 7.2.3 these spectra, which are representative of very distant galaxies, are characterized by single, very faint and narrow lines, known as Ly$-\alpha$ lines, and correspond to objects that were not detected by previous approaches (cf. subsection 7.3.3).

## 8.2 Multiple-round detection strategy applied to MUSE data

We present here the detection results obtained when performing the multiple-round detection strategy introduced in subsection 6.2.3 on MUSE hyperspectral data. We first

illustrate the principle of the approach in two steps on a MUSE sub-cube (Figure 8.1). Then, we will turn to the multiple-step detection strategy.

### Results of the two-step approach

Numerical results are illustrated on a part ($14 \times 10$ spectra of $N = 2048$ components each) of a simulated MUSE data cube shown in Figure 8.1. On this data sub-cube, the following *two-step* approach is implemented: after a first detection round with the $\text{GLR}_{1\text{s}}^{(1\text{D})}$ detection test (5.35) for a target $\text{P}_{\text{FA}}$, a second detection-round follows performing the LR-MP$\beta$ test (6.31) on all those pixel-vectors not detected by the $\text{GLR}_{1\text{s}}^{(1\text{D})}$ but contiguous to bright detected one (see subsection 6.2.3 and Figure 8.1(d)).

As a reference, we show in Figure 8.1(a) the SNR map of the considered scene: spectra with large, energetic features are shown in yellow, and spectra with fainter features in darker red. Figure 8.1(b) and 8.1(c) show the detection results obtained running the unconstrained GLR (5.9) and the $\text{GLR}_{1\text{s}}^{(1\text{D})}$ tests respectively, at a $\text{P}_{\text{FA}} = 0.01$. As already seen, at a same $\text{P}_{\text{FA}}$ the $\text{GLR}_{1\text{s}}^{(1\text{D})}$ outperforms the unconstrained GLR test (22 detections for the $\text{GLR}_{1\text{s}}^{(1\text{D})}$ versus 10 detections for the unconstrained GLR). In Figure 8.1(d), the spectra where detection was claimed using the $\text{GLR}_{1\text{s}}^{(1\text{D})}$ for a target $\text{P}_{\text{FA}_0} = 0.01$ (the $\mathbf{\Gamma}$ set, discussed in subsection 6.2.3) are highlighted in green (these are also the white pixels of Figure 8.1(c)). In the same image, the blue and brown spectra correspond to the set $\bar{\mathbf{\Gamma}}$: in brown are shown the spectra where nothing was detected, but for which a contiguous neighbor was. We take those spectra to perform a *second round* of detection with the LR-MP$\beta$ test for a $\text{P}_{\text{FA}} = 0.01$. This means that for each (faint) pixel in the brown contour:
i) we identify the brightest neighbor spectrum;
ii) we use the bright spectrum to estimate the salient spectral content of the faint one;
iii) based on this faint spectrum estimate we perform the LR-MP$\beta$ test of (6.31).

Figure 8.1(e) reports the detection results of such a two-step approach: the spectra for which a detection was found are shown in red (a total of 35 detections were claimed with the cascade of the $\text{GLR}_{1\text{s}}^{(1\text{D})}$ and LR-MP$\beta$ tests). A closer analysis of the spectra flagged in red and of their estimates shows that most of the time, the features detected in the faint spectra are real. An example was shown in Figure 6.3, for the green and blue crossed spectra of Figure 8.1(a).

Figure 8.1(f) shows the resulting FA rate, after performing the cascade of the two tests. The FA rates for spectra tested only once with the $\text{GLR}_{1\text{s}}^{(1\text{D})}$ test are reported in gradations of blue while spectra tested twice are easily recognizable in red. The FA variation from one spectrum to another is due to the fact that, when performing the first detection round with the $\text{GLR}_{1\text{s}}^{(1\text{D})}$ test, according to model (6.1) one normalized dictionary per spectrum is computed (because $\mathbf{\Sigma}$ varies from one spectrum to another). For simplicity, instead of computing the value of the $\eta$ threshold that would give $\text{P}_{\text{FA}} = 0.01$ for each spectrum in the data cube (which would involve the computation of one $\text{P}_{\text{FA}}$ vs. $\eta$ curve (cf. Figure 7.9(a)) for each spectrum), we fix a value of $\eta$ corresponding to a mean $\text{P}_{\text{FA}}$ of 0.01 for all spectra. For this value of $\eta$, corresponding to a mean $\text{P}_{\text{FA}}$ of 0.01, we obtain FA rates which are slightly different, depending on the $\mathbf{\Sigma}$ of the spectrum we are testing. The $\text{P}_{\text{FA}}$ map of Figure 8.1(f) reports a minimum FA value of $\text{P}_{\text{FA}min} = 0.009$ and a maximum value equal to $\text{P}_{\text{FA}Max} = 0.023$. Hence, the overall $\text{P}_{\text{FA}}$ of the two-step strategy follows in

the range $0.009 \leq {P_{FA}}^{2tests} \leq 0.023$. In particular, the mean ${P_{FA}}^{2tests}$ over all spectra in the data cube corresponds to 0.014, which is similar to the FA rate set on each detection round, independently. 35 detections (green and red pixels in Figure 8.1(e)) were claimed performing the cascade of the $GLR_{1s}^{(1D)}$ and LR-MP$\beta$ tests at the mean ${P_{FA}}^{2tests} = 0.014$. These detection performances would not have been reached performing the sole $GLR_{1s}^{(1D)}$ test at $P_{FA} \approx 0.014$ (a total of 33 detections is obtained with the $GLR_{1s}^{(1D)}$ test for $\eta = 4.2$, corresponding to an only $P_{FA} = 0.12$ that is, ten times higher), the two-step strategy shows a higher overall detection power.

**Results with the multiple-step procedure on a whole data-cube**

Results on the whole dry-run data cube of Figure 7.11 are illustrated in Figure 8.2 when applying the multiple-step procedure of subsection 6.2.3. This time, after a first detection round using the sparsity constrained $GLR_{1s}^{(1D)}$ test at $P_{FA} \approx 0.0018$ (corresponding to a value of the threshold $\eta = 5.2$, for this data cube), successive detection rounds are performed on faint contour pixels with the LR-MP$\beta$ test at $P_{FA} = 0.0018$ until no new detection is found. As a reference, Figure 8.2(a) reports the SNR map of the considered data cube (same as Figure 7.12(b)). Figure 8.2(b) shows in white the pixels where detection is claimed after performing the first detection round with the $GLR_{1s}^{(1D)}$ test, for a total of 767 detections on the cube. These pixels are those in green in Figure 8.2(c). Similarly to the results shown in Figure 8.1(d) and 8.1(e), a second detection round is performed with the LR-MP$\beta$ test at $P_{FA} = 0.0018$ on all the faint pixels (not detected with the $GLR_{1s}^{(1D)}$ test but contiguous to detected ones) in the brown contour in Figure 8.2(c). At this point, the new detected pixels (flagged in red) are aggregated to the set $\mathbf{\Gamma}$ of already detected ones (in green, in Figure 8.2(d)) and another detection round with the LR-MP$\beta$ on the pixels of the new brown contour is performed. This procedure iterates until no more faint pixels in brown contours are detected. Figure 8.2(e) shows in red the totality of pixels detected after three successive rounds with the LR-MP$\beta$ test (since no new pixels are detected at the end of the third round with the LR-MP$\beta$ test, the algorithm stops). A total of 959 detections is obtained. The corresponding $P_{FA}$ map is illustrated in Figure 8.2(f). The minimum FA value is of $P_{FA\,min} = 3 \cdot 10^{-4}$ and a maximum value equals $P_{FA\,Max} = 0.008$. The mean $P_{FA}$ over all spectra in the data cube corresponds to 0.0018, which is similar to the FA rate set on each detection round, independently. At this indicative FA rate, the multi-step strategy shows better detection power (959 detections) with respect to only one detection round with the $GLR_{1s}^{(1D)}$ (767 detections). Detection performances similar to that of the cascade of the two tests would be obtained with the sole $GLR_{1s}^{(1D)}$ test by setting a much higher FA rate: when setting $\eta = 4.8$, corresponding to $P_{FA} \approx 10^{-2}$, 947 detections are claimed with this test.

From these results we see that with the multi-round detection strategy introduced in subsection 6.2.3, improved detection performances can be achieved accounting for simple spatial dependency essentially: unknown proportionality in (6.24), while maintaining the $P_{FA}$ rate reasonably low. Nevertheless, performing such multiple-round detection approach is complicated, especially with regard to the analytical computation of the overall $P_{FA}$ and $P_{DET}$ (cf. subsection 6.2.3) and the variation of the $P_{FA}$ from one spectrum to another. Moreover, since it is based on the exploitation of brighter spectra's estimated spectral con-

tent for the detection of faint contiguous ones, this approach does not allow any improvement on the detection of spectrally significant but very faint pixel-vectors that were not detected at the first detection round with the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ test and for which none of the eight neighbors was. This is the case for example, of objects number $6, 7, 13$ in the dry-run (cf. Figure 7.11). Improved detection results on such objects are presented in the next section, by applying to MUSE data the spatio-spectral model based technique developed in section 6.3.

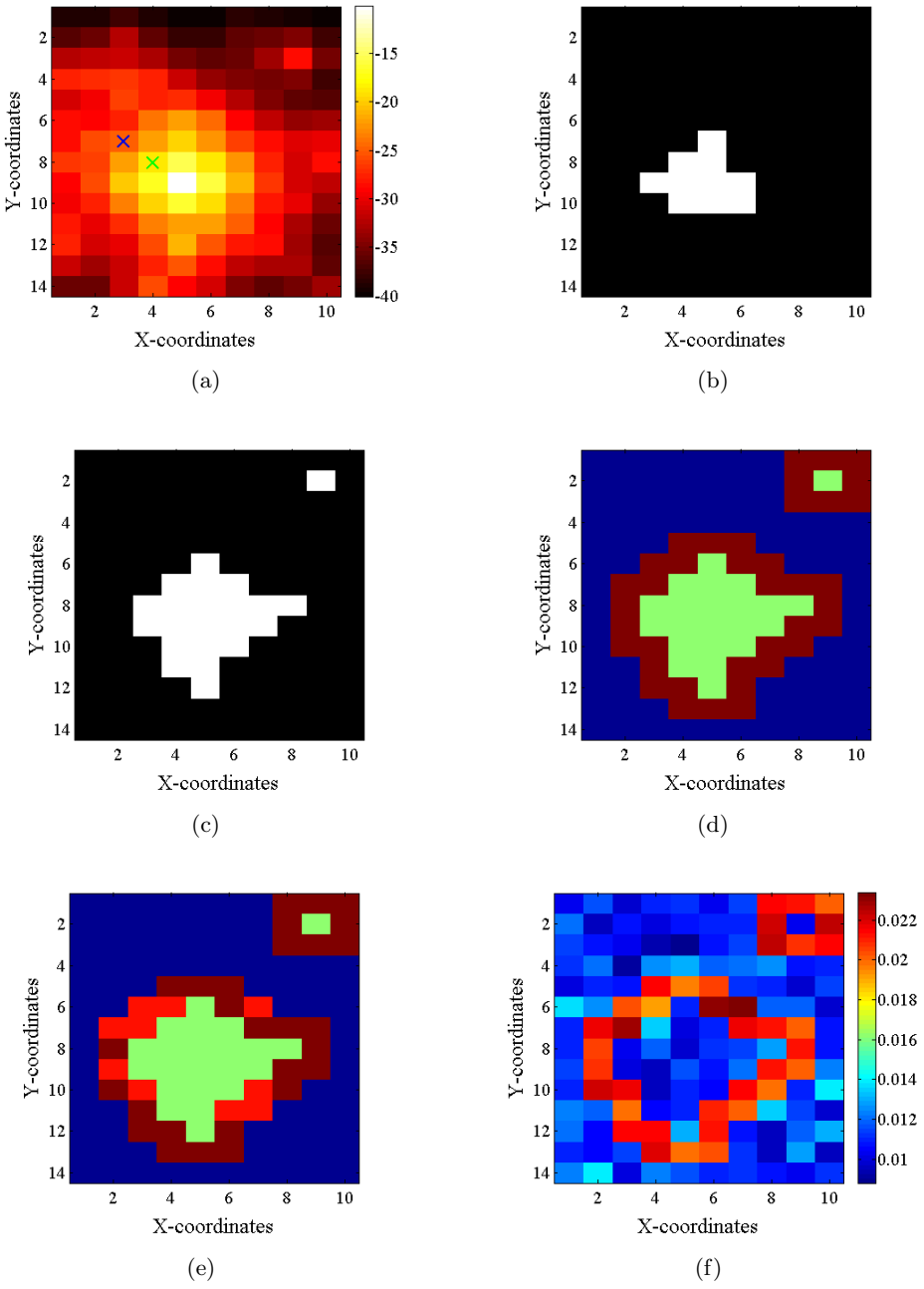



Figure 8.1: Detection results using a two-step detection strategy on a MUSE sub-cube. 8.1(a) Reference scene: SNR map in dB. 8.1(b) Result performing a single detection round with the unconstrained GLR test at $P_{FA} = 0.01$. 8.1(c) Result performing a single detection round with the $GLR_{1s}^{(1D)}$ test at $P_{FA} = 0.01$. 8.1(d) Selection of the image contour to perform a two-step approach: spectra where $GLR_{1s}^{(1D)}$ claimed detection at significance level $P_{FA_0} = 0.01$ are shown in green, while neighbor spectra on which the LR-MP$\beta$ test is performed are in brown. 8.1(e) Detection result with the LR-MP$\beta$ test at $P_{FA} = 0.01$ (detected pixels in red). 8.1(f) Corresponding $P_{FA}$ map of the two-step approach. The FA rates for spectra tested only once with the $GLR_{1s}^{(1D)}$ test appear in gradations of blue while spectra tested twice are easily recognizable in red. The mean $P_{FA}$ over all spectra is $P_{FA} = 0.014$.

