Author/s:
Miri Rekavandi, Aref

Title:
From Robust to Efficient Detection and Estimation: Applicable to Brain Activity Detection

Date:
2021

Persistent Link:
https://hdl.handle.net/11343/282492

From Robust to Efficient Detection and Estimation: Applicable to Brain Activity Detection

Aref Miri Rekavandi

ORCID: 0000-0001-9542-759X

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

Department of Electrical and Electronic Engineering
THE UNIVERSITY OF MELBOURNE

September 2021

Abstract

Functional magnetic resonance imaging (fMRI) is a non-invasive imaging technique that has been extensively used in recent years to localize neural activities in the brain. Assuming a general linear model (GLM) to approximate recorded signals, classical maximum likelihood based estimators and detectors are used in the literature and their design is based on strong assumptions (e.g., Gaussianity of the noise). In the first part of this thesis, we extend the existing detectors and estimators using an alternative measure from information geometry called "$\alpha-$divergence". This measure is used to obtain robustness against possible outliers or mismatches in observations and achieve more reliable results. Finally, in the second part, we consider existing knowledge about the data as prior information (e.g., low rank data, locally correlated noise) and develop detectors that are more adapted for such data types.

After validation of our proposed techniques using simulations and synthetic data, we compare the performance of the proposed methods with existing methods in the literature in different imaging applications with the same data modeling (e.g., fMRI data, hyperspectral images, etc.)

Declaration of Authorship

This is to certify that:

- the thesis comprises only my original work towards the PhD,
- due acknowledgement has been made in the text to all other material used; and
- the thesis is less than 100,000 words in length, exclusive of tables, maps, bibliographies and appendices.

Aref Miri Rekavandi, March 2021.

This page intentionally left blank.

Preface

The research work presented in this thesis was primarily conducted by the student. The supervisors also contributed towards the thesis through supervision, technical advice, and manuscript proofreading through regular meetings. The student gratefully appreciates the financial support provided by the University of Melbourne, via Melbourne Research Scholarship (MRS). The outcomes of this thesis are published or under review in the following conferences and journal papers:

- Chapter 3
  - [1] **A. Miri Rekavandi**, A. K. Seghouane, and R. J. Evans, (2021), Robust Subspace Detectors Based on $\alpha-$Divergence with Application to Detection in Imaging, IEEE Transactions on Image Processing, 30, 5017-5031.
  - [2] **A. Miri Rekavandi**, A. K. Seghouane, and R. J. Evans, Robust Likelihood Ratio Test Using $\alpha-$Divergence, 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1150-1154.
- Chapter 4
  - [3] **A. Miri Rekavandi**, A. K. Seghouane, and R. J. Evans, Learning Robust and Sparse Principal Components with the $\alpha-$Divergence, to be submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence.
  - [4] **A. Miri Rekavandi**, and A. K. Seghouane, Robust Principal Component Analysis Using Alpha Divergence, 2020 EEE International Conference on Image Processing (ICIP). pp. 6-10.
- Chapter 5

- [5] **A. Miri Rekavandi**, A. K. Seghouane, and R. J. Evans, Adaptive Matched Filter using Non-Target Free Training Data, 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1090-1094.

- Chapter 6

  - [6] **A. Miri Rekavandi**, A. K. Seghouane, and R. J. Evans, Adaptive Subspace Detector in High Dimensional Space with Insufficient Training Data, 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1164-1168.

- Chapter 7

  - [7] **A. Miri Rekavandi**, A. K. Seghouane, and R. J. Evans, Adaptive Brain Activity Detection in Structured Interference and Partially Homogeneous Locally Correlated Disturbance, to be submitted to IEEE Transactions on Medical Imaging.

This page intentionally left blank.

Acknowledgements

I would like to express my gratitude to my supervisors Dr. Abd-Krim Seghouane and Prof. Robin J. Evans, for their continuous support, encouragement and constructive technical and non technical feedback that helped me to finish this journey on track and in a right direction. I also want to thank my PhD committee chair, A/Prof. Marcus Brazil for his interest and effort.

At last but not least, I would like to thank my parents Shaban and Rahimeh, my lovely wife Shima, my brother Shoeib, my sister Zahra and all my friends Masoud, Ehsan, Salman, Saeed, Navid, Amir, Jeanne, Marc, Asif, Ali, Seyed Mostafa, Masoud, Shadi, Afshin for supporting and helping me to overcome the challenges in both my normal and academic life. Truly, finishing this PhD was not possible without their encouragement.

This page intentionally left blank.

To my parents, my wife and my siblings.

This page intentionally left blank.

Contents

List of Figures

List of Tables

Abbreviations

fMRI	Functional Magnetic Resonance Imaging
GLM	General Linear Model
SAR	Synthetic-Aperture Radar
BOLD	Blood Oxygen Level Dependent
EEG	Electroencephalography
MEG	Magnetoencephalography
IFFT	Inverse Fast Fourier Transform
HRF	Hemodynamic Response Function
LTI	Linear Time Invariant
PCA	Principal Component Analysis
ICA	Independent Component Analysis
DL	Dictionary Learning
SNR	Signal to Noise Ratio
LRT	Likelihood Ratio Test
UMP	Uniformly Most Powerful
GLRT	Generalized Likelihood Ratio Test
MLE	Maximum Likelihood Estimator/Estimate
MSD	Matched Subspace Detector
ASD	Adaptive Subspace Detector
AMF	Adaptive Matched Filter
CRLO	Cramér–Rao Lower Bound
LFD	Least Favorable Distribution
BP	Breakdown Point
IF	Influence Function
SC	Sensitivity Curve

MBC	Maximum Bias Curve
KL	Kullback-Leibler
ML	Maximum Likelihood
CFAR	Constant False Alarm Rate
ROC	Receiver Operating Characteristic
HYDICE	Hyperspectral Digital Imagery Collection Experiment
RCS	Radar Cross Section
GMM	Gaussian Mixture Model
RPCA	Robust Principal Component Analysis
PC	Principal Component
FB	Foreground-Background
AI	Artificial Intelligence
PPCA	Probabilistic Principal Component Analysis
CCA	Canonical Correlation Analysis
AC	Adjacent Cell
CUT	Cell Under the Test
SVD	Singular Value Decomposition
ARSD	Adaptive Reduced Subspace Detector
ACE	Adaptive Coherence Estimator
BPRFT	Block-Paradigm Right Finger Tapping
ERRFT	Event Related Right Finger Tapping

Notation Summary

a, $\mathbf{a}$	Scalar and vector values
$\mathbf{A}$, $\mathbf{A}^{i:j}$, $a_{i,j}$	Matrix, matrix including row i to j of matrix $\mathbf{A}$, entry on the $i-$th row and $j-$th column of matrix $\mathbf{A}$
$*$, $\lvert.\rvert$	Convolution operator, absolute value
$\langle.\rangle$,$\langle.\rangle^{\perp}$	Subspace spanned by columns of input matrix, the orthogonal subspace
$f : \mathbf{D} \to \mathbf{Y}$	Mapping function that maps from the domain $\mathbf{D}$ to $\mathbf{Y}$
$\mathbf{E}_p[.]$	Expectation with respect to the distribution p
$\mathcal{N}(\boldsymbol{\mu}, \mathbf{R})$	Multivariate normal density with the mean $\boldsymbol{\mu}$ and the covariance $\mathbf{R}$
$\mathcal{CN}(\boldsymbol{\mu}, \mathbf{R})$	Multivariate complex normal distribution with the mean $\boldsymbol{\mu}$ and the covariance $\mathbf{R}$
$\mathbf{I}_N$, $\mathbf{1}_N$	$N \times N$ identity matrix and a vector with elements of 1 with size N
$\mathbf{I}$	Fisher Information Matrix
$(.)^{\top}$, $(.)^{*}$, $(.)^{\dagger}$, $(.)^{-1}$	Transpose, complex conjugate, complex conjugate transpose and inverse
$\to_d$, $\to_p$	Distributed as, convergence in probability
Var(.), tr(.), det(.)	Variance, trace of a matrix, determinant of a matrix
diag(.), eig(.), SVD(.)	Diagonalization function, eigenvalues, singular value decomposition
$\mathbf{P}_D$,$\mathbf{P}_{D^{\perp}}$	Orthogonal projection on the subspace spanned by columns of $\mathbf{D}$, orthogonal projection on the orthogonal subspace to $\langle\mathbf{D}\rangle$
log(.)	Natural logarithm
$\frac{\partial}{\partial \mathbf{x}}$, $\nabla \mathbf{x}$	Partial derivative with respect to $\mathbf{x}$, Gradient in the direction of $\mathbf{x}$
$\chi^2_p(\lambda)$	Noncentral Chi-squar distribution with p degrees of freedom and non-centrality parameter λ
$\lVert.\rVert_2$, $\lVert.\rVert_1$, $\lVert.\rVert_F$, $\lVert.\rVert_*$	ℓ_2^1 norm, ℓ_1 norm, Frobenius norm and nuclear norm
$D(.\lVert.)$, $(.)_+$	Divergence between two arguments, positive part of the argument

Chapter 1

Introduction

Detection and estimation theory have wide applications in different areas such as target or object detection from recorded noisy signals from a radar [11], hyperspectral imaging devices [12] and synthetic-aperture radar (SAR)[13]. It is also useful for symbol detection in the receiver side of a noisy communication channel [14, 15], and activity detection in various biomedical images [16, 17]. All mentioned applications deal with sets of recorded and corrupted signals which need statistical tools for better understanding of their underlaying mechanisms and for making inferences. Throughout this thesis, the main objective is to develop and propose estimation and detection methods which achieve a better interpretation from the provided data. In this thesis, when talking data, this refers to functional magnetic resonance imaging (fMRI) data, unless specified otherwise. However, the proposed methods are not limited to this modality and can be applied to any type of data that follow the same linear model. To prove this claim, anywhere possible in this thesis, the superiority of the proposed methods in other applications are demonstrated. However, the studied applications are mostly related to detection and estimation in imaging systems.

Prior to studying existing tools and techniques for analyzing fMRI data, blood oxygen level dependent (BOLD) fMRI data is described. BOLD fMRI is one of the most popular techniques of acquiring fMRI data [18]. This information can help to have a better understating of origin and nature of fMRI data.

1.1 BOLD fMRI

Functional MRI or especially BOLD fMRI is a noninvasive technique for studying local changes in deoxyhemoglobin concentration in the brain by measuring the blood oxygenation level dependent (BOLD) contrast. It is an indirect method of measuring neural activity [19]. The oxygenated and deoxygenated hemoglobin have different magnetic properties. So, acquiring MR images during a task and evaluating them, helps to find areas of the brain which show neural activation during experiments [8].

In recent years, this modality has drawn researcher's attention to study and discover complex functionality of the brain parts given different tasks. Currently, a number of different techniques (direct/indirect, invasive/non-invasive, electrical/magnetic field based) are developed to acquire medical data that enable researchers to look at the brain in different temporal and spatial resolutions [20]. For example, electroencephalography (EEG) and magnetoencephalography (MEG) are among of those techniques which provide high temporal resolution (in the order of a few milliseconds) but low spatial resolution. In contrast, fMRI is among one of the high spatial and moderate temporal resolution techniques [8].

In order to construct an MR image, the subject is placed inside the magnetic field of a large electromagnet. Typically, the data over a slice in the brain can be acquired by radiating a radio frequency pulse in different frequencies by uniformly or non uniformly sweeping a 2D space (k space) and measuring the induced current in receiver coil [21]. To obtain the image form of the MR signal, an inverse fast Fourier transform (IFFT) is needed to be applied on the acquired data. To have a 3D image of the brain, it is necessary to repeat the same steps for all slices over the brain. On average, the whole process of recording a brain fully, with a high enough resolution (for example $64 \times 64 \times 30$), takes approximately 2 seconds [8]. That is the reason which causes the fMRI data to have a roughly low temporal resolution. Finally, to have a full study on the data, a few minutes recording of the full brain is required to obtain a reliable result.

Generally, the fMRI data is recorded in two scenarios. First, *task-related fMRI* (t-fMRI) in which a subject is asked to perform a specific activity like moving right or left fingers to find areas of the brain that are responsible for such activities [22]. Second scenario is the *resting state fMRI* (rs-fMRI) in which the subject is asked to relax and stop doing any physical task [23].

The increased metabolic demands due to neuronal activity cause an increase in the

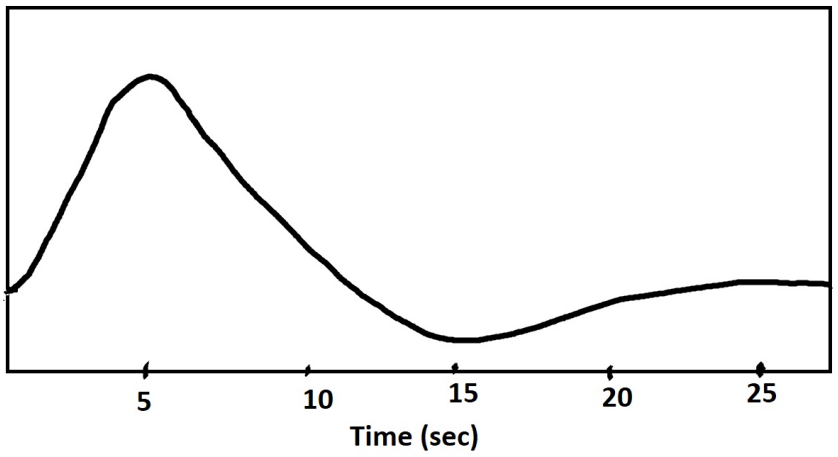

FIGURE 1.1: The standard canonical model for HRF [8]

pumping of oxygenated blood to active regions of the brain. As a result, it is expected to see an increase in the recorded signal after 2 seconds from start of the neural activity. The signal often reaches to its peak value after 5-8 seconds and gradually reaches to its stable point [24]. Therefore, measuring this signal is an indirect indicator of neural activity in the brain. This behavior of neural response is referred to as the hemodynamic response function (HRF) or canonical HRF. Figure 1.1 shows the standard shape of such responses. Note that this response is a reflection to a quick and fast activity. Thus, it can be seen as an impulse response function.