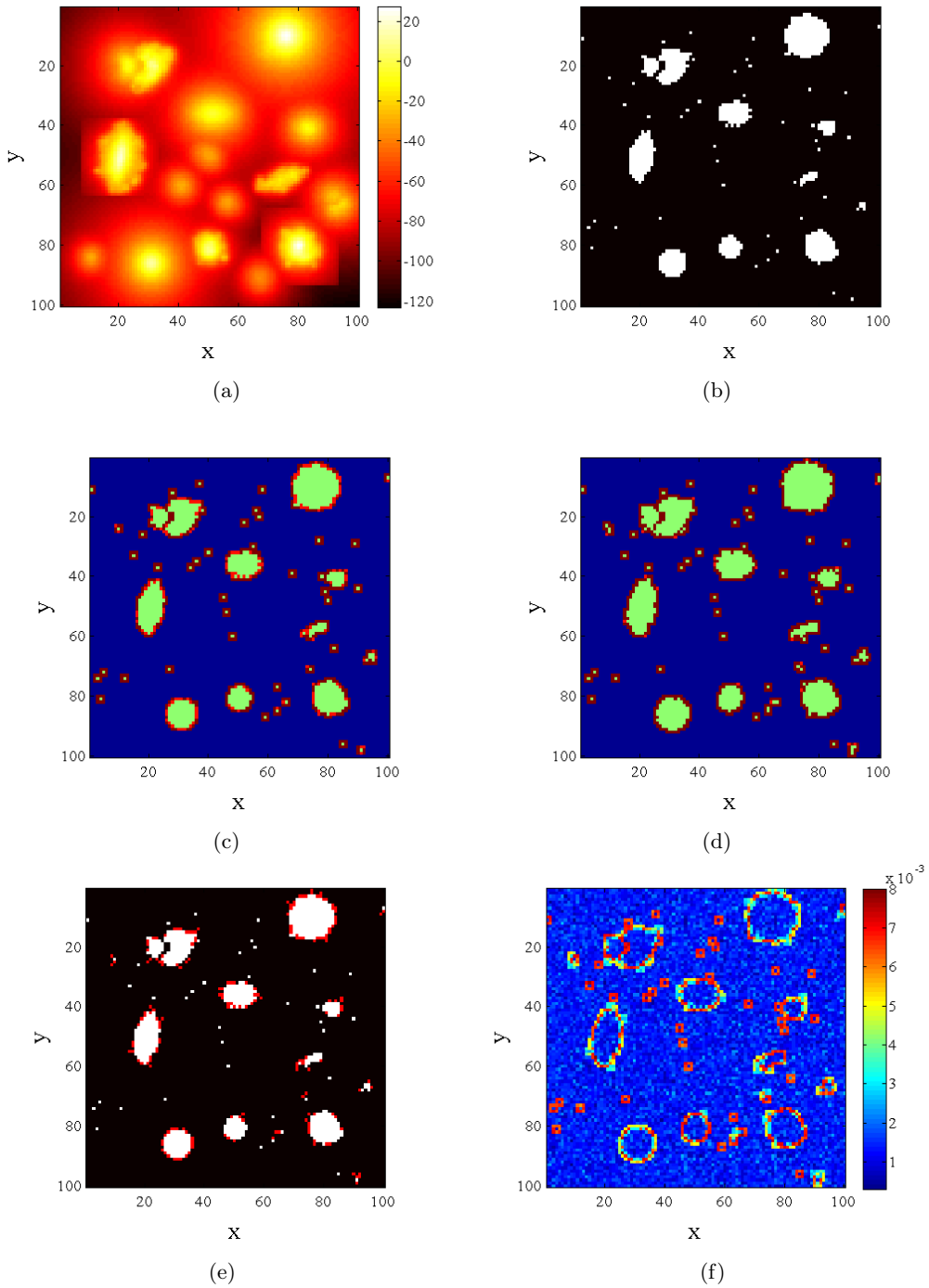



Figure 8.2: Detection results using a multiple-step detection strategy on MUSE dry-run. 8.2(a) Reference scene: SNR map in dB (same as Figure 7.12(b)). 8.2(b) Result performing a single detection round with the sparsity-constrained $\mathrm{GLR}_{1s}^{(1D)}$ test at $\mathrm{P_{FA}} = 0.0018$: 767 detections were claimed (white pixels). 8.2(c) Results cascading two detection rounds: a first round with the $\mathrm{GLR}_{1s}^{(1D)}$ at $\mathrm{P_{FA}} = 0.0018$ (detected pixels in green) and a second round with the LR-MP$\beta$ test at $\mathrm{P_{FA}} = 0.0018$ on brown contour pixels (detected pixels in red). 8.2(d) Results cascading three detection rounds: pixels in red in Figure 8.2(c) are aggregated to the set $\mathbf{\Gamma}$ of detected pixels (in green) and another round with the LR-MP$\beta$ test at $\mathrm{P_{FA}} = 0.0018$ is performed on new brown contour pixels (detected pixels in red). 8.2(e) In white: pixels detected after a single detection round with the sparsity-constrained $\mathrm{GLR}_{1s}^{(1D)}$ test (detection map of Figure 8.2(b)); in red: set of faint contour pixels detected after two successive detection rounds performing the LR-MP$\beta$ test. A total number of 959 detections were claimed. 8.2(f) Corresponding $\mathrm{P_{FA}}$ map after three detection rounds on the dry-run cube. The mean $\mathrm{P_{FA}}$ over all spectra is $\mathrm{P_{FA}} = 0.0018$.

## 8.3 Detection results of the spatio-spectral model based approach on MUSE data

As an illustration of the sparsity-constrained GLR using a spatio-spectral model, the $\mathrm{GLR}_{1\mathrm{s}}^{(3\mathrm{D})}$ detection method (6.48) proposed in section 6.3, we show here that we can improve on the results of section 7.3 by using spatio-spectral dictionaries. These dictionaries exploit both the sparsity of the data along the $\lambda$ dimension and the information leakage over several voxels, caused by the instrument 3D PSF.

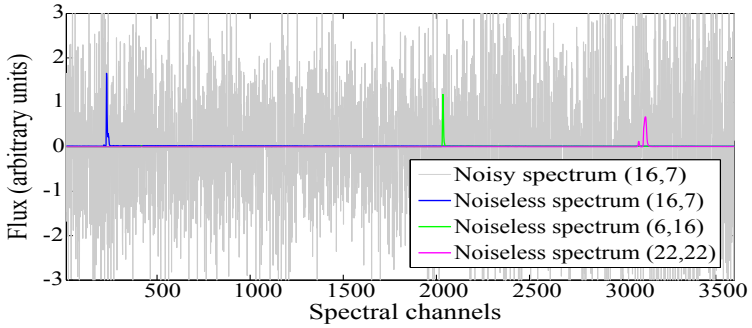



Figure 8.3: Typical spectra under $\mathcal{H}_1$ and corresponding data for $(16, 7)$. Coordinates $(x, y)$ correspond to spatial position in Figure 8.8(a).

One of the major challenges of the MUSE instrument concerns the detection of very distant galaxies. Those galaxies present spectra composed by a sole, faint and narrow line, known as Lyman-alpha line (Ly-$\alpha$). Figure 8.3 reports examples of Ly-$\alpha$ spectra (blue, magenta and green) and the corresponding data (grey) for one of them (see also Figure 8.8). These spectra actually correspond to sources number $6, 7$ and $13$ in MUSE dry-run cube of Figure 7.11. The highly noisy character of these data is striking: the lines are totally buried in noise and the SNR of the spectra is

$$\mathrm{SNR} = 10 \log_{10} \frac{\|\mathbf{FHR}\alpha\|_2^2}{\mathrm{Tr}\{\mathbf{\Sigma}\}} \approx -30\mathrm{dB}. \tag{8.1}$$

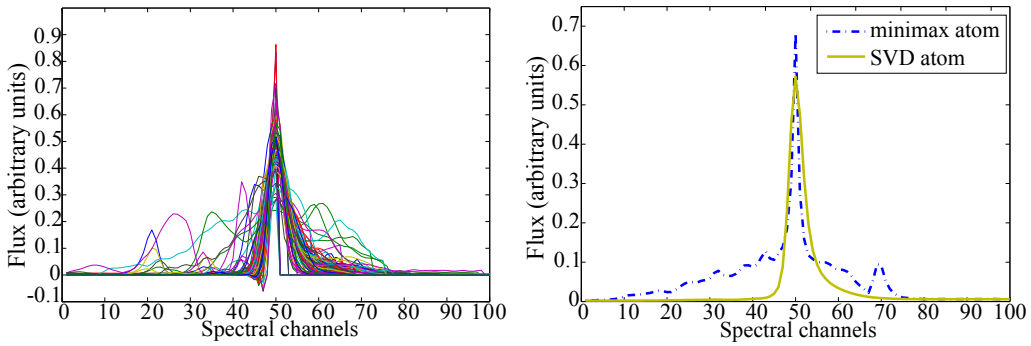
### 8.3.1 3D (spatio-spectral) dictionary

For such sources, our approach tries to limit the difficulties caused by the composite nature of the model. Although the spectral pattern to be detected is not known, we made the request to the MUSE consortium to dispose of a set of training signals that represents well the variety of all possible features to be detected. The possible shapes of the signals under $\mathcal{H}_1$ can, in this case, be accurately simulated using astrophysical models: we dispose of a library of $\mathcal{L} = 10^4$ line profiles[10] (100 of which are shown on Figure 8.4(a)), which extend on about $N_\lambda = 100$ spectral channels. In real data, the location in spectral channel of the lines is however arbitrary and unknown (it can be anywhere in the $\approx 3600$ spectral channels), because it depends on the distance of the source (Doppler effect[11]). Performing

[10] provided by Roland Bacon, Principal Investigator of MUSE consortium.

[11] The Doppler effect (from Christian Doppler, who discovered it in 1842) is the change in frequency of a wave (or other periodic event) for an observer moving relative to its source [Doppler 1842].

the tests (5.35) and (6.48) over $X \times Y = (300 \times 300)$ spectra in a MUSE data cube, testing for each one of these pixel-vectors the full library of $\mathcal{L} = 10^4$ spectral signatures (extended over $N_\lambda = 100$ spectral channels) at each spectral band in the $\Lambda = 3600$ dimension, entails performing $\mathcal{L} \times \Lambda \times X \times Y \times N_\lambda \approx 3 \times 10^{14}$ multiplications and additions. If we assume $3 \cdot 10^9$ operations/sec, this represents 6 months of computation time for the 3D test.




(a) 100 spectral lines in $\mathcal{L}$ (b) SVD and minimax atoms

Figure 8.4: Examples of spectral lines and trained atoms.

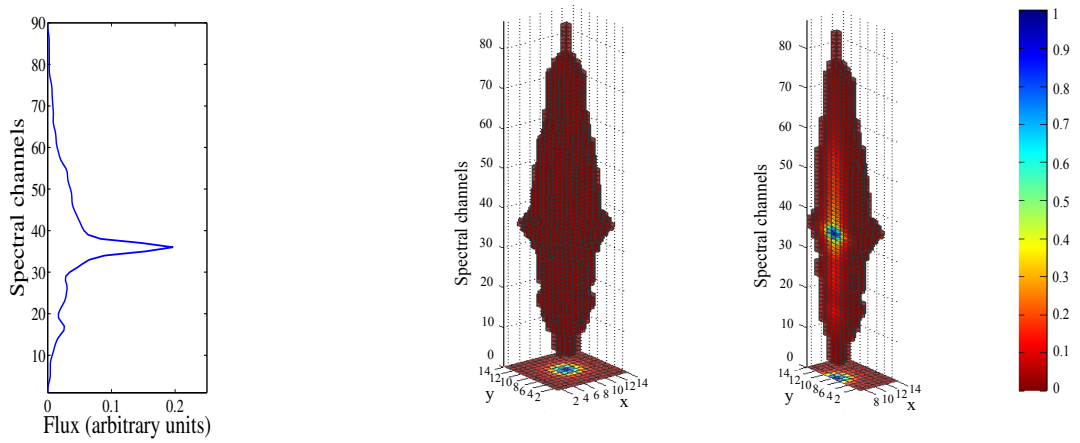
We thus build specific dictionaries $\mathbf{R}$, different from the one used up to now, designed according to the K-SVD (cf. Appendix B.3) and minimax (cf. subsection 2.2.2 and [Suleiman 2013]) approaches. In particular, we optimize three dictionaries (SVD, K-SVD with 7 atoms and minimax) of reduced dimensions (resp. $\Lambda \times 1$, $\Lambda \times 7$ and $\Lambda \times 1$) on the whole set of lines centered at a wavelength $\Lambda_{50}$ (Figure 8.4(a)). The resulting SVD and minimax atoms are shown in Figure 8.4(b). Note that while SVD and K-SVD learning techniques provide atoms that reflect average (most probable) spectral characteristics in the whole library of line profiles, the minimax approach accounts for (by maximizing the worst $\mathrm{P_{DET}}$) the most marginal spectral form, which could be helpful in detecting sources with spectral features dissimilar from the others.

**- Construction of the 1D spectral dictionaries by translation:**

Because of the unknown location of the line sought in spectral channel, we generated translation invariant versions of the SVD, K-SVD and minimax spectral profiles by shifting the corresponding atoms at all possible wavelengths. This yielded the 1D $\mathbf{R}_{\mathrm{SVD}}$, $\mathbf{R}_{\mathrm{KSVD}_7}$ and $\mathbf{R}_{\mathrm{minimax}}$ spectral dictionaries used to perform the unconstrained GLR (1.55) and the 1D (spectral) $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ (5.35) test (see Figures 8.8(b) and 8.8(c)).

**- Construction of the 3D spatio-spectral dictionaries:**

The SVD, K-SVD and minimax dictionaries of 3D atoms are obtained as follows:
i) we compute the optimized SVD, K-SVD and minimax spectral profiles;
ii) each such optimized dictionary profile is composed with the instrument WSF, which spreads the atom's content over 7 spectral elements;
iii) the atom obtained at step ii) is composed with the instrument SSF, which blurs the atom's content at each spectral band over $13 \times 13$ pixels in the spatial domain;
iv) translation invariant dictionary of this 3D atom is obtained by centering it at each posi-




(a) (b)

Figure 8.5: Example of 3D minimax atom: (a) minimax spectral profile (b) Left: corresponding 3D atom after composition by the SSF and the WSF. Right: a cut of the 3D atom. This atom is whitened and then correlated at each position in $(x, y, \lambda)$ in the data cube.

tion $(x, y, \lambda)$ in the data cube and by whitening it with the corresponding noise variances. This leads to specific 3D atoms, which account for the WFS and SSF effects and for target spectral shapes.

These atoms are used to perform the sparsity-constrained spatio-spectral model-based $\mathrm{GLR}_{1\mathrm{s}}^{(3\mathrm{D})}$ test (6.48) (see Figure 8.8).