The HRF can be directly used in t-fMRI analysis. For this purpose, it is assumed that each voxel in the brain (small cubic parts of the brain at which the individual fMRI time series are representing their net responses) acts as a linear time invariant (LTI) system [25]. In the t-fMRI the subjects are asked to do the task based on a time scheme. A sampled version of the time scheme can be used as the input of the system or stimulus function $\mathbf{v}$. Finally, the expected response in data ($\mathbf{x}$), can be estimated by convolving the HRF with the stimulus, i.e.,

$$\mathbf{x} = \mathbf{v} * \mathbf{a}, \tag{1.1}$$

where $\mathbf{a}$ is the sampled version of HRF and $*$ stands for the convolution operator. Figure 1.2 shows how the convolution operation can result in the response signal of the voxels in the brain for a specific task given two different time schemes of event-related and block-paradigm. $\mathbf{x}$ can also be estimated using data-driven techniques such as principal component analysis (PCA) [26], independent component analysis (ICA) [27] or dictionary learning (DL) [28, 29] for achieving an accurate model of expected target signal.

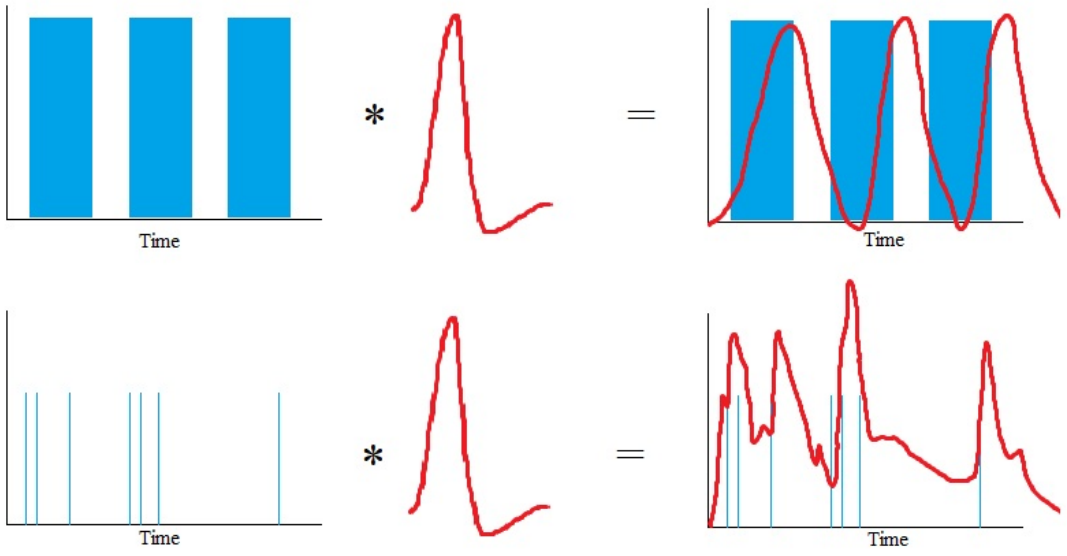

FIGURE 1.2: The BOLD response for Block-Paradigm and Event-Related stimulus functions [8].

Before applying statistical analysis methods on the raw fMRI data, some preprocessing steps are needed to prepare and validate data for further processes. These steps are, but not limited to, the following:

- **Slice timing correction:** During the analysis of fMRI data, it is assumed that measurements of all voxels are recorded simultaneously. However, when looking at the data acquiring process, we notice that recordings are captured sequentially over the slices [30]. This means a time shifting should be applied to all voxels, based on their corresponding slice location in space, to make this assumption valid.

- **Motion correction:** It is very important to acquire data from the same voxel during the experiment. Even a small movement of subject can cause recording data from neighbor voxels which can make the result invalid. So, in this step, it is tried to track head movements and compensate this movements by spatial shifting in 3D space by comparing with a reference image (e.g., mean image) [31].

- **Coregistration:** Typically, fMRI data does not have enough spatial resolution to clearly show the anatomy of the brain. So for the presentation purposes, we map acquired data to a high resolution structural MR image.

- **Normalization:** Different people have different brain sizes and shapes. Therefore, for a group analysis, it is required to map measurements from different subjects to

a standard template brain, e.g., the Talairach or Montreal Neurological Institute (MNI) brain.

- **Smoothing** It is very common in fMRI data analysis to apply a spatial Gaussian kernel for spatial smoothing [8]. In general this step can increase the signal to noise ratio (SNR) and reduce the effect of different anatomies in people for further analysis.

After applying aforementioned prepossessing steps on the raw fMRI data, statistical analysis methods should be applied to localize brain activity which is the main goal of fMRI data analysis through this thesis. But before achieving this goal, first it is needed to model the acquired data. In the next section, the commonly used model for fMRI data analysis is described.

1.2 Data Modelling

In order to localize brain activity, related to a specific task in t-fMRI datasets, a priori knowledge about the subspace of the signals is needed. It was suggested to use the convolution between the HRF and its derivatives with the task stimuli to estimate this subspace [8]. Data driven methods also can be used to extract this subspace in a more reliable approach with more adaptation with data [32]. This subspace can be fed into a general linear model (GLM) as a design matrix or regressors to make the inference. GLM is widely used in different signal processing applications, especially fMRI data analysis due to its simplicity, interpretability and reasonable structure [33]. In a GLM for fMRI data analysis, time courses of any given voxel can be represented by a linear combination of columns in the design matrix together with additive noise term. At given time t, for each voxel in the brain, three main time series are added, i.e.,

$$y(t) = x(t) + i(t) + \xi(t), \tag{1.2}$$

where $x(t)$ is the signal of interest, $i(t)$ is the interference term and $\xi(t)$ represents the noise term. Interference or drift usually models bias and low frequency unwanted additional terms [34]. These low frequency terms can be produced by the scanner instabilities, head motion or physiological noise. Looking at the GLM model of fMRI data,

we can see that for a given time series $\mathbf{y} \in \mathbb{R}^N$ in a specific voxel, the data can be represented by

$$\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} h_{1,1} & \cdots & h_{1,p} \\ \vdots & \ddots & \vdots \\ h_{N,1} & \cdots & h_{N,p} \end{bmatrix} \times \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_p \end{bmatrix} + \begin{bmatrix} b_{1,1} & \cdots & b_{1,t} \\ \vdots & \ddots & \vdots \\ b_{N,1} & \cdots & b_{N,t} \end{bmatrix} \times \begin{bmatrix} \phi_1 \\ \vdots \\ \phi_t \end{bmatrix} + \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_N \end{bmatrix} \tag{1.3}$$

or compactly

$$\mathbf{y} = \underbrace{\mathbf{H}\boldsymbol{\theta}}_{\mathbf{x}} + \underbrace{\mathbf{B}\boldsymbol{\phi}}_{\mathbf{i}} + \boldsymbol{\xi} = \mathbf{C}\boldsymbol{\beta} + \boldsymbol{\xi}, \tag{1.4}$$

where $\mathbf{H} \in \mathbb{R}^{N\times p}$ is a known matrix whose columns span the subspace containing the target signal $\mathbf{x}$, $\mathbf{B} \in \mathbb{R}^{N\times t}$ is a known matrix whose columns span the subspace containing the interference signal $\mathbf{i}$ and $\boldsymbol{\theta} \in \mathbb{R}^p$ and $\boldsymbol{\phi} \in \mathbb{R}^t$ are unknown coordinates. In this model, $\mathbf{C} \in \mathbb{R}^{N\times(p+t)}$ is the design matrix where part of its columns represent the target subspace and the rest represent the interference subspace. Note that the subspaces $\langle\mathbf{H}\rangle$ and $\langle\mathbf{B}\rangle$ are assumed to be linearly independent, and the above model holds for all voxels and spatial correlations between voxels are ignored. $\boldsymbol{\xi} \in \mathbb{R}^N$ is the error or noise term which can generally have any characteristics. But for simplicity or having exact and guaranteed results, often it is assumed that this noise follows a parametric density model, for example Gaussian density which is parametrized by the mean and the variance. These assumptions are just for making the problem mathematically tractable. In reality it is always possible to have deviations in the mentioned model, particularly in the noise characteristics.

1.3 Problem Formulation

The problem of localizing neural activity can be seen as a binary hypothesis testing problem, i.e.,

$$\mathcal{H}_0 : \textbf{Given voxel is not active.} \quad \text{vs} \quad \mathcal{H}_1 : \textbf{Given voxel is active}. \tag{1.5}$$

In binary hypothesis testing, the absence or presence of an event (in this case neural activity) is under investigation. For example, for the model of (1.4), the main objective of fMRI activation detection task is to check if in a given observation vector $\mathbf{y}$, any

columns of the matrix $\mathbf{H}$ are present. This can be achieved by looking at the entries of $\boldsymbol{\theta}$. In other words, the binary hypothesis test in its explicit form is deciding between the following two hypotheses:

$$\mathcal{H}_0 : \boldsymbol{\theta} = \mathbf{0}, \quad \text{vs} \quad \mathcal{H}_1 : \boldsymbol{\theta} \neq \mathbf{0}, \tag{1.6}$$

where other unknown parameters (e.g., $\boldsymbol{\phi}$) can take any arbitrary values in both hypotheses. When $\boldsymbol{\theta} = \mathbf{0}$, it means there is no sign of target signal ($\mathbf{x}$) in the given time series and in contrast, when $\boldsymbol{\theta} \neq \mathbf{0}$ it means that the target signal has some contributions in forming $\mathbf{y}$. The first case is known as the *null* hypothesis while the latter is known as an *alternative* hypothesis. Note that there is no intersection between these two hypotheses and they are totally distinct, otherwise the problem is not solvable. There are numerous methods in detection theory which can solve the above problem using different assumptions in the model.

1.4 Motivation

Despite the existence of various methods in signal processing for solving the problem of (1.6), still those methods are dealing with either restricted or very general assumptions; in both cases such methods are not suitable for studying fMRI or other types of imaging data. For example, it is commonly accepted that the noise in such data can be modeled by an independent and identically distributed (i.i.d.) Gaussian density while this is not always correct in the real world [35], and investigation of robust methods is required. However, most of the methods in robust detection literature is dealing with non-parametric methods around a nominal known density which is not applicable to our application due to lack of information about the exact nominal density. In the field of statistics, there are some techniques for robust parameter estimation which are directly used in parametric detection framework (for example Huber's loss function), but they are not developed enough in signal processing community to be used in real world problems. On the other side of spectrum, the methods which use classical but irrelevant assumptions do not work well either. However, using the information obtained from the data generating process, some modifications can be made on those methods to improve

their performance significantly. Thus, in this thesis we set to use our statistical knowledge and establish a link between existing methods in the signal processing literature and the tools form information geometry and statistics to design and develop suitable methods for using detection and estimation theory in imaging systems especially for fMRI data analysis.

1.5 Thesis Outline

The rest of this thesis is organized as follows.

Chapter 2: Literature Review

This chapter starts by categorizing estimation and detection approaches. Then, in particular two categories of the works are presented, 1: parametric detectors which are based on estimation theory, 2: robust detectors which consist of parametric and non-parametric methods. Both methods are overviewed and their advantages and disadvantages are discussed.

Chapter 3: Robust Subspace Detectors Based on α-Divergence with Application to Detection in Imaging

In this chapter, a new highly robust detector is proposed using likelihood ratio, Wald and Rao tests framework in i.i.d. nominally Gaussian noise with unwanted interference. These detectors are developed using existing link between the maximum likelihood based approaches and divergences in information geometry. Classical methods are generalized using the general family of divergences called $\alpha-$divergence. Moreover, their asymptotic behavior is discussed, and some automatic approaches for tuning hyperparameter α are presented. Finally, the superiority of the methods are presented in simulated data and real applications including fMRI, hyperspectral and SAR imaging data.

Chapter 4: Learning Robust and Sparse Principal Components with the $\alpha-$Divergence

In this chapter, the main shortcoming of robust subspace detectors which is the assumption that the signal subspace is perfectly known is addressed. A data-driven approach is presented which is based on principal component analysis and $\alpha-$divergence to derive existing patterns in data without any concerns about outliers. Our methods are based on different frameworks and assumptions. Using broad simulation scenarios and real world implementations, it is shown that the proposed techniques can be beneficial.

These approaches can be used as the initial steps of robust subspace detectors to recover the true subspace of data.