Examples of atoms obtained in the 3D case are shown in Figure 8.5 and Figure 8.7. In Figure 8.5, Figure 8.5(a) illustrates the minimax spectral profile (the same as in Figure 8.4(b)) while Figure 8.5(b) shows the corresponding 3D atom (on the left) and a vertical cut (on the right). Similarly, Figure 8.7(a) reports the plots of three different spectral profiles and Figure 8.7(b) the corresponding 3D structures. The top raw figures shows the spectral profile obtained by SVD approach and the corresponding 3D atom, while the center and the bottom raws figures show two spectral profiles obtained by KSVD and the corresponding 3D structures. These examples of 3D atoms illustrate how the energy of Ly$-\alpha$ emitters can typically be distributed in MUSE data cube because of the blurring effects of the instrument WSF and SSF. For the 3D atoms in Figure 8.7 for example, we see that while first two (top and center raw) look very similar except for a slight difference in the tail amplitude at lower spectral channels, the third 3D atom has a different shape characterized by two distinct energy blobs.

### 8.3.2 Spatial $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ vs spatio-spectral $\mathrm{GLR}_{1\mathrm{s}}^{(3\mathrm{D})}$: Detection results on three faint Ly-$\alpha$ emitters

We compare the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ (5.35) and the $\mathrm{GLR}_{1\mathrm{s}}^{(3\mathrm{D})}$ (6.48) detection tests on a small sub-

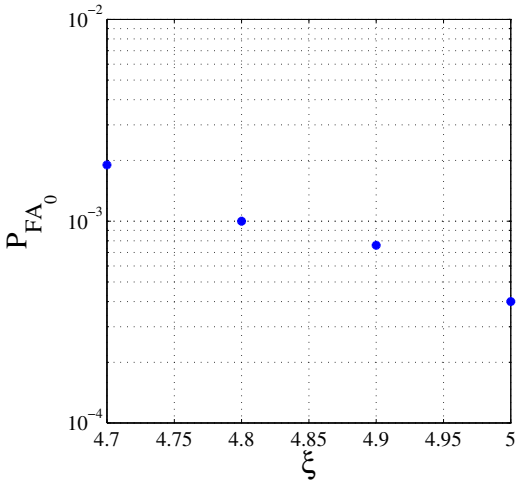



Figure 8.6: Empirical correspondence between $P_{FA_0}$ and $\xi$ (5.23) for the SVD 3D atom for $\xi = 4.7, 4.8, 4.9, 5$ obtained over $10^2$ noise realizations.

cube of the dry-run of size $X \times Y = 26 \times 28$ with $\Lambda = 3600$. Figure 8.8(a) illustrates the *noiseless* reference scene (arbitrary units), which contains three spatially localized Ly-$\alpha$ sources, highlighted by white circles (corresponding to objects number 6, 7 and 13 in Figure. 7.11). These sources appear spread because of the SSF. The spectra associated to the blue, magenta and green-crossed pixels are those shown in Figure 8.3.

We have run the detection tests at a same $P_{FA_0} = 10^{-3}$ for each spectral pixel. Similarly to Figure 7.9(a) for the $GLR_{1s}^{(1D)}$ detection test, the correspondence between the maximal probability of false alarm $P_{FA_0}$ and the $\xi$ test threshold values (5.23) for each spatio-spectral dictionary model-based test considered here is obtained by numerical simulation. As an illustration, Figure 8.6 shows the $P_{FA_0}$ vs. $\xi$ curve obtained for the $GLR_{1s}^{(3D)}$ test when performed with the SVD 3D atom, taking four different values of the test threshold $\xi = 4.7, 4.8, 4.9, 5$. A FA rate of $P_{FA_0} = 10^{-3}$ is here obtained for $\xi = 4.8$. To obtain the same $P_{FA_0}$, we found $\xi = 4.8$ for the $GLR_{1s}^{(3D)}$ test using the minimax 3D atom, and $\xi = 5.1$ for the $GLR_{1s}^{(3D)}$ using the $KSVD_7$ 3D dictionary.

Note that we use *estimated* noise variances provided by MUSE consortium, to make these simulations realistic as only estimated variances will be available on real data.

For the reasons listed below, the value of $P_{FA}$ varies from one test spectrum to another so we refer to this $P_{FA}$ as an indicative mean value for all the spectra of the data cube. First of all, our models assume knowledge of the true $\mathbf{\Sigma}$ but only estimates are available. Second, noise variances are spatially and spectrally variable in MUSE data. Hence, for each considered vectorized data sub-cube $\mathbf{s}_V(x, y)$ (6.38), different vectorized whitening sub-cubes $\mathbf{\Sigma}_V^{-\frac{1}{2}}$ (6.44) are considered and thus different whitened 3D atoms $\mathbf{\Sigma}_V^{-\frac{1}{2}}\mathbf{FHR}$ (6.44) are obtained at each $(x, y, \lambda)$ position, with which whitened sub-cubes $\mathbf{s}_V(x, y)$ are correlated. Note also that constant approximations (mean overall spectral bands) of the WSF and the SSF are considered in the design of the 3D atoms (cf. subsection 7.2.4). These approximations are not critical given the small variations of the spreading functions in $\Lambda$

and the high noise level affecting the data. According to this, not only an indicative one but one $P_{FA}(\xi)$ value per pixel should have been computed. This would be computationally very long and not really necessary (we checked by simulation). Third, these realistic simulated MUSE data cubes reflect possible errors in the estimation of the sky background (as this may happen in practice) and on its subtraction from the considered scene, with consequent possible variations on the variance and mean values of the whitened data distribution. All these reasons contribute to the variability of the $P_{FA}$. Nevertheless, as will be shown in Figure 8.8 this estimated $P_{FA}$ is fairly accurate.

Figure 8.8(b) and Figure 8.8(c) report the 1D model-based tests outcomes, respectively for the unconstrained GLR test (5.9), and the $GLR_{1s}^{(1D)}$ test (5.29) using the spectral $\mathbf{R}_{KSVD_7}$ dictionary. In these figures, the values of the test statistics above the threshold $\xi(P_{FA})$ are shown in color: in the black pixels, no detection is found at $P_{FA} = 10^{-3}$. As clearly visible, at this $P_{FA}$, none of the three objects in the scene is detected (similar null results are obtained with 1D tests using $\mathbf{R}_{SVD}$ and $\mathbf{R}_{minimax}$).

The detection results obtained with the 3D model-based $GLR_{1s}^{(3D)}$ test are shown in Figure 8.8(d) using a 3D atom with minimax spectral profile, in Figure 8.8(e) using a 3D atom with SVD spectral profile, and in Figure 8.8(f) using a dictionary of 7 3D atoms with K-SVD spectral profiles, at $P_{FA} = 10^{-3}$. This time, two among the three Ly-$\alpha$ *objects* in the scene are clearly detected. We checked that these are true detections by a comparison of the detected structures with the corresponding true spectra of objects $6, 7$ and 13 in the true scene of Figure 7.11. We note that some pixels (in the top right corner of the scene) lead to a slight increase in false alarm, probably because of estimation errors of the background, as mentioned before. The minimax and (K)-SVD approaches yield essentially comparable results (with slightly worse performances for the minimax approach on these three sources). The third source (bottom right corner of the scene) remains undetected: the features detected around pixel $(x, y) = (22, 22)$ do not actually correspond to real spectral features (the two other do). Due to its very weak signal level with respect to noise (pixel-vector $(x, y) = (22, 22)$ (in magenta, in Figure 8.3) is the faintest of the three considered Ly-$\alpha$ emitters) and to the fact that all important informations about the considered data (e.g. library of adapted line profiles, high and variable noise level, spatial and spectral spreading effects on the sources) were taken into account in the 3D model (6.46), we believe that this kind of signal illustrates a detection limit for this instrument.

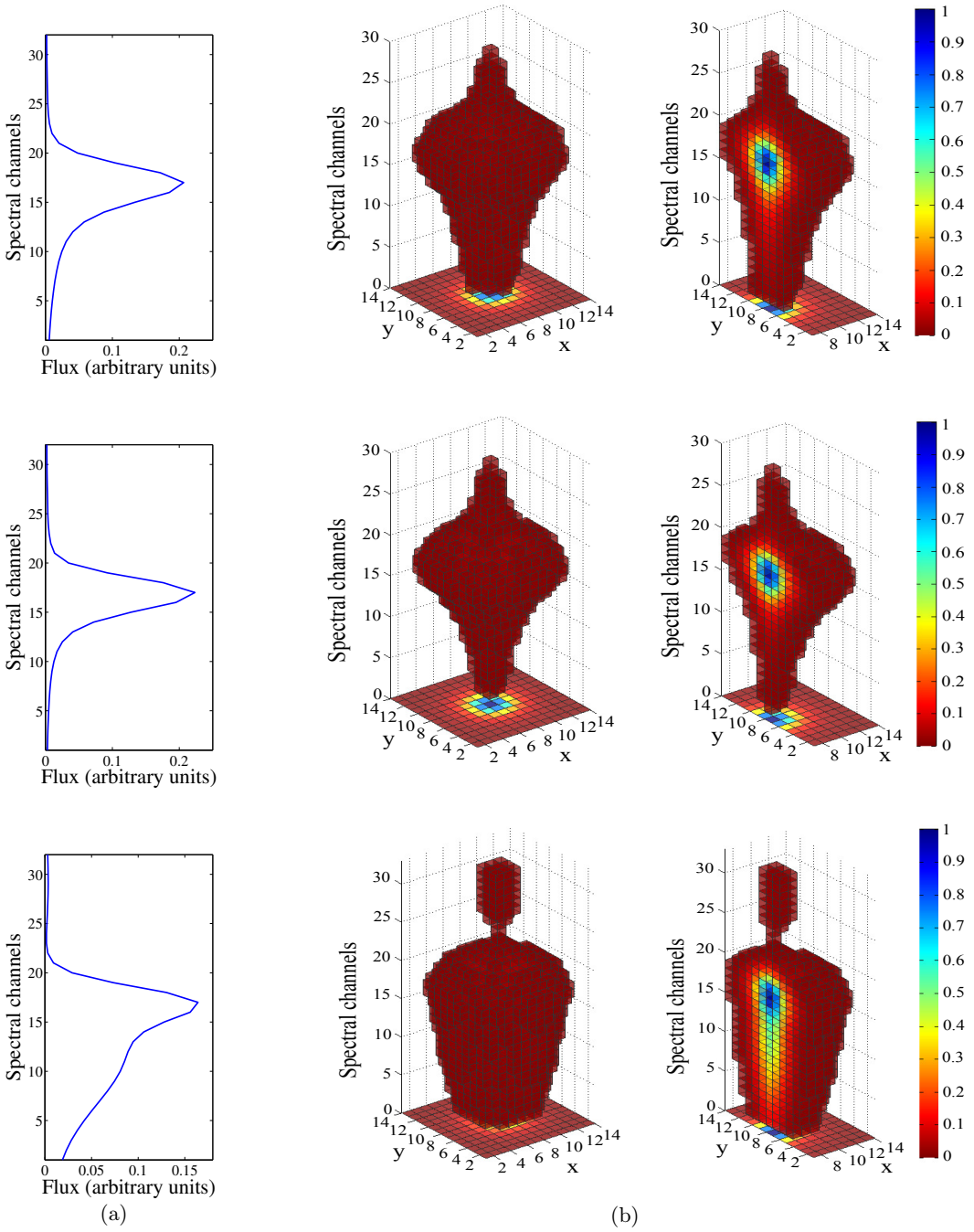



Figure 8.7: Examples of 3D SVD atoms. 8.7(a) three 1D learned atoms. Top: atom obtained by the SVD of $\mathcal{L}$; Middle and bottom: two atoms of the optimized $\text{KSVD}_7$ dictionary. 8.7(b) Left column: corresponding 3D atoms after composition by the SSF and the WSF. Right: a cut of these 3D atoms.

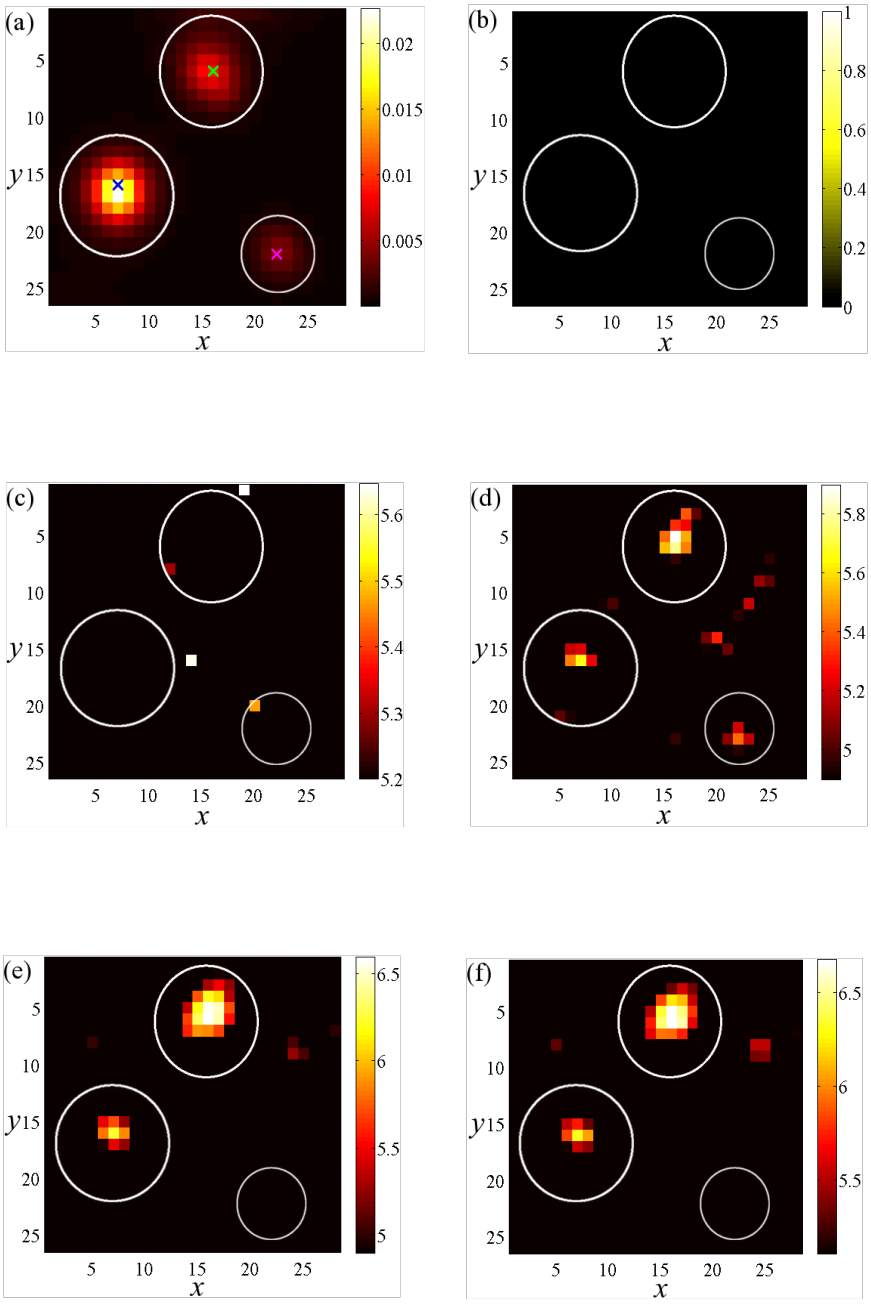



Figure 8.8: Compared detection performances of the unconstrained GLR (1.55), the $\mathrm{GLR}_{1s}^{(1D)}$ test (5.35) and the $\mathrm{GLR}_{1s}^{(3D)}$ test (6.48)) on a simulated MUSE data sub-cube. (a): Reference map (mean over all wavelengths). (b,c): GLR and $\mathrm{GLR}_{1s}^{(1D)}$ with $\mathbf{R}_{\mathrm{KSVD}_7}$. (d-f): $\mathrm{GLR}_{1s}^{(3D)}$ with minimax, SVD and $\mathrm{KSVD}_7$ 3D atoms.