Chapter 5: Robust Adaptive Matched Filter

This chapter deals with multivariate noise where there is no guarantee that the secondary data is signal free and so data cannot directly be fed into the maximum likelihood based covariance estimation block. Using α-divergene as a robust measure, an adaptive detector is developed which can significantly reduce the impact of possible existing target signals in the secondary observed data. This can be useful for any scenario in environment where the signal free assumption cannot be met for the training data.

Chapter 6: Adaptive Subspace Detector in High Dimensional Space with Insufficient Training Data

In contrast to the previous chapters, in this chapter extra prior information about the data (in this case low rank assumption) is used to significantly reduce the number of unknown parameters. This can help to increase the estimation accuracy and consequently reduce the estimation variance. Simulation results show the superiority of the algorithm compared to well known adaptive subspace detectors.

Chapter 7: Adaptive Brain Activity Detection in Structured Interference and Partially Homogeneous Locally Correlated Disturbance

In this chapter, a detector in complex space is developed, which takes the advantages of prior information, in this case locally correlation, to develop a detector with a better detection rate in a finite sample regime. Partially homogeneous disturbance with unwanted interference is considered to address a general problem. The covariance matrix is estimated using conditional likelihoods where individual maximizing of each factor can lead toward an explicit estimator of banded covariance. A wide range of simulations are provided to show the superiority of this detector over other well known detectors.

Chapter 8: Conclusion and Future Work

In this Chapter, the dissertation is concluded and future research directions are presented.

Chapter 2

Literature Review

2.1 Introduction

Modern signal detection theory is now a fundamental design step of many electronic systems in different applications of radar, sonar, digital communication, image processing, speech recognition and medicine [36]. Due to its vast utility, researchers from different areas, try to develop the best possible or good enough detectors (or tests) for each specific application. Depending on how much prior information is imposed to the problem, detection can be a simple task (e.g., simple binary hypothesis test with exactly known distributions) or a hard task (e.g., multiple hypothesis testing with stochastic unknown signal in noise with unknown distributions). In a binary hypothesis testing which is the main focus of this thesis, similarity of the distributions under hypotheses leads to confusion and increasing error rate of detection [37].

In order to design a detector, different objectives can be used, among them the Neyman Pearson detector [38] and Bayesian approach [39] are the most popular ones. The latter is based on minimizing a risk function and the former is based on fixing one of the error types and minimizing the other one. To be able to compare two given arbitrary detectors, some metrics are needed. To measure the error rate of a detector, two types of error are introduced in literature (false alarm and miss detection) which will be introduced in next sections. Usually, in a parametric distribution case, the hypothesis testing turns to parameter testing. For example using the GLM in fMRI data analysis in previous chapter, it was observed that the problem of localizing the neural activity

task is equivalent to the following parameter testing

$$\mathcal{H}_0 : \boldsymbol{\theta} = \mathbf{0}, \quad \text{vs} \quad \mathcal{H}_1 : \boldsymbol{\theta} \neq \mathbf{0}, \tag{2.1}$$

where $\boldsymbol{\theta}$ is the parameter of interest. Depending on the type of problem and the prior information on the distributions in two hypotheses, different techniques exist in literature to solve the problem. In next section, we try to categorize these techniques based on their characteristics.

2.2 From Parametric to Robust Detection

In some research areas, people are interested to have an exact and precise model for the data to be able to achieve the best possible performance under those strict assumptions [9]. Under those exact assumptions, lots of nice theorems and mathematical proofs can be developed which are true just for aforementioned conditions. On the other hand, such methods are truly dependent on the assumptions that have been made and cannot tolerate even small deviations in the models [40]. This makes them super sensitive to the existing constraints. Typically, parametric models are extensively used in these methods, where the model is parametrized with limited number of parameters that can be estimated using measured data. Alternatively, there exist other approaches that are general, and are designed to have the minimum restricted assumptions [41]. Usually, these techniques are universal and work for different scenarios. Disadvantageous of the second category is their weak performance at the cost of validity for a wider data spectrum. These methods are usually non-parametric approaches which are not in our interest due to their weak performance and ignoring existing prior information in the model. So, a new category is required that has the benefits of the both categories while avoiding their weaknesses.

Robust estimation and detection are the research fields that can fill this gap. As is shown in figure 2.1, robust methods have wider ranges of validity compared to parametric methods and have almost the same performance level as parametric approaches and close to the optimal performance. In this category, an uncertainty set around the nominal model is often formed to tolerate against deviations in the models and assumptions using variety of definitions on the uncertainty set. A wider uncertainty set can tolerate

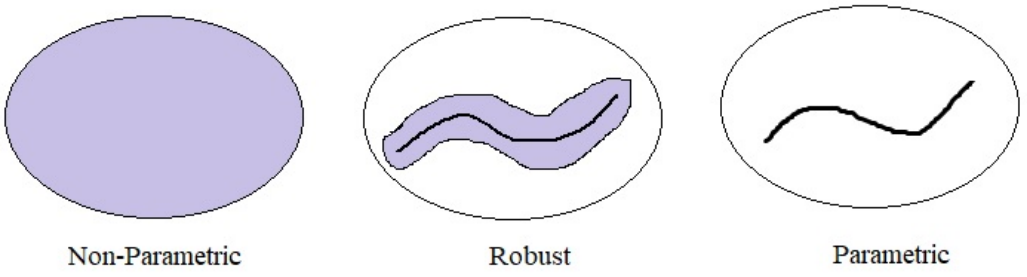

FIGURE 2.1: Categories in detection and estimation theory [9].

against heavier deviations and smaller uncertainty set achieves a better performance level. Therefore in robust methods, performance in the nominal model and the level of robustness are in trade off with each other and increasing both of them simultaneously is not possible [37].

At limit points, robust methods can be changed to a pure parametric approach or non-parametric method. This often can be obtained by changing the user defined parameters like the maximum distance around the nominal distributions [42]. Since robust methods are similar to both parametric and non-parametric approaches, it is possible to develop them either as parametric or non-parametric. In signal processing community, mostly the non-parametric approaches are developed and most of them are based on *minimax* approach [39]. The parametric version of robust detectors mostly are investigated in statistics and are the modified versions of the parametric methods using robust estimation techniques [42]. However, early approaches in both methods are developed by Huber [43–45]. Although the latter approach is the main focus of this thesis, but we present a full literature review in both techniques together with a review on parametric detectors. Before presenting the literature, we need to get familiar with the meaning of some repeated terms in detection literature like *power* and *error rate* which are shortly described in the following section.