## 8.4 Conclusion

This chapter has illustrated the detection results obtained by implementing the spatio-spectral model based detection strategies of chapter 6 in the context of the astrophysical hyperspectral data of the MUSE instrument.

Results on a whole MUSE data cube obtained performing the multi-step strategy (cascade of the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ test with several LR-MP$\beta$ detection rounds) were shown, evidencing that accounting for spatial dependencies is indeed useful to improve detection performances on very faint spectra, with respect to the use of the sole $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ test at a same $\mathrm{P_{FA}}$. While keeping the $\mathrm{P_{FA}}$ to reasonably low levels, the reason is that those faint spectra are indeed characterized by a spectral content similar to that of contiguous brighter pixel-vectors.

We then have shown the detection performances of the $\mathrm{GLR}_{1\mathrm{s}}^{(3\mathrm{D})}$ test (6.48) on very faint objects in the MUSE data cube. When applied to the hyperspectral data of the MUSE instrument, the $\mathrm{GLR}_{1\mathrm{s}}^{(3\mathrm{D})}$ detection method clearly improves on $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ approach. We believe there is not much room for amelioration, because:
i) the signals may appear at unknown location in $\lambda$ and are of unknown nature. We limited this drawback by using a specific dictionary;
ii) noise is high and variable in the spectral and spatial dimensions. We properly accounted for this by appropriate weighting.
The major informations on the data thus seem to have been taken into account in this 3D data model through the introduction of adapted dictionaries and the modeling of the instrument's WSF and SSF blurring effects. To decrease the $\mathrm{P_{FA}}$ while keeping the power essentially constant, possible optimizations are of course to put "cosmological constraints" on the translations of the sources in $\lambda$, and to shift the learned atoms at multiples of the sampling channels (instead of all channels). In addition, all tests could use the true (instead of approximated) SSF and WSF, to the detriment of a considerable increase of the approach complexity, for limited expected improvements.

# Conclusions and perspectives

# Summary and Conclusions

In this work we have presented new methods for the detection of weak sparse signals in high level of noise. The proposed detection approaches were based on statistical hypothesis testing principles. In the first part of the manuscript a general overview of hypothesis tests has been reported. A statistical hypothesis test calls for a reliable hypothetical description of the considered data and for the selection of an appropriate decision criterion, the test statistic, according to which decision is made on the veracity of the assertions we test. In order to control the probability of making wrong decisions (or probability of false alarm $\mathrm{P_{FA}}$), a test threshold also has to be chosen. Many examples of widely known hypothesis tests approaches have been provided, for both scalar and vector models, highlighting the differences between frequentist methods based on likelihood ratios (the LR and the GLR tests, whose detection surfaces in the two-dimensional case have also been analysed) and a Bayesian alternative, the Bayes Factor.

We have focused our attention on binary composite hypothesis tests, that is, tests composed by only two different claims that depend on unknown parameters: the null hypothesis $\mathcal{H}_0$, under which the data have been described as an only noise signal, and the alternative hypothesis $\mathcal{H}_1$, under which the data have appeared as noise plus an unknown signal vector.

Interested in relatively modeling and locating the signal vector's subspace hosting the information of interest, we have concentrated our attention on the computation of sparsity-promoting estimates. An efficient solution was found computing a MAP estimate with suitable prior law, an approach also interpretable as a PLS problem: when calculated using double exponential-like shaped priors with flat tails, the MAP estimate actually leads to sparsity-promoting thresholding functions. In particular, in the case of a Laplacian prior law, the MAP estimate yields the soft-thresholding operator. This fact has been exploited in the design of the proposed detection approaches.

The second part of the manuscript has been divided into two sub-parts. In the fist sub-part, we have introduced the proposed detection tests, called the LRMAP and the PDR tests, for a simple data model. Directly inspired from the GLR test, the LRMAP test statistic computes the ratio of the likelihoods under the two hypotheses of the model with a thresholded sparse parameter estimate under $\mathcal{H}_1$. On the other hand, the ratio of the posterior densities was considered for the PDR. This has led to the identification of two thresholding test statistics, $T_{\mathrm{LRMAP}}$ and $T_{\mathrm{PDR}}$, well adapted for the detection of sparse signals. Connection was found and mentioned with existing approaches in the literature. The LRMAP and the PDR tests have been studied and their properties analyzed in both the scalar and the vector cases. Numerical comparisons with two widely used classical detection methods, the frequentist GLR test on one side and the Bayesian Bayes Factor on the other, have shown in the vector case the higher detection power of the proposed tests.

Interested in better understanding the relative behavior of the LRMAP and PDR tests, which is difficult to analyse because of their strong parameter dependency, we have conducted a detailed comparative analysis based on the geometry of the two tests detection regions. This study has allowed to specify analytical conditions on the value of the tests' parameters (the $\gamma$ and $\eta$ thresholds), for which one approach is better adapted than the other towards sparse signals detection goals. Both tests have shown to be equivalent and

adapted to 1−sparse signals for $\gamma = 0$. Nevertheless, for $\eta \neq 0$ and$\gamma \neq 0$, the PDR tends to better detect highly sparse signals while the LRMAP is more adapted to less sparse signals.

In the second sub-part, we have formulated the proposed detection methods in the case of more refined data models accounting for adapted redundant dictionaries. One-dimensional (spectral) dictionary-based data models were firstly considered. In this framework, we have shown that for $\gamma = 0$ the LRMAP and the PDR resume to a same test, which we called the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ detector. We proved that this test is equivalent to the *Max* test of [Arias-Castro 2010], an optimal detector in the case of high sparsity levels. We have also numerically shown that for highly sparse signals our approach has similar detection power to that of the HC method, another powerful detector for faint sparse signals in noise, coming from the statistical literature.

We secondly presented detection methods seeking to exploit both the spectral similarities of neighbors pixels in the spatial domain, and the greater accuracy of dictionary-based models taking into account the spatio-spectral blur of information caused by instrumental Point Spread Functions. First, the following two-step methodology, exploiting both the spatial and spectral dependencies of neighbors pixels in a data cube, has been developed: after a first detection round using the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ test, we have introduced a multiple detection round performing a matched filter-like test, which we have called the LR-MP$\beta$ test, on spectra not detected by the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ test but contiguous to bright detected ones. This approach was shown to be able to exploit the neighbors' common spectral information for the detection of faintest signals that would not be detected with the sole use of the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ strategy. Second, we presented a detection method based on a three-dimensional (spatio-spectral) data model, accounting for the information spread caused by the instrument PSF not only in the spectral direction but also in the spatial ones. This new data model has led to a test named the $\mathrm{GLR}_{1\mathrm{s}}^{(3\mathrm{D})}$ test. We underline that all the proposed detection strategies allowed to take account of noise characteristics, strongly variable both spectrally and spatially, and to reliably control the FA rate.

The last part of the manuscript has concerned the application of the proposed dictionary-based detection tests to simulated but very realistic astrophysical hyperspectral data, provided by the ESO's MUSE instrument consortium. In the case of one-dimensional (spectral) dictionary-based data models, we have shown that thanks also to the use of an adapted redundant dictionary designed in accordance to the spectral characteristics of the signals sought, the sparsity-based $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ outperforms the classical (unconstrained) GLR test when considering a collection of spectra in a hyperspectral data cube.

In the case of spatio-spectral detection approaches, when using the multiple-step strategy accounting for simple spatial dependencies between neighbors pixels, improved detection performances have been achieved on very faint spectra contiguous to brighter ones, while maintaining the $\mathrm{P_{FA}}$ rate reasonably low. Although the detection performances of the $\mathrm{GLR}_{1\mathrm{s}}^{(1\mathrm{D})}$ test and of the multiple-step approach were satisfactory with respect to other techniques (e.g. the unconstrained GLR or the HC tests) for most test sources, these kind of strategies are still insufficient to detect extremely faint and sparse sources such as the Ly$-\alpha$ objects present in the considered MUSE hyperspectral data cubes.

In response to this challenge, we have set up and compared a test called $\mathrm{GLR}_{1\mathrm{s}}^{(3\mathrm{D})}$ for three different dictionaries of 3D atoms computed using dictionary learning techniques such as KSVD and minimax methods. The test has exhibited increased detection power with

respect to the previous tests, allowing thus to detect some of the faintest Ly$-\alpha$ objects in MUSE data cube and evidencing probable detection limit of the instrument for such sources.

# Perspectives

All the signal detection approaches developed in this thesis have been validated on a number of numerical simulations. Their numerous tests allowed us to identify several improvements.

The first element concerns the study of the PDR and LRMAP tests for prior laws different from Laplacian. MAP-based thresholding functions can be generally obtained by using prior distributions that are mono-modal and not differentiable at the origin. Hence, it would be interesting to undertake a comparative study of the thresholding properties of the MAP estimate using diverse prior laws. Nevertheless, we believe that substantially different results would not be obtained (due to the relative similarity of the resulting thresholding functions) while considering other types of priors would heavily complicate the analysis and implementation of the detection tests, which we showed to be already difficult in the Laplacian case considered in this work.

In the footsteps of the study proposed in chapter 4, another possible way to compare the performances of the PDR and LRMAP tests could be to geometrically analyze, through a maximization/minimization approach, the corresponding values of the $P_{DET}$ for a same value of the $P_{FA}$, numerically evaluated in function of both the $\eta$ and $\gamma$ parameters. The bases for such a problem, which is under investigation, have been presented in Appendix B.5.

It would be also very helpful to refine the dictionary $\mathbf{R} = [\mathbf{R}^{\ell}\,\mathbf{R}^{c}\,\mathbf{R}^{b}]$, which is still somewhat schematic, and to further investigate on the design and computation of specific dictionaries using appropriate learning techniques. In fact, the design of dictionaries for the detection of specific sources, such as for example Ly$-\alpha$ objects, plays an important role for increasing the performances of the tests. Encouraging results have been shown in the third part of this thesis where improved detection results (with respect to not-learned dictionaries) have been obtained on MUSE hyperspectral data cubes, with dictionaries learned using KSVD and minimax approaches and provided by [Suleiman 2013].

Concerning the spatio-spectral dictionary-based test introduced in section 6.3, possible optimizations to decrease the $P_{FA}$ while keeping the power constant would be to put cosmological constraints on the translations of the sources in the spectral dimension, and to shift the learned atoms at multiples of the sampling channels (instead of all channels). In addition, all tests could use the true (instead of the approximated) SSF and WSF, to the detriment of a considerable increase of complexity.

At last, but not least, it could be very interesting (and somehow rewarding) to see how the detection approaches developed during these three years of research will perform on MUSE's real hyperspectral data, as soon as the instrument is operating at the VLT (a version of the code is under preparation for its implementation in MUSE's pipeline).

## A.1 Proof of NP Lemma

We seek to find the critical region $\mathcal{Z}_1 = \{\mathbf{x} : \mathcal{H}_1\}$ in which we decide to accept $\mathcal{H}_1$ so that the

$$\mathrm{P}_{\mathrm{DET}} = \int_{\mathcal{Z}_1} p(\mathbf{x} \mid \mathcal{H}_1)\mathrm{d}\mathbf{x} \tag{A.1}$$

is maximum for a given

$$\mathrm{P}_{\mathrm{FA}} = \int_{\mathcal{Z}_1} p(\mathbf{x} \mid \mathcal{H}_0)\mathrm{d}\mathbf{x} = \alpha. \tag{A.2}$$

Here, $p(\mathbf{x} \mid \mathcal{H}_0)$ and $p(\mathbf{x} \mid \mathcal{H}_1)$ represent the known PDFs' of $\mathbf{x}$ under the null and the alternative hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$. We consider the following Lagrangian $\mathcal{L}$

$$\mathcal{L} = \mathrm{P}_{\mathrm{DET}} + \lambda\left(\mathrm{P}_{\mathrm{FA}} - \alpha\right), \tag{A.3}$$

where $\lambda \in \mathbb{R}$ is a Lagrange multiplier. According to (A.1) and (A.2), (A.3) is equal to

$$\begin{aligned} \mathcal{L} &= \int_{\mathcal{Z}_1} p(\mathbf{x} \mid \mathcal{H}_1)\mathrm{d}\mathbf{x} + \lambda\left(\int_{\mathcal{Z}_1} p(\mathbf{x} \mid \mathcal{H}_0)\mathrm{d}\mathbf{x} - \alpha\right) \\ &= \int_{\mathcal{Z}_1} \left(p(\mathbf{x} \mid \mathcal{H}_1) + \lambda p(\mathbf{x} \mid \mathcal{H}_0)\right)\mathrm{d}\mathbf{x} - \lambda\alpha. \end{aligned} \tag{A.4}$$

To maximize $\mathcal{L}$ and hence the $\mathrm{P}_{\mathrm{DET}}$ for a $\mathrm{P}_{\mathrm{FA}} = \alpha$, we should include $\mathbf{x} \in \mathcal{Z}_1$ if

$$p(\mathbf{x} \mid \mathcal{H}_1) + \lambda p(\mathbf{x} \mid \mathcal{H}_0) > 0. \tag{A.5}$$

We thus decide $\mathcal{H}_1$ if

$$\frac{p(\mathbf{x} \mid \mathcal{H}_1)}{p(\mathbf{x} \mid \mathcal{H}_0)} > -\lambda. \tag{A.6}$$

Since the LR is always nonnegative, we let $\gamma = -\lambda$ so that we decide $\mathcal{H}_1$ when

$$\frac{p(\mathbf{x} \mid \mathcal{H}_1)}{p(\mathbf{x} \mid \mathcal{H}_0)} > \gamma, \tag{A.7}$$

where the threshold $\gamma > 0$ is found from $\mathrm{P}_{\mathrm{FA}} = \alpha$.

## A.2 Proof of BF as the minimum $\mathrm{P_E}$ detector

The proof is addressed in terms of the minimization of the Bayes Risk $\mathcal{R}$ [Kay 1998b]

$$\begin{aligned} \mathcal{R} &= \sum_{i=0}^{1}\sum_{j=0}^{1} K_{ij} P(\mathcal{H}_i \mid \mathcal{H}_j) P(\mathcal{H}_j) \\ &= K_{00}P(\mathcal{H}_0 \mid \mathcal{H}_0)P(\mathcal{H}_0) + K_{01}P(\mathcal{H}_0 \mid \mathcal{H}_1)P(\mathcal{H}_1) \\ &\quad + K_{10}P(\mathcal{H}_1 \mid \mathcal{H}_0)P(\mathcal{H}_0) + K_{11}P(\mathcal{H}_1 \mid \mathcal{H}_1)P(\mathcal{H}_1), \end{aligned} \tag{A.8}$$

where $K_{ij}$ are constants representing the cost if we decide $\mathcal{H}_i$ but $\mathcal{H}_j$ is true. Since for $K_{00} = K_{11} = 0$ and $K_{10} = K_{01} = 1$ we have that $\mathcal{R} = \mathrm{P_E}$, this proof also applies to the minimum $\mathrm{P_E}$ problem. We define $\mathcal{Z}_1 = \{\mathbf{x} : \text{decide } \mathcal{H}_1\}$ the critical region and $\mathcal{Z}_0$ its complement. Then,

$$\begin{aligned}\mathcal{R} &= K_{00}P(\mathcal{H}_0)\int_{\mathcal{Z}_0} p(\mathbf{x} \mid \mathcal{H}_0)\mathrm{d}\mathbf{x} + K_{01}P(\mathcal{H}_1)\int_{\mathcal{Z}_0} p(\mathbf{x} \mid \mathcal{H}_1)\mathrm{d}\mathbf{x} \\ &\quad + K_{10}P(\mathcal{H}_0)\int_{\mathcal{Z}_1} p(\mathbf{x} \mid \mathcal{H}_0)\mathrm{d}\mathbf{x} + K_{11}P(\mathcal{H}_1)\int_{\mathcal{Z}_1} p(\mathbf{x} \mid \mathcal{H}_1)\mathrm{d}\mathbf{x}. \end{aligned} \tag{A.9}$$

Since $\mathcal{Z}_1$ and $\mathcal{Z}_0$ partition the entire space, using

$$\int_{\mathcal{Z}_0} p(\mathbf{x} \mid \mathcal{H}_i)\mathrm{d}\mathbf{x} = 1 - \int_{\mathcal{Z}_1} p(\mathbf{x} \mid \mathcal{H}_i)\mathrm{d}\mathbf{x}, \tag{A.10}$$

we have

$$\begin{aligned}\mathcal{R} &= K_{00}P(\mathcal{H}_0) + K_{01}P(\mathcal{H}_1) \\ &\quad + \int_{\mathcal{Z}_1} \left[\left(K_{10}P(\mathcal{H}_0) - K_{00}P(\mathcal{H}_0)\right) p(\mathbf{x} \mid \mathcal{H}_0) + \left(K_{11}P(\mathcal{H}_1) - K_{01}P(\mathcal{H}_1)\right) p(\mathbf{x} \mid \mathcal{H}_1)\right] \mathrm{d}\mathbf{x}.\end{aligned}$$

At this point, we decide $\mathcal{H}_1$ if

$$\left(K_{10} - K_{00}\right) P(\mathcal{H}_0)p(\mathbf{x} \mid \mathcal{H}_0) < \left(K_{01} - K_{11}\right) P(\mathcal{H}_1)p(\mathbf{x} \mid \mathcal{H}_1). \tag{A.11}$$

Assuming $K_{10} > K_{00}, K_{01} > K_{11}$, we obtain finally

$$\frac{p(\mathbf{x} \mid \mathcal{H}_1)}{p(\mathbf{x} \mid \mathcal{H}_0)} > \frac{(K_{10} - K_{00})P(\mathcal{H}_0)}{(K_{01} - K_{11})P(\mathcal{H}_1)} = \gamma, \tag{A.12}$$

which corresponds to the definition of the minimum $\mathrm{P_E}$ detector given in (1.68) that is, the BF.

## A.3 Proof of Property 3.4.1

*Prop. 3.4.1.1:*

Since $T(u)$ is even and strictly increasing for $u \geq u_0$, for $|u| \geq u_0$ we have that

$$\begin{aligned} T(u) \underset{H_0}{\overset{H_1}{\gtrless}} \gamma_{test} &\Leftrightarrow |u| \underset{H_0}{\overset{H_1}{\gtrless}} T^{-1}(\gamma_{test}) \\ &\Leftrightarrow u^2 \underset{H_0}{\overset{H_1}{\gtrless}} \left[T^{-1}(\gamma_{test})\right]^2, \end{aligned}$$

which is the same test as GLR of (3.21) by choosing $\gamma_{GLR}$ as in (3.26). Since the condition $|u| \geq u_0$ is equivalent to $T(u) > T(u_0)$, the test $T(u) \underset{H_0}{\overset{H_1}{\gtrless}} \gamma_{test}$ and the

GLR coincide for $\gamma_{test} \geq \gamma_{test_0}$, with $\gamma_{test_0} = T(u_0)$. The false alarm of the test is consequently

$$\begin{aligned} P_{FA_{test}}(\gamma_{test}) &= P_{FA_{GLR}}\left(\left[T^{-1}(\gamma_{test})\right]^2\right) \\ &= Pr\left(u^2 > \left[T^{-1}(\gamma_{test_0})\right]^2 \mid \mathcal{H}_0\right) \\ &= 1 - \Phi_{\chi^2_1}\left(\left[T^{-1}(\gamma_{test_0})\right]^2\right), \end{aligned}$$

which shows (3.28). Computing $P_{FA_{test}}(\gamma_{test_0} = T(u_0))$ directly yields $P_{FA_0}$, which proves Prop. 3.4.1.1.

*Prop. 3.4.1.2:*

Since $T(u_0) \geq 0$, $\gamma_{test_0} \geq 0$. Then $\forall \gamma_{test} \in [0; \gamma_{test_0}]$, we have

$$\begin{aligned} T(u) \underset{H_0}{\overset{H_1}{\gtrless}} \gamma_{test} \;&\Leftrightarrow\; |u| \underset{H_0}{\overset{H_1}{\gtrless}} u_0 \\ &\Leftrightarrow\; T(u) \underset{H_0}{\overset{H_1}{\gtrless}} T(u_0), \end{aligned}$$

which is the test obtained with $\gamma_{test_0}$. Consequently,

$$P_{FA_{test}} = P_{FA_0}, \qquad \forall \gamma_{test} \in [0; \gamma_{test_0}],$$

which completes the proof.

## A.4 Proof of Proposition 4.3.1

The case of $n \geq 2$ components $u_i > \eta$ is firstly treated, before considering the particular case of $n = 1$.
Case $n \geq 2$: Similarly to (4.3) and (4.5), the $\mathcal{S}^{(n)}_{PDR}$ and $\mathcal{S}^{(n)}_{LRMAP}$ domains, for $n \geq 2$, represent two $n$-spheres of equations

$$\mathcal{S}^{(n)}_{PDR}: \quad \left\{ \boldsymbol{u} \mid \sum_{i=1}^{n} (u_i - \eta)^2 = \gamma_P,\ u_i \leq \eta, i > n \right\}, \tag{A.13}$$

$$\mathcal{S}^{(n)}_{LRMAP}: \quad \left\{ \boldsymbol{u} \mid \sum_{i=1}^{n} u_i^2 = \gamma_L + n\eta^2,\ u_i \leq \eta, i > n \right\}. \tag{A.14}$$

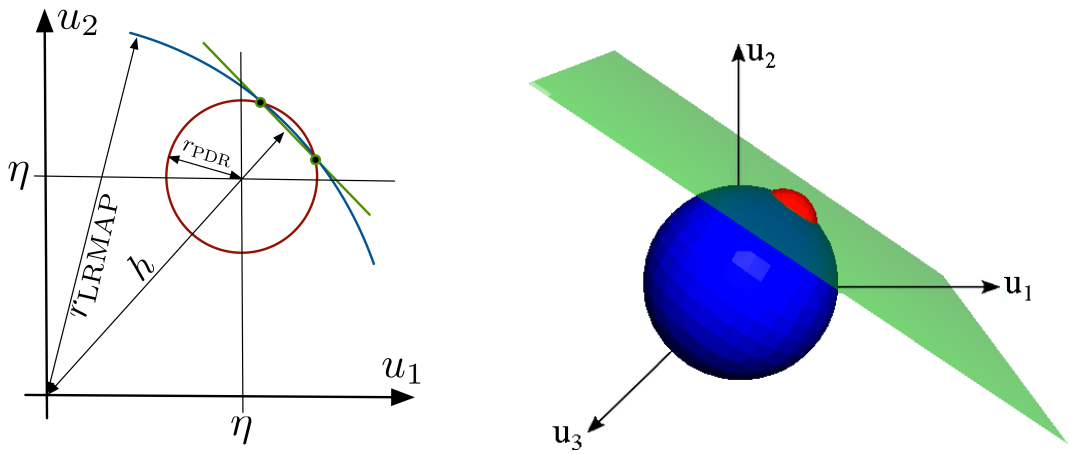
When intersections between $\mathcal{S}^{(n)}_{PDR}$ and $\mathcal{S}^{(n)}_{LRMAP}$ occur, two conditions are verified: first of all,

$$\mid r_{LRMAP} - r_{PDR} \mid < d < r_{LRMAP} + r_{PDR}, \tag{A.15}$$

where $r_{LRMAP}$ and $r_{PDR}$ are the radii of $\mathcal{S}^{(n)}_{PDR}$ and $\mathcal{S}^{(n)}_{LRMAP}$, and $d$ represents the distance between the centers of the two $n$-spheres; secondly,

$$r^2_{LRMAP} - h^2 < (h - d)^2, \tag{A.16}$$

that comes from the fact that the intersection between two $n$-spheres is an $n$-circle which center-origin distance is equal to $h = (r^2_{LRMAP} - r^2_{PDR} + d^2)/2d$ [Kern 1948]. Figure




(a) (b)

Figure A.1: Intersections between the $\mathcal{S}^{(n)}_{PDR}$ and $\mathcal{S}^{(n)}_{LRMAP}$ domains when $n \geq 2$. Figure A.1(a) reports the $2D$ case. In this case the two $n$-spheres reduce to two circles intersecting in two points (black points in the figure) that lay on a same line, reported in green. Figure A.1(b) shows the $3D$ case. Here, $\mathcal{S}^{(3)}_{PDR}$ and $\mathcal{S}^{(3)}_{LRMAP}$ are two spheres, the intersection of whom is a circle lying on the intersection plane plotted in green.

A.1 illustrates the case of intersection between the PDR (in red) and LRMAP (in blue)

subsurfaces. In particular, Figure A.1(a) illustrates the case of a $N = 2$ component vector, while in Figure A.1(b) we consider the case $N = 3$. Figure A.1(a) shows that in the $2D$ case, the two $n$-spheres reduce to two circles intersecting in two points (in black), lying on a same line (in green). On the other hand, as reported in Figure A.1(b), when $N = 3$, $\mathcal{S}^{(3)}_{PDR}$ and $\mathcal{S}^{(3)}_{LRMAP}$ constitute two intersecting spheres. This time, the corresponding intersection is a circle, which lies on the plane reported in green. According to definitions in (A.13) and (A.14), the condition in (A.15) leads to

$$\sqrt{\gamma_L + n\eta^2} - \sqrt{\gamma_P} < \sqrt{n}\eta < \sqrt{\gamma_L + n\eta^2} + \sqrt{\gamma_P}. \tag{A.17}$$

The right inequality of (A.17) is always verified. On the other hand, expressing $\gamma_L$ as a function of $\gamma_P$, the left side of the condition gives the following upper bound to the $\gamma_L$ values: $\gamma_L < \gamma_P + 2\eta\sqrt{n\gamma_P}$. Condition (A.16) finally gives the following lower bound on $\gamma_L$: $\gamma_L > \gamma_P + \eta\sqrt{2n\gamma_P}$. We thus obtain that, in order to have intersections between $\mathcal{S}^{(n)}_{PDR}$ and $\mathcal{S}^{(n)}_{LRMAP}$, we must choose $\gamma_L > \gamma_P$ and so that $\gamma_P + \eta\sqrt{2n\gamma_P} < \gamma_L < \gamma_P + 2\eta\sqrt{n\gamma_P}$.
Case $n = 1$: In this particular case, the LRMAP and PDR detection regions reduce to $\mathcal{S}^{(1)}_{PDR} : (u_1 - \eta)^2 = \gamma_P$ and $\mathcal{S}^{(1)}_{LRMAP} : u_1^2 = \gamma_L + \eta^2$. Combining these two equations, the condition we obtain on $\gamma_L$, in order to have $\mathcal{S}^{(1)}_{PDR} < \mathcal{S}^{(1)}_{LRMAP}$, is: $\gamma_L > \gamma_P + 2\eta\sqrt{\gamma_P}$. On the contrary, when $\gamma_L < \gamma_P + 2\eta\sqrt{\gamma_P}$, we must have $\mathcal{S}^{(1)}_{PDR} > \mathcal{S}^{(1)}_{LRMAP}$.