2.3 Related Terminologies

Similar to the classification task, in a binary decision making, two types of error exist. The first error type occurs when we decide $\mathcal{H}_1$ when $\mathcal{H}_0$ is true which is the meaning of the terms *false alarm* (in detection theory), false positive (in computer science) or *significance level* (in statistics). The second type of error occurs when in contrast we

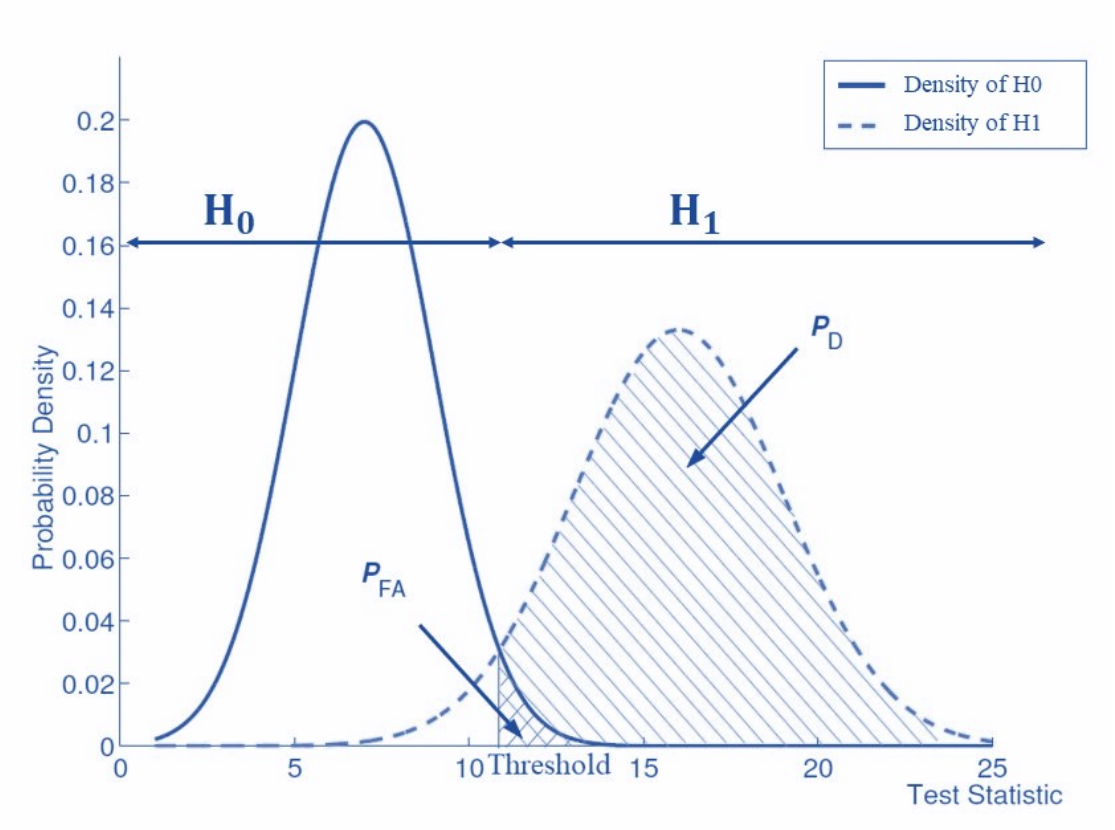

FIGURE 2.2: Visualization of P_{FA} and P_D for a binary hypothesis test.

decide $\mathcal{H}_0$ when $\mathcal{H}_1$ is true, and it's often referred by the terms *miss detection* (in detection theory) or *false negative* (in computer science) [36]. The decision making task is often performed by a randomized decision rule $\boldsymbol{\delta}(\mathbf{T}(\mathbf{Y})) = (\delta_0, \delta_1) : \mathbf{Y} \rightarrow [0,1]^2$, where each entry in this rule assigns a probability value to both hypotheses given a test statistic function ($\mathbf{T}$) of observed input $\mathbf{Y}$ and they are complementary of each other, i.e., $\delta_0 + \delta_1 = 1$. The hypothesis $\mathcal{H}_1$ will be decided only if $\delta_1 \geq \gamma$, where γ is an appropriate scalar threshold value. Let $\boldsymbol{\Delta}$ be a set of all the possible decision rules, then using above definition and let $R_i = \{\mathbf{Y} : \mathit{Decide}\ \mathcal{H}_i\}, i = 0, 1$, be the region in the space where using the decision making rule, $\mathcal{H}_i$ is decided. For any $\boldsymbol{\delta} \in \boldsymbol{\Delta}$, the false alarm and miss detection rates mathematically can be written as:

$$P_{FA}(\boldsymbol{\delta}, f_0) = \mathbf{E}_0[\delta_1] = \int_{R_1} f_0^N d\mathbf{Y} = \int_{\mathbf{Y}} \delta_1 f_0^N d\mathbf{Y}, \tag{2.2}$$

and

$$P_M(\boldsymbol{\delta}, f_1) = \mathbf{E}_1[1 - \delta_1] = \int_{R_0} f_1^N d\mathbf{Y} = \int_{\mathbf{Y}} (1 - \delta_1) f_1^N d\mathbf{Y}, \tag{2.3}$$

where f_i^N and $\mathbf{E}_i[.], i = 0, 1$, show the joint probability function in $\mathcal{H}_i$ and the expectation with respect to i^{th} density function, respectively. The overall error probability can be derived based on these two error types by $P_E(\boldsymbol{\delta}, f_0, f_1) = P_{\mathcal{H}_0} P_{FA}(\boldsymbol{\delta}, f_0) + P_{\mathcal{H}_1} P_M(\boldsymbol{\delta}, f_1)$. The *power* of a test also is another term which shows the probability that a test decide $\mathcal{H}_1$ when $\mathcal{H}_1$ is true, and naturally it is equal to $P_D = 1 - P_m$ where P_D stands for probability of detection or power of the test. Figure 2.2 shows these probabilities when the test statistic follows two Gaussian densities under two hypotheses.
Now we are ready to present literature on parametric detection.

2.4 Parametric Detectors

In a general parametric detection setup, it is assumed that a random variable $\mathbf{y} \in \mathbb{R}^m$ can be caused by a null hypothesis with density function $f(\mathbf{y}; \boldsymbol{\omega}_0)$ or an alternative hypothesis with density function $f(\mathbf{y}; \boldsymbol{\omega}_1)$ where for short they can be denoted by f_0 and f_1. In general, f_0 and f_1 can be different distribution models where $\boldsymbol{\omega}_0 \in \boldsymbol{\Omega}_0$ and $\boldsymbol{\omega}_1 \in \boldsymbol{\Omega}_1$. $\boldsymbol{\Omega}_i$, $i = 0, 1$, defines the possible parameter set for each hypothesis. We are given N i.i.d. samples, i.e., $\mathbf{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_N\} \in \mathbb{R}^{m \times N}$, the binary hypothesis test is to decide which hypothesis distribution is the source of given data, i.e.,

$$\mathcal{H}_0 : \mathbf{Y} \sim f_0^N, \quad \text{vs} \quad \mathcal{H}_1 : \mathbf{Y} \sim f_1^N. \tag{2.4}$$

In computer science point of view, this problem is a classification task where there is no need to have a labeled dataset. Instead of usual dataset, the models are known in advance (either with known or unknown parameters) and detection task is performed only using the test data. In the case of known parameters, using both Neyman Pearson and Bayesian approaches, the likelihood ratio test (LRT) is the uniformly most powerful (UMP) test.