## B.1 Computation of the Hard Thresholding operator

The scalar HT operator can be derived solving the PLS problem

$$\hat{\theta}_{\text{PLS}} = \arg\min_{\theta} J(\theta) \tag{B.1}$$

with cost function

$$J(\theta) = \frac{1}{2}(x-\theta)^2 + \varphi_\mu(\theta), \tag{B.2}$$

and non-convex clipped $l_1$ penalty function of the form (2.15) [Antoniadis 2001]

$$\varphi_\mu(\theta) = \mu^2 - \frac{1}{2}(|\theta| - \mu)^2 I(|\theta| < \mu).$$

Since $J(\theta)$ is an even function, we focus only on the case $\theta \geq 0$ (symmetric results are obtained with a similar reasoning in the case of $\theta \geq 0$).

- $\theta > \mu$: we want to find $\theta$ so that the following cost function is minimized

  $$J(\theta) = \frac{1}{2}(x-\theta)^2 + \mu^2. \tag{B.3}$$

  This problem has straightforward solution that is $\hat{\theta} = x$, leading to

  $$J(\mu) = \mu^2. \tag{B.4}$$

- $0 \leq \theta \leq \mu$: the cost function is in this case

  $$\begin{aligned} J(\theta) &= \frac{1}{2}(x-\theta)^2 + \mu^2 - \frac{1}{2}(\theta-\mu)^2 \\ &= \frac{1}{2}(x^2+\mu^2) + \theta(\mu - x). \end{aligned} \tag{B.5}$$

  For $x > \mu$, the quantity $\mu - x < 0$. The largest value of $\theta$ that minimizes (B.5) is thus equal to $\hat{\theta} = \mu$. For this value, the cost function (B.5) is

  $$J(\mu, x) = \frac{1}{2}(x-\mu)^2 + \mu^2, \tag{B.6}$$

  but since $J(\mu, x)$ in (B.6) is always larger than $J(\mu)$ in (B.4), the optimal solution is actually $\hat{\theta} = x$.

  For $x \leq \mu$, the quantity $\mu - x \geq 0$. In this case, the lowest value of $\theta$ that minimizes (B.5) is $\hat{\theta} = 0$. This solution yields

  $$J(\mu, x) = \frac{1}{2}(x^2 - \mu^2), \tag{B.7}$$

  which is always smaller than $J(\mu)$ in (B.4).

For $\theta \geq 0$ we thus get

$$\hat{\theta} = \begin{cases} 0, \text{ if } 0 \leq x \leq \mu \\ x, \text{ if } x > \mu. \end{cases} \tag{B.8}$$

Combining these results with the ones considering $\theta \leq 0$, finally yields the HT estimator of (2.18).

## B.2 Computation of the BF test statistic for model (3.18)

According to model (3.18) the BF test is defined as

$$\mathrm{BF}(x) = \frac{\int_{\mathbb{R}} p\left(x \mid \theta\right) \pi\left(\theta\right) \mathrm{d}\theta}{p\left(x \mid 0\right)} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma. \tag{B.9}$$

For a gaussian likelihood and a Laplacian prior, the computation of the BF test statistic leads to

$$\begin{aligned} T_{\mathrm{BF}}(x) &= \frac{1}{2\lambda} \exp\left(\frac{x^2}{2\sigma^2}\right) \int_{\mathbb{R}} \exp\left[-\frac{(x-\theta)^2}{2\sigma^2} - \frac{|\theta|}{\lambda}\right] \\ &= \frac{1}{2\lambda} \exp\left(\frac{x^2}{2\sigma^2}\right) \left\{\int_{\mathbb{R}^+} \exp\left[-\frac{(x-\theta)^2}{2\sigma^2} - \frac{\theta}{\lambda}\right] \mathrm{d}\theta + \int_{\mathbb{R}^-} \exp\left[-\frac{(x-\theta)^2}{2\sigma^2} + \frac{\theta}{\lambda}\right] \mathrm{d}\theta\right\}. \end{aligned} \tag{B.10}$$

The two integrals in (B.10) can be re-written as

$$\begin{aligned} \int_{\mathbb{R}^+} \exp\left[-\frac{(x-\theta)^2}{2\sigma^2} - \frac{\theta}{\lambda}\right] \mathrm{d}\theta &= \int_{\mathbb{R}^+} \exp\left\{-\frac{1}{2\sigma^2}\left[\left(\theta - \left(x - \frac{\sigma^2}{\lambda}\right)\right)^2 - \left(x - \frac{\sigma^2}{\lambda}\right)^2 + x^2\right]\right\} \mathrm{d}\theta \\ &= \exp\left(\frac{\sigma^2}{2\lambda^2} - \frac{x}{\lambda}\right) \int_{\mathbb{R}^+} \exp\left\{-\frac{1}{2\sigma^2}\left[\left(\theta - \left(x - \frac{\sigma^2}{\lambda}\right)\right)^2\right]\right\} \mathrm{d}\theta \end{aligned} \tag{B.11}$$

and

$$\begin{aligned} \int_{\mathbb{R}^-} \exp\left[-\frac{(x-\theta)^2}{2\sigma^2} + \frac{\theta}{\lambda}\right] \mathrm{d}\theta &= \int_{\mathbb{R}^-} \exp\left\{-\frac{1}{2\sigma^2}\left[\left(\theta - \left(x + \frac{\sigma^2}{\lambda}\right)\right)^2 - \left(x + \frac{\sigma^2}{\lambda}\right)^2 + x^2\right]\right\} \mathrm{d}\theta \\ &= \exp\left(\frac{\sigma^2}{2\lambda^2} + \frac{x}{\lambda}\right) \int_{\mathbb{R}^-} \exp\left\{-\frac{1}{2\sigma^2}\left[\left(\theta - \left(x + \frac{\sigma^2}{\lambda}\right)\right)^2\right]\right\} \mathrm{d}\theta \end{aligned} \tag{B.12}$$

Using the definition of the right tail probability of a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ with mean $\mu$ and standard deviation $\sigma$ (also called $Q-$function)

$$\begin{aligned} Q\left(\frac{\gamma - \mu}{\sigma}\right) &= 1 - \Phi\left(\frac{\gamma - \mu}{\sigma}\right) \\ &= \frac{1}{\sqrt{2\pi}} \int_{\gamma}^{+\infty} \exp\left[-\frac{(\theta - \mu)^2}{2\sigma^2}\right] \mathrm{d}\theta, \end{aligned} \tag{B.13}$$

and taking it for $\gamma = 0$ and $\mu = x - \frac{\sigma^2}{\lambda}$, we obtain for the integral over $\mathbb{R}^+$

$$\int_{\mathbb{R}^+} \exp\left\{-\frac{1}{2\sigma^2}\left[\left(\theta - \left(x - \frac{\sigma^2}{\lambda}\right)\right)^2\right]\right\} \mathrm{d}\theta = \sqrt{2\pi}\left[1 - \Phi\left(-\frac{x}{\sigma} + \frac{\sigma}{\lambda}\right)\right]. \tag{B.14}$$

Replacing (B.14) in (B.11) gives

$$\int_{\mathbb{R}^+} \exp\left[-\frac{(x-\theta)^2}{2\sigma^2} - \frac{\theta}{\lambda}\right] \mathrm{d}\theta = \sqrt{2\pi} \exp\left(\frac{\sigma^2}{2\lambda} - \frac{x}{\lambda}\right)\left[1 - \Phi\left(-\frac{x}{\sigma} + \frac{\sigma}{\lambda}\right)\right]. \tag{B.15}$$

A similar reasoning is done for the integral over $\mathbb{R}^-$. For the integral in (B.12), using definition (B.13) with $\gamma = 0$ and $\mu = x + \frac{\sigma^2}{\lambda}$, we have that

$$\begin{aligned}\int_{\mathbb{R}^-} \exp\left\{-\frac{1}{2\sigma^2}\left[\left(\theta - \left(x + \frac{\sigma^2}{\lambda}\right)\right)^2\right]\right\} \mathrm{d}\theta &= \int_{\mathbb{R}^+} \exp\left\{-\frac{1}{2\sigma^2}\left[\left(\theta - \left(-x - \frac{\sigma^2}{\lambda}\right)\right)^2\right]\right\} \mathrm{d}\theta \\ &= \sqrt{2\pi}\left[1 - \Phi\left(\frac{x}{\sigma} + \frac{\sigma}{\lambda}\right)\right]. \end{aligned} \tag{B.16}$$

Replacing (B.16) in (B.12) gives

$$\int_{\mathbb{R}^-} \exp\left[-\frac{(x-\theta)^2}{2\sigma^2} + \frac{\theta}{\lambda}\right] \mathrm{d}\theta = \sqrt{2\pi} \exp\left(\frac{\sigma^2}{2\lambda} + \frac{x}{\lambda}\right)\left[1 - \Phi\left(\frac{x}{\sigma} + \frac{\sigma}{\lambda}\right)\right]. \tag{B.17}$$

Now, inserting in (B.10) the expressions found in (B.15) and (B.17) we finally have

$$\begin{aligned} T_{\mathrm{BF}}(x) &= \frac{\sqrt{2\pi}}{2\lambda} \exp\left(\frac{x^2}{2\sigma^2} + \frac{\sigma^2}{2\lambda^2}\right)\left\{\exp\left(-\frac{x}{\lambda}\right)\left[1 - \Phi\left(-\frac{x}{\sigma} + \frac{\sigma}{\lambda}\right)\right] + \exp\left(\frac{x}{\lambda}\right)\left[1 - \Phi\left(\frac{x}{\sigma} + \frac{\sigma}{\lambda}\right)\right]\right\} \\ &= \frac{\sqrt{2\pi}}{2\lambda} \exp\left(\frac{x^2}{2\sigma^2} + \frac{\sigma^2}{2\lambda^2}\right)\left\{2\cosh\left(\frac{x}{\lambda}\right) - \exp\left(-\frac{x}{\lambda}\right)\Phi\left(-\frac{x}{\sigma} + \frac{\sigma}{\lambda}\right) - \exp\left(\frac{x}{\lambda}\right)\Phi\left(\frac{x}{\sigma} + \frac{\sigma}{\lambda}\right)\right\}. \end{aligned} \tag{B.18}$$

## B.3 Learning techniques for dictionary selection: the K-SVD approach

A way to design dictionaries is to use learning techniques. Learning techniques allow to obtain dictionaries whose content is adapted to fit a considered set of signals. In this Appendix, we present the general dictionary learning problem for sparse representation of signals. A detailed analysis of the KSVD approach introduced in [Aharon 2006] follows.

Consider a set of K *training* signals in $\mathbb{R}^N$, $\{\mathbf{x}_k\}_{k=1,\dots,K}$. In matrix form, the set of all the training signals is represented by a matrix $\mathbf{X}$ of dimension $N \times K$, in which each $\mathbf{x}_k$ is a column vector. We seek to sparsely approximate the matrix $\mathbf{X}$ as a linear combination of few elementary features in a set noted $\mathbf{R}$:

$$\mathbf{X} \approx \mathbf{R}\mathbf{A}. \tag{B.19}$$

The columns of matrix $\mathbf{R}$, which is of dimensions $N \times P$, represent the $P$ elements of the dictionary $\mathbf{R} = [\boldsymbol{r}_1, \dots, \boldsymbol{r}_P]$ we seek to learn, while the matrix $\mathbf{A}$, of dimensions $P \times K$, contains all the synthesis coefficients $\{\boldsymbol{a}_k\}_{k=1,\dots,K}$ of $\mathbf{X}$ on $\mathbf{R}$.

The problem here is thus to find matrices $\mathbf{R}$ and $\mathbf{A}$ such that the following optimization problem is solved

$$\min_{\mathbf{R},\mathbf{A}} ||\mathbf{X} - \mathbf{RA}||_F^2 , \tag{B.20}$$

where $||\cdot||_F$ represents the Frobenius norm[12], and the term $||\mathbf{X} - \mathbf{RA}||_F^2$ the overall Mean Square Error (MSE)

$$||\boldsymbol{E}||_F^2 = ||\mathbf{X} - \mathbf{RA}||_F^2 .$$

To restrict the search space, some constraints can be imposed. As we are looking for sparse decomposition of $\mathbf{X}$ over $\mathbf{R}$, the most appropriate constraint is on the number of nonzero coefficients in the $\mathbf{A}$ matrix column vectors $\boldsymbol{a}_i$:

$$\min_{\mathbf{R},\mathbf{A}} ||\mathbf{X} - \mathbf{RA}||_F^2 , \text{ subject to } ||\boldsymbol{a}_i||_0 \leq n, \forall i. \tag{B.21}$$

Recently, the K-SVD approach to dictionary learning was introduced in [Aharon 2006] to solve problem (B.21). K-SVD is an iterative algorithm alternating the two following steps:

- a *sparse coding* (sparse decomposition) step: computing the best $\mathbf{A}$, with $\mathbf{R}$ fixed;
- a *dictionary update* step: searching for the best dictionary $\mathbf{R}$, with $\mathbf{A}$ fixed.

In the sparse coding step, the problem we seek to solve is

$$\min_{\mathbf{A}} ||\mathbf{X} - \mathbf{RA}||_F^2 , \text{ subject to } ||\boldsymbol{a}_i||_0 \leq n, \forall i. \tag{B.22}$$

For each one of the $K$ columns in $\mathbf{A}$, problem (B.22) can be written as

$$\min_{\boldsymbol{a}_i} ||\mathbf{x}_i - \mathbf{R}\boldsymbol{a}_i||_2^2 , \text{ subject to } ||\boldsymbol{a}_i||_0 \leq n. \tag{B.23}$$

Finding the sparsest representation for the problem (B.23) is a combinatorial problem [Davis 1997]. The authors propose to resort to approximate solutions by using for instance a well known improvement of the MP algorithm (see section 6.2) that is its *orthogonal* version. Differently from the MP algorithm, in the Orthogonal Matching Pursuit (OMP) [Pati 1993] after every step the sparse vector of coefficients is updated (a new atom of the dictionary is chosen) on the criterion of best MSE signal approximation of the residual, computed as the orthogonal projection of the signal onto the set of atoms selected so far. The OMP algorithm generally provides better results than the MP, but requires more computation.

After the sparse coding step, dictionary updating is performed. At $\mathbf{A}$ fixed, the $P$ dictionary atoms $\boldsymbol{r}_i$ are updated one by one freezing, at each time, all columns in $\mathbf{R}$ except one, say $\boldsymbol{r}_k$. Each updated column $\tilde{\boldsymbol{r}}_k$ is found jointly to its corresponding coefficients (the $k-$th row in $\mathbf{A}$) using a rank-1 approximation.

[12] Given a matrix $\boldsymbol{A}$ of dimensions $(M \times N)$, the Frobenius norm is equal to $||\boldsymbol{A}||_F = \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N} a_{ij}^2}$.

## B.4 Asymptotic approximation of $T_{\mathbf{LRMAP}}$

An asymptotic approximation of $T_{\mathrm{LRMAP}}$ under $\mathcal{H}_0$, for $N \to \infty$, is proposed. A first condition must be set on the $\eta$ threshold. In order to avoid the accumulation of *"only noise"* components in (3.5), it is necessary to choose $\eta = \sqrt{2 \log N a_N}$, where $a_N$ (e.g. $a_N = 1/\log(N)^2$) will enable to conserve only the *"signal"* components.

The asymptotic normality of the LRMAP test statistic under the null hypothesis can be proved using the central limit theorem by verifying the Lyapunov's condition[13]. As in [Fan 1996], the $T_{\mathrm{LRMAP}}$ $k-$th moments involve

$$\mathbb{E}[(u^2 - \eta^2) I(|u| > \eta)]^k \tag{B.24}$$

and can be computed through the integration by parts of

$$\int_{\eta}^{\infty} u^{2k} \exp^{-\frac{u^2}{2}} \mathrm{d}u \tag{B.25}$$

followed by an approximation for $\eta \to \infty$. For a given $P_{FA}$, we obtain the asymptotic LRMAP test:

$$\frac{1}{\sigma_{\mathrm{LRMAP}}} (T_{\mathrm{LRMAP}}(\boldsymbol{u}) - \mu_{\mathrm{LRMAP}}) \underset{H_0}{\overset{H_1}{\gtrless}} \Phi^{-1}(1 - P_{FA}),$$

with

$$\mu_{\mathrm{LRMAP}} = \sqrt{\frac{2}{\pi}} \exp^{-\eta^2/2} \left( \frac{2}{\eta} - \frac{4}{\eta^3} + O(\eta^{-5}) \right), \tag{B.26}$$

and, after similar calculations,

$$\sigma^2_{\mathrm{LRMAP}} = \sqrt{\frac{2}{\pi}} \exp^{-\eta^2/2} \left( \frac{8}{\eta} - \frac{24}{\eta^3} + O(\eta^{-5}) \right). \tag{B.27}$$

[13] Lyapunov's Central Limit Theorem: for a sequence of independent random variables $X_i$ with $i = 1, \ldots, N$, each with finite expected value $\mu_i$ and variance $\sigma_i^2$, define $s_N^2 = \sum_{i=1}^{N} \sigma_i^2$ the sum of variances. If for some $\delta > 0$, the Lyapunov's condition $\lim_{N \to \infty} 1/s_N^{2+\delta} \sum_{i=1}^{N} \mathbb{E}[|X_i - \mu_i|^2 + \delta] = 0$ is satisfied, then a sum of $(X_i - \mu_i)/s_n$ converges in distribution to a standard normal random variable, as $n$ goes to infinity: $1/s_n \sum_{i=1}^{N} (X_i - \mu_i) \xrightarrow{d} \mathcal{N}(0, 1)$. For simplicity, $\delta$ is usually chosen equal to 1.

## B.5 A tentative of comparison of the tests by analysis of the $P_{FA}$ and $P_{DET}$ bounds

We seek to compare the performances of the two tests by evaluating, for comparable values of the $P_{FA}$, the corresponding values of the $P_{DET} = Pr(T(\boldsymbol{u}, \eta) > \gamma \mid \mathcal{H}_1)$.

The dependency of the $P_{FA}$ and the $P_{DET}$ with respect to the ($\eta$,$\gamma$) parameters make their computation inextricable, even in the case of the simple model (3.1). Hence, based on the expressions (3.48) and (3.49) and using suitable maximizing-minimizing analytical functions, we propose an approximation aiming at bounding the $P_{FA}$ (resp. $P_{DET}$) in a range $P_{FAm} < P_{FA} < P_{FAM}$ (resp. $P_{DETm} < P_{DET} < P_{DETM}$). For both tests, the idea is to compare the $P_{FA}$ and the $P_{DET}$ bounds to assess relative power.

Figure B.1 illustrates the computation of the $P_{FA}$ maximizing-minimizing surfaces for the PDR (left) and the LRMAP (right) tests for $N = 2$. In particular, Figure B.1(a) shows the computation of the $P_{FA}$ upper and the lower *bounds*, respectively $P_{FAM}$ and $P_{FAm}$. Here, the $P_{FAM}$ upper bound corresponds to the $P_{FA}$ of the closest surface to the tests' detection surfaces $\mathcal{S}$ (the one that gives minimum $P_{FA}$ value) chosen between the inscribed circle (in green, dashed line) and the inscribed square (in green, continuous line). We thus have that,

$$P_{FAM} = \min(P_{FA\circ_i}, P_{FA\square_i}), \tag{B.28}$$

where

$$P_{FA\circ_i} = 1 - Pr\left(\sum_{j=1}^{N} u_j^2 < r_i^2\right) = 1 - \Phi_{\chi^2_N}(r_i^2), \tag{B.29}$$

with $r_i$ the radius of the circle inscribed in $\mathcal{S}$ under $\mathcal{H}_0$, and

$$P_{FA\square_i} = 1 - Pr(|u_j| < l_i \ \forall j) = 1 - (2\Phi(l_i) - 1)^N, \tag{B.30}$$

with $l_i$ the side of the square inscribed in $\mathcal{S}$ under $\mathcal{H}_0$. $\Phi_{\chi^2_N}(\cdot)$ and $\Phi(\cdot)$ represent the CDFs of a centered $\chi^2$ with $N$ degrees of freedom and of a standard normal distribution, respectively. Similarly, the $P_{FAm}$ lower bound corresponds to the $P_{FA}$ of the closest surface to the tests' detection surfaces $\mathcal{S}$ (the one that gives maximum $P_{FA}$ value) chosen between the circumscribed circle (in orange, dashed line) and the circumscribed square (in orange, continuous line). We thus have that

$$P_{FAm} = \max(P_{FA\circ_c}, P_{FA\square_c}), \tag{B.31}$$

where $P_{FA\circ_c}$ and $P_{FA\square_c}$ are computed considering the radius/side of the circle/square circumscribed to $\mathcal{S}$ under $\mathcal{H}_0$.

An equivalent reasoning yields, for the bounds of the $P_{DET}$:

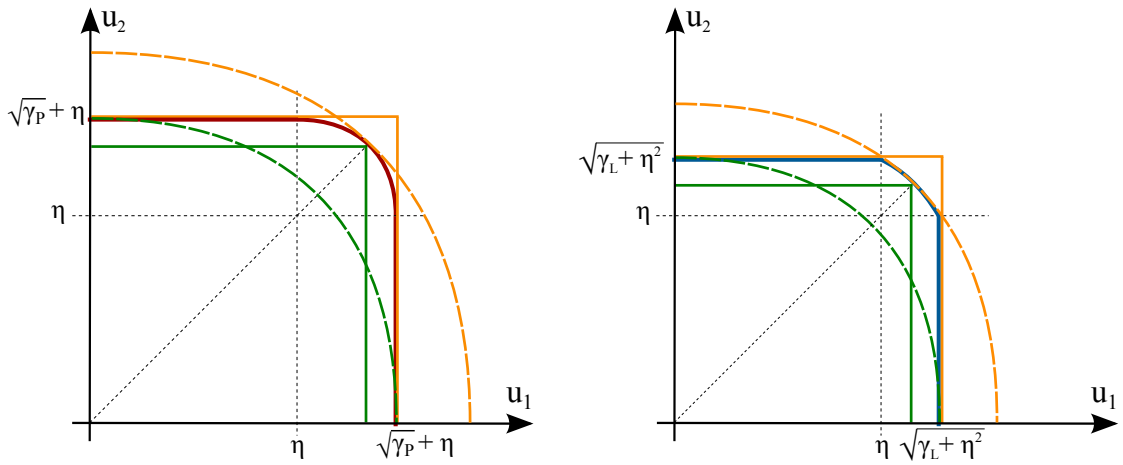
$$P_{DETM} = \min(P_{DET\circ_i}, P_{DET\square_i}), \tag{B.32}$$

where

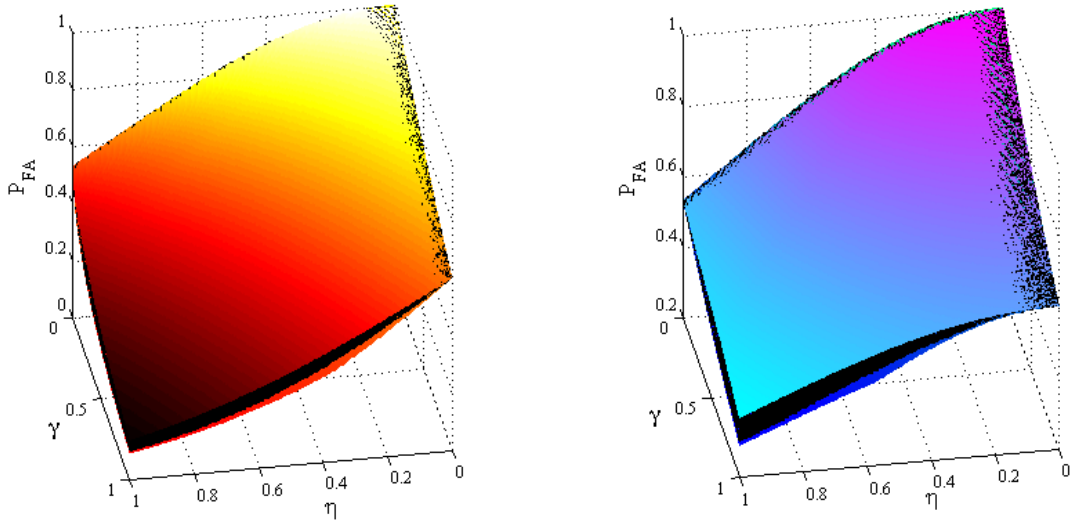
$$P_{DET\circ i} = 1 - \Phi_{\chi^2_N(\lambda)}(r_i^2) \tag{B.33}$$

and

$$P_{DET\square i} = 1 - \prod_{j=1}^{N} [\Phi(l_i - \theta_j) + \Phi(l_i + \theta_j) - 1], \tag{B.34}$$




(a)




(b)

Figure B.1: Figure B.1(a) Maximizing (orange) and minimizing (green) surfaces of the PDR (left) and the LRMAP (right) tests, for $N = 2$. Figure B.1(b) Analytical $P_{FA_m}$ and $P_{FA_M}$ approximating bounds plotted as functions of the ($\eta$, $\gamma$) parameters, for the PDR (red) and the LRMAP (blue) tests (see text). Empirical values of the PDR and LRMAP tests $P_{FA}$ computed as functions of the ($\eta$,$\gamma$) parameters (in black for both tests) follow within the tight $P_{FA}$ envelopes defined by the approximating *bounds*. This confirms the validity of the approach.

and for the minimum $P_{DET}$:

$$P_{DETm} = \max(P_{DET\circ_c}, P_{DET\square_c}), \tag{B.35}$$

with $P_{DET\circ_c}$ and $P_{DET\square_c}$ computed en considering the radius/side of the circle/square

circumscribed to $\mathcal{S}$ under $\mathcal{H}_1$.

Figure B.1(b) reports the $\mathrm{P}_{\mathrm{FA}m}$ and $\mathrm{P}_{\mathrm{FA}M}$ bounds of the PDR (left) and the LRMAP (right) tests in function of the $\eta$ and $\gamma$ parameters, for $N = 2$. The true $\mathrm{P}_{\mathrm{FA}}$ values obtained by MC simulations are represented in black. These values are obtained computing the tests' $\mathrm{P}_{\mathrm{FA}}$ for different couples $(\eta, \gamma)$, over $10^5$ noise realizations. The empirical $\mathrm{P}_{\mathrm{FA}}$ values remain within the *narrow* envelope defined by the approximating *bounds*. This confirm the validity of the approach and opens to the possibility of analytically refining the comparative results between the two tests. The study of the $\mathrm{P}_{\mathrm{DET}}$ envelopes and the analytical comparison of the tests are currently under investigation.

# Bibliography


[Aharon 2006] M. Aharon, M. Elad and A. M. Bruckstein. *K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation.* IEEE Trans. Signal Processing, vol. 54, no. 11, pages 4311–4322, November 2006. (Cited on pages 36, 112, 163 and 164.)

[Antoniadis 1997] A. Antoniadis. *Wavelets in Statistics: A Review (with discussion).* Italian Journal of Statistics, vol. 6, pages 97–144, 1997. (Cited on page 37.)

[Antoniadis 2001] A. Antoniadis and J. Fan. *Regularization of Wavelet Approximations.* Journal of the American Statistical Association, vol. 96, no. 455, pages 939–967, 2001. (Cited on pages 36, 37, 38, 39, 45, 67 and 161.)

[Arias-Castro 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, no. 5, pages 2533–2556, 2010. (Cited on pages 6, 48, 81, 85, 90, 91, 95, 111 and 150.)

[Bajwa 2013] W. U. Bajwa and A. Pezeshki. *Finite Frames for Sparse Signal Processing.* Finite Frames: Theory and Applications, 2013. (Cited on page 48.)

[Barrett 1998] H.H. Barrett, C.K. Abbey and E. Clarkson. *Objective assessment of image quality. III. ROC metrics, ideal observers, and likelihood-generating functions.* Journal of the Optical Society of America A: Optics, Image Science, and Vision, vol. 15, no. 6, pages 1520–35, June 1998. (Cited on page 81.)

[Basu 1996] S. Basu. *Bayesian hypotheses testing using posterior density function.* Statistics and Probability Letters, vol. 30, no. 1, pages 79–86, 1996. (Cited on pages 47, 61 and 63.)

[Bourguignon 2011] S. Bourguignon, D. Mary and E. Slézak. *Restoration of Astrophysical spectra with sparsity constraints: models and algorithms.* Selected Topics in Signal Processing, IEEE Journal of, vol. 5, pages 1002 – 1013, September 2011. (Cited on pages 36, 59, 99, 120 and 122.)

[Bourguignon 2012] S. Bourguignon, D. Mary and E. Slézak. *Processing MUSE hyperspectral data: Denoising, deconvolution and detection of astrophysical sources.* Statistical Methodology, vol. 9, pages 32–43, 2012. (Cited on page 53.)