2.4.1 Bayes Risk

In a Bayesian decision making framework, all actions are assigned a corresponding cost value. In particular, $C_{i,j}$ shows the cost of deciding the i^{th} hypothesis when j^{th} hypothesis is true. Thus, defining a risk function, we are trying to minimize the risk with

respect to decision making rules. For example in a binary hypothesis testing case the risk function is

$$\mathcal{R} = \sum_{j=0}^{1}\sum_{i=0}^{1} C_{i,j} P(\mathcal{H}_i|\mathcal{H}_j) P_{\mathcal{H}_j}. \tag{2.5}$$

For a better understanding, consider the case that correct assignment of labels, does not have any cost and wrong assignment has a cost of 1, i.e., $C_{1,1} = C_{0,0} = 1$ and $C_{1,0} = C_{0,1} = 0$. Note that this means $\mathcal{R} = P_E(\boldsymbol{\delta}, f_0, f_1)$ and minimization of Bayesian risk is equivalent to minimization of error rate in decision making rule.

Let $R_1 = \{\mathbf{Y} : Decide \mathcal{H}_1\}$ be the region in the space where using the decision making rule, $\mathcal{H}_1$ is decided and R_0 is its complement region in the space i.e., $\int_{R_0} p(\mathbf{Y}|\mathcal{H}_i)d\mathbf{Y} = 1 - \int_{R_1} p(\mathbf{Y}|\mathcal{H}_i)d\mathbf{Y}$. Then, the risk can be written as,

$$\begin{aligned}
\mathcal{R} &= C_{0,0}P_{\mathcal{H}_0}\int_{R_0} p(\mathbf{Y}|\mathcal{H}_0)d\mathbf{Y} + C_{1,0}P_{\mathcal{H}_0}\int_{R_1} p(\mathbf{Y}|\mathcal{H}_0)d\mathbf{Y} \\
&+ C_{0,1}P_{\mathcal{H}_1}\int_{R_0} p(\mathbf{Y}|\mathcal{H}_1)d\mathbf{Y} + C_{1,1}P_{\mathcal{H}_1}\int_{R_1} p(\mathbf{Y}|\mathcal{H}_1)d\mathbf{Y} \\
&= C_{0,1}P_{\mathcal{H}_1} + C_{0,0}P_{\mathcal{H}_0} + \int_{R_1}(C_{1,0}P_{\mathcal{H}_0} - C_{0,0}P_{\mathcal{H}_0})p(\mathbf{Y}|\mathcal{H}_0) \\
&+ (C_{1,1}P_{\mathcal{H}_1} - C_{0,1}P_{\mathcal{H}_1})p(\mathbf{Y}|\mathcal{H}_1)d\mathbf{Y}.
\end{aligned} \tag{2.6}$$

For given $\mathbf{Y}$, it is decided $\mathcal{H}_1$ only if

$$(C_{1,0}P_{\mathcal{H}_0} - C_{0,0}P_{\mathcal{H}_0})p(\mathbf{Y}|\mathcal{H}_0) \leq (C_{1,1}P_{\mathcal{H}_1} - C_{0,1}P_{\mathcal{H}_1})p(\mathbf{Y}|\mathcal{H}_1), \tag{2.7}$$

or by assuming $C_{1,0} \geq C_{0,0}$ and $C_{0,1} \geq C_{1,1}$ we decide $\mathcal{H}_1$ when

$$\frac{p(\mathbf{Y}|\mathcal{H}_1)}{p(\mathbf{Y}|\mathcal{H}_0)} \geq \frac{(C_{1,0} - C_{0,0})P_{\mathcal{H}_0}}{(C_{0,1} - C_{1,1})P_{\mathcal{H}_1}} = \gamma. \tag{2.8}$$

Based on the above result, the LRT ($p(\mathbf{Y}|\mathcal{H}_1)/p(\mathbf{Y}|\mathcal{H}_0)$) is the best detector in Bayesian setup [39] which minimizes the Bayes risk function. However, in signal processing we are interested to the case to have control on specific error type (for example false alarm) not the combination of the errors. For instance, in radar community or biomedical data analysis, it is always interesting to design a detector that has a bounded false alarm rate while it has a reasonable level of detection. The reason is that for example detection of a wrong target in sky means an enemy aircraft is detected and need to do some military reactions which usually has a high financial cost. Also in medical data analysis, false diagnosing can gives a high stress to the patient which sometimes can have even worse

consequences compared to the detected illness. To address this requirement, Neyman Pearson test has been proposed which is described in next section.

2.4.2 Neyman Pearson Test

Neyman and Pearson have shown that LRT is the UMP test among all possible decision making rules when the false alarm rate is fixed to $\bar{\alpha}$ [38]. Mathematically, in this setup we seek to find the solution of

$$\boldsymbol{\delta} = \sup_{\boldsymbol{\delta} \in \boldsymbol{\Delta}} P_D(\boldsymbol{\delta}, f_1) \ \textbf{s.t.} \ P_{FA}(\boldsymbol{\delta}, f_0) = \bar{\alpha}. \tag{2.9}$$

Using the Lagrangian multipliers the problem can be changed to an unconstrained optimization of

$$\begin{aligned} \boldsymbol{\delta}^* &= \sup_{\boldsymbol{\delta} \in \boldsymbol{\Delta}} F \\ &= P_D + \lambda(P_{FA} - \bar{\alpha}) \\ &= \int_{R_1} p(\mathbf{Y}|\mathcal{H}_1) + \lambda p(\mathbf{Y}|\mathcal{H}_0) d\mathbf{Y} - \lambda\bar{\alpha}. \end{aligned} \tag{2.10}$$

To maximize F, we should include $\mathbf{Y}$ in R_1 when the integral is positive [36]. In other words, we decide R_1 if

$$p(\mathbf{Y}|\mathcal{H}_1) + \lambda p(\mathbf{Y}|\mathcal{H}_0) \geq 0, \tag{2.11}$$

or

$$\frac{p(\mathbf{Y}|\mathcal{H}_1)}{p(\mathbf{Y}|\mathcal{H}_0)} \geq -\lambda, \tag{2.12}$$

where the λ is a negative scalar which can be defined based on the required false alarm rate. Similar to Bayes risk minimization case, again it is shown that LRT is the best possible test. However, the above results are valid when distributions under both hypotheses are perfectly known or does not depend on any unknown value [39]. Usually this ideal condition is not met in many signal processing applications especially the fMRI data that is under our investigation. Thus, more investigation for finding some suboptimal tests are needed. In next section, some of these parametric methods will be introduced.