[Brand 1995] J. C. D. Brand. Lines of light: The sources of dispersive spectroscopy, 1800 - 1930. (pp. 37-42). Gordon and Breach Publishers, 1995. (Cited on page 116.)

[Chan 2012] K. Chan, J. Li, W. Eichinger and E. Bai. *A new physics-based method for detecting weak nuclear signals via spectral decomposition.* Nuclear Instruments and Methods in Physics Research A, vol. 667, pages 16–25, 2012. (Cited on page 109.)

[Chen 2001] S.S. Chen, D.L. Donoho and M.A. Saunders. *Atomic decomposition by basis pursuit.* SIAM Review, vol. 43, no. 1, pages 129–159, 2001. (Cited on pages 48 and 88.)

[Chen 2011a] Y. Chen, N. Nasrabadi and T. D. Tran. *Simultaneous Joint Sparsity Model for Target Detection in Hyperspectral Imagery.* IEEE Geoscience and Remote Sensing Letters, vol. 8, pages 676–680, 2011. (Cited on page 53.)

[Chen 2011b] Y. Chen, N. Nasrabadi and T. D. Tran. *Sparse Representation for Target Detection in Hyperspectral Imagery.* IEEE Selected Topics in Signal Processing, vol. 5, no. 3, pages 629–640, 2011. (Cited on pages 52 and 53.)

[Davis 1997] G. Davis, S. Mallat and M. Avellaneda. *Adaptive greedy approximations.* Journal of Constructive Approximation, vol. 13, pages 57–98, 1997. (Cited on page 164.)

[Donoho 1994] D. L. Donoho and I. M. Johnstone. *Ideal Spatial Adaptation by Wavelet Shrinkage.* Biometrika, vol. 81, no. 3, pages 425–455, 1994. (Cited on pages 37, 39 and 47.)

[Donoho 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 6, 48, 49, 85, 90, 91 and 95.)

[Doppler 1842] C. J. Doppler. *Über das farbige Licht der Doppelsterne und einiger anderer Gestirne des Himmels (About the coloured light of the binary stars and some other stars of the heavens).* Proceedings of the Royal Bohemian Society of Sciences, vol. 2, 1842. (Cited on page 139.)

[Elad 2010] M. Elad. Sparse and redundant representations: From theory to applications in signal and image processing. Springer, 2010. (Cited on page 34.)

[Fan 1996] J. Fan. *Test of Significance Based on Wavelet Thresholding and Neyman's Truncation.* Journal of the American Statistical Association, vol. 91, no. 434, pages 674–688, June 1996. (Cited on pages 47, 63, 69, 70 and 165.)

[Fan 1997] J. Fan. *Comment on 'Wavelets in Statistics: A Review' by A. Antoniadis.* Italian Journal of Statistics, vol. 6, pages 97–144, 1997. (Cited on page 41.)

[Fan 2001] J. Fan and R. Lie. *Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties.* Journal of the American statistical Association, vol. 96, pages 1348–1360, 2001. (Cited on page 41.)

[Fuchs 2004] J.J. Fuchs. *On sparse representations in arbitrary redundant bases.* IEEE Transactions on Information Theory, vol. 50, no. 6, pages 1341 – 1344, June 2004. (Cited on pages 48 and 88.)

[Fuchs 2010] J. J. Fuchs. *The Generalized Likelihood Ratio Test and the Sparse Representation Approach.* In ICISP'10 Proceedings of the 4th International Conference on Image and Signal Processing, pages 245–253, 2010. (Cited on pages 48, 61 and 90.)

[Good 1965] I. J. Good. *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. MIT Press, Cambridge, MA, 1965. (Cited on pages 47 and 61.)

[Haupt 2007] J. Haupt and R. Nowak. *Compressive sampling for signal detection*. In Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on, volume 3, pages 1509–1512, 2007. (Cited on pages 48 and 52.)

[Herranz 2010] D. Herranz, F. Argüeso, J. L. Sanz and M. Lopez-Caniego. *A sparse detection approach to astronomical point source detection*. In 18th European Signal Processing Conference (EUSIPCO), pages 139–143, 2010. (Cited on page 53.)

[Herschel 1800] W. Herschel. *Experiments on the Refrangibility of the Invisible Rays of the Sun*. Philosophical Transactions of the Royal Society of London, no. 90, pages 284–292, 1800. (Cited on page 116.)

[Huygens 1690] C. Huygens. Traité de la lumière. 1690. (Cited on page 116.)

[Jeffreys 1935] H. Jeffreys. *Some Tests of Significance, Treated by the Theory of Probability*. Proceedings of the Cambridge Philosophy Society, vol. 31, pages 203–222, 1935. (Cited on page 31.)

[Kay 1998a] S. M. Kay. Fundamentals of statistical signal processing: Detection theory, volume 2. Prentice-Hall PTR, 1998. (Cited on pages 14, 32 and 59.)

[Kay 1998b] S. M. Kay. Fundamentals of statistical signal processing: Estimation theory, volume 1. Prentice-Hall PTR, 1998. (Cited on pages 41, 43 and 155.)

[Kern 1948] W. F. Kern and J. R. Bland. Solid mensuration with proofs. New York: Wiley, 1948. (Cited on page 158.)

[Kosmalski 2011] J. Kosmalski, L. Parès and al. *Optical design of the VLT/MUSE instrument*. In Proceedings of the SPIE, volume 8167, 2011. (Cited on page 119.)

[Kraut 2001] S. Kraut, L. L. Scharf and L. T. McWhorter. *Adaptive Subspace Detectors*. IEEE Transactions on Signal Processing, vol. 49, no. 1, pages 1–16, 2001. (Cited on page 52.)

[Larsson 2007] E. G. Larsson and Y. Selen. *Linear Regression With a Sparse Parameter Vector*. Signal Processing, IEEE Transactions on, vol. 55, no. 2, pages 451–460, 2007. (Cited on page 48.)

[Lehmann 2005] E.L. Lehmann and J.P. Romano. Testing statistical hypotheses. Springer, 2005. (Cited on pages 36 and 112.)

[Mallat 1993] S. Mallat and Z. Zhang. *Matching pursuits with time-frequency dictionaries*. IEEE Trans. Signal Processing, vol. 41, no. 12, pages 3397–3415, 1993. (Cited on page 99.)

[Mallat 2008] S. Mallat. A wavelet tour of signal processing : the sparse way (3rd edition). Academic Press, 2008. (Cited on pages 34 and 35.)

[Malloy 2011] M. Malloy and R. Nowak. *Sequential Analysis in High Dimensional Multiple Testing and Sparse Recovery.* In Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on, pages 2661–2665, 2011. (Cited on page 48.)

[Manolakis 2001] D. Manolakis, C. Siracusa and G. A. Shaw. *Adaptive matched subspace detectors for hyperspectral imaging applications.* In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), volume 5, pages 3153–3156, 2001. (Cited on page 52.)

[Manolakis 2002] D. Manolakis and G. A. Shaw. *Detection algorithms for hyperspectral imaging applications.* IEEE Signal Processing Magazine, vol. 19, no. 1, pages 29–43, 2002. (Cited on page 51.)

[Manolakis 2003] D. Manolakis, D. Marden and G. A. Shaw. *Hyperspectral Image Processing for Automatic Target Detection Applications.* Lincoln Lab. J., vol. 14, no. 1, pages 79–116, 2003. (Cited on page 115.)

[Manolakis 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 page 112.)

[Moulin 1999] P. Moulin and J. Liu. *Analysis of multiresolution image denoising schemes using generalized Gaussian and complexity priors.* IEEE Transactions on Information Theory, vol. 45, no. 3, pages 909–919, April 1999. (Cited on pages 36, 37, 41, 45 and 67.)

[Newton 1672] I. Newton. *Where Light is declared to be not Similar.* Philosophical Transactions of the Royal Society, no. 80, pages 3075–3087, 1672. (Cited on page 116.)

[Nikolova 2000] M. Nikolova. *Local strong homogeneity of a regularized estimator.* SIAM Journal on Applied Mathematics, vol. 61, no. 2, pages 633–658, August 2000. (Cited on pages 36 and 37.)

[Paredes 2009] J. Paredes, Z. Wang, G. R. Arce and B. M. Sadler. *Compressive matched subspace detection.* In 17th European Signal Processing Conference (EUSIPCO 2009), pages 120–124, 2009. (Cited on page 48.)

[Paris 2011a] S. Paris, D. Mary and A. Ferrari. *Composite Hypothesis Tests for Sparse Parameters.* In Statistical Signal Processing Workshop (SSP), 2011 IEEE, pages 737–740, 2011. (Cited on page 4.)

[Paris 2011b] S. Paris, D. Mary and A. Ferrari. *PDR and LRMAP detection tests applied to massive hyperspectral data.* In Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2011 4th IEEE International Workshop on, pages 93–96, 2011. (Cited on page 5.)

[Paris 2011c] S. Paris, D. Mary and A. Ferrari. *Sparsity-based composite detection tests. Application to astrophysical hyperspectral data.* In 19th European Signal Processing Conference (EUSIPCO 2011), pages 1909–1913, 2011. (Cited on pages 4 and 5.)

[Paris 2011d] S. Paris, D. Mary and A. Ferrari. *Sur des tests d'hypothèses composites pour des paramètres parcimonieux*. In XXIII Colloque GRETSI, 2011. (Cited on page 5.)

[Paris 2012] S. Paris, D. Mary and A. Ferrari. *MAP-based sparse detection strategies. Application to the hyperspectral data of the MUSE instrument*. In Conference on Astronomical Data Analysis, (ADA 7), 2012. (Cited on page 5.)

[Paris 2013a] S. Paris, D. Mary and A. Ferrari. *Detection tests using sparse models, with application to hyperspectral data*. IEEE Trans. Signal Processing, vol. 61, no. 6, pages 1481–1494, 2013. (Cited on page 5.)

[Paris 2013b] S. Paris, R.F.R. Suleiman, D. Mary and A. Ferrari. *Constrained Likelihood Ratios for detecting sparse signals in highly noisy 3D data*. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), page Accepted, 2013. (Cited on page 5.)

[Pati 1993] Y.C. Pati, R. Rezaiifar and P. S. Krishnaprasad. *Orthogonal Matching Pursuit: Recursive Function Approximation with Applications to Wavelet Decomposition*. In Proceedings of the 27 th Annual Asilomar Conference on Signals, Systems, and Computers, 1993. (Cited on page 164.)

[Pesquet 1996] J.-C. Pesquet. *Bayesian approach to best basis selection*. In IEEE International Conference on In Acoustics, Speech and Signal Processing (ICASSP), pages 2634–2637, 1996. (Cited on page 45.)

[Ramos 2012] A. Asensio Ramos and R. Manso Sainz. *Signal detection for spectroscopy and polarimetry*. Astronomy & Astrophysics, vol. 547, 2012. (Cited on page 109.)

[Robey 1992] F. C. Robey, D. R. Fuhrmann, E. J. Kelly and R. Nitzberg. *A CFAR adaptive matched filter detector*. IEEE Transactions on Aerospace and Electronic Systems, vol. 28, no. 1, pages 208–216, 1992. (Cited on page 51.)

[Scharf 1994] L. L. Scharf and B. Friedlander. *Matched Subspace Detectors*. IEEE Transactions on Signal Processing, vol. 42, no. 8, pages 2146–2157, 1994. (Cited on pages 51 and 52.)

[Schweizer 2001] S. Schweizer and J.F. Moura. *Efficient detection in hyperspectral imagery*. IEEE Transactions on Image Processing, vol. 10, no. 4, pages 584–597, April 2001. (Cited on page 53.)

[Skolnik 1980] M. I. Skolnik. Introduction to radar systems. McGraw-Hill, 1980. (Cited on page 3.)

[Starck 2009] J.L. Starck, J.M. Fadili, S. Digel, B. Zhang and J.Chiang. *Source detection using a 3D sparse representation: application to the Fermi gamma-ray space telescope*. Astronomy & Astrophysics, vol. 504, no. 2, pages 641–652, 2009. (Cited on pages 109 and 115.)

[Starck 2010] J.L. Starck, F. Murtagh and J. Fadili. Sparse image and signal processing: Wavelets, curvelets, morphological diversity. Cambridge University Press, 2010. (Cited on pages 34 and 36.)

[Suleiman 2013] R.F.R. Suleiman, D. Mary and A. Ferrari. *Minimax sparse detection based on one-class classifiers*. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013. (Cited on pages 36, 112, 140 and 153.)

[Theiler 2012] J. Theiler and B. Wohlberg. *Detection of spectrally sparse anomalies in hyperspectral imagery*. In Proc. IEEE Southwest Symposium on Image Analysis and Interpretation, pages 117–120, 2012. (Cited on page 52.)

[Tibshirani 1996] R. Tibshirani. *Regression Shrinkage and Selection via the Lasso*. Journal of the Royal Statistical Society. Series B (Methodological), vol. 58, no. 1, pages 267–288, 1996. (Cited on pages 48 and 88.)

[Tipping 2012] M.E. Tipping. *The Relevance Vector Machine*. In S. A. Solla, T. K. Leen and K.-R. Müller (Eds.), editeurs, Advances in Neural Information Processing Systems 12, pages 652–658. MIT Press, 2012. (Cited on page 109.)

[Trees 1968] H. L. Van Trees. Detection, estimation and modulation theory, volume 1. John Wiley and Sons, New York, 1968. (Cited on page 51.)

[Tukey 1976] J. W. Tukey. *The Higher Criticism. Course Notes, Statistics 411*. Rapport technique, Princeton University, 1976. (Cited on pages 48, 49, 85 and 95.)

[Vo-Dinh 2004] T. Vo-Dinh and al. *A hyperspectral imaging system for in vivo optical diagnostics*. In Engineering in Medicine and Biology Magazine, IEEE, volume 23, pages 40–49, 2004. (Cited on page 115.)

[Wang 2008] Z. Wang, G. R. Arce and B. M. Sadler. *Subspace Compressive detection for sparse signals*. In Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on, pages 3873–3876, 2008. (Cited on page 48.)

[Williams 1996] R. E. Williams, B. Blacker and al. *The Hubble Deep Field: Observations, Data Reduction, and Galaxy Photometry*. Astronomical Journal, vol. 112, page 1335, 1996. (Cited on page 119.)

[Young 1802] T. Young. *Bakerian Lecture: On the Theory of Light and Colours*. Philosophical Transactions of the Royal Society of London, no. 92, pages 12–48, 1802. (Cited on page 116.)