2.4.3 Generalized Likelihood Ratio Test

In any signal processing applications, usually there are some unknown parameters in the assumed model which need to be estimated using existing data. Therefore, in most of these cases there is no UMP test [46]. When such optimal test does not exist, we need to find tests with suboptimal performances. The first alternative can be the generalized LRT (GLRT) where in this test all unknown parameters in both $\mathcal{H}_0$ and $\mathcal{H}_1$ will be replaced with their maximum likelihood estimators (MLEs) [47]. Partitioning the parameter vector $\boldsymbol{\omega}$ in (2.4) to sub parameter vectors of $\boldsymbol{\theta}$ (paramete of interest) and $\boldsymbol{\lambda}$ (nuisance parameter), i.e., $\boldsymbol{\omega} = [\boldsymbol{\theta}^\top, \boldsymbol{\lambda}^\top]^\top$ and considering the same distribution models f for both hypotheses with different parametr sets, the GLRT can be written as [48]

$$T_G(\mathbf{Y}) \quad = \quad \frac{2}{N}[\sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_1} \sum_{i=1}^{N} \log\left(f(\mathbf{y}_i, \boldsymbol{\omega})\right) - \sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_0} \sum_{i=1}^{N} \log\left(f(\mathbf{y}_i, \boldsymbol{\omega})\right)] \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma, \qquad (2.13)$$

where $\boldsymbol{\Omega}_0 = [\mathbf{0}^\top, \boldsymbol{\lambda}^\top]^\top$ is the restricted parameter set, but $\boldsymbol{\Omega}_1$ can get any arbitrary values. The log(.) function is a monotone increasing function that does not have any impact on the detector result but can significantly make the estimation part easier especially for the general family of exponential noise densities. This test is presenting a measure that quantifies how much estimated densities in two hypotheses are different from each other. Naturally, high test value assigned to an observation means that we are highly confident that the recorded data is coming from $\mathcal{H}_1$ rather than $\mathcal{H}_0$. The appropriate threshold value can be computed based on the desired false alarm rate. To estimate and find the likelihood difference of the above expression, different assumptions are used in the literature which lead to different famous detectors. In next sections, we briefly introduce four well known detectors of matched subspace detector (MSD) [49], Kelly's GLRT [50], adaptive subspace detector (ASD)[51] and finally adaptive matched filter (AMF) [52] which are used for the comparison purpose in this thesis. These detectors are based on Gaussian noise distribution in both complex domain and real domain with both known and unknown covariance matrices assumptions.

2.4.3.1 Matched Subspace Detector

One of early classical approaches based on GLRT is the MSD. For the linear model given by

$$\mathbf{y} = \underbrace{\mathbf{H}\boldsymbol{\theta}}_{\mathbf{x}} + \underbrace{\mathbf{B}\boldsymbol{\phi}}_{\mathbf{i}} + \boldsymbol{\xi} = \mathbf{C}\boldsymbol{\beta} + \boldsymbol{\xi}, \tag{2.14}$$

in MSD, it is assumed that $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_N)$ which means noise components are uncorrelated or correlated with known covariance [49]. Therefore, all time courses of a given voxel can be put in entries of vector $\mathbf{y}$ to form the observation vector. The parameter of interest is $\boldsymbol{\theta}$, and the nuisance parameters are σ^2 and $\boldsymbol{\phi}$ which are unknown. At first step, all parameters are estimated with MLEs in both hypotheses to be fed into the log likelihood functions.

In fact when we use the term *subspace* it means that the signal can be built with more than one modal and it is actually a linear combination of those with any arbitrary weights. Before extracting the test, we consider following extra assumptions :

1. The subspaces $\mathbf{H}$ and $\mathbf{B}$ are known. In this case, $\mathbf{H}$ can be obtained by for example convolving stimuli with HRF and some of its derivatives. Since $\mathbf{B}$ should model the drift terms, we can use a set of discrete cosine functions with low frequency variations in columns of $\mathbf{B}$.

2. The subspaces $\mathbf{H}$ and $\mathbf{B}$ are not orthogonal but they are independent. That means we cannot model each column of $\mathbf{H}$ as a linear combination of columns of $\mathbf{B}$.

Depending on the source $\mathbf{y}$, it can have one of two following densities

$$\mathcal{H}_0 : \mathbf{y} \sim \mathcal{N}(\mathbf{B}\boldsymbol{\phi}, \sigma^2\mathbf{I}_N) \quad \text{vs} \quad \mathcal{H}_1 : \mathbf{y} \sim \mathcal{N}(\mathbf{H}\boldsymbol{\theta} + \mathbf{B}\boldsymbol{\phi}, \sigma^2\mathbf{I}_N). \tag{2.15}$$

The probability density function for the multivariate normal (MVN) vector $\mathbf{y}$ with the covariance of $\sigma^2\mathbf{I}_N$ is

$$f(\mathbf{y}; \boldsymbol{\omega}) = (2\pi\sigma^2)^{-N/2} \exp\{-\frac{1}{2\sigma^2}\|\boldsymbol{\xi}\|_2^2\} \tag{2.16}$$

where the noise $\boldsymbol{\xi}$ is

$$\boldsymbol{\xi} = \mathbf{y} - \mathbf{H}\boldsymbol{\theta} - \mathbf{B}\boldsymbol{\phi}. \tag{2.17}$$

This function is dependent on values $(\boldsymbol{\theta}, \boldsymbol{\phi}, \sigma^2)$. For estimated values of $(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\phi}}_1, \hat{\sigma}_1^2)$ and $(\hat{\boldsymbol{\phi}}_0, \hat{\sigma}_0^2)$ the GLRT is given by

$$\ell(\mathbf{y}) = \frac{f(\mathbf{y}; \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\phi}}_1, \hat{\sigma}_1^2)}{f(\mathbf{y}; \boldsymbol{\theta} = \mathbf{0}, \hat{\boldsymbol{\phi}}_0, \hat{\sigma}_0^2)}. \tag{2.18}$$

With thresholding on $\ell(\mathbf{y})$, we can detect the presence of signal of interest. Plugging MLEs into (2.18) and using monotone increasing function, the well known MSD is given by

$$\ell_{MSD}(\mathbf{y}) \quad = \quad \frac{\hat{\sigma}_0^2 - \hat{\sigma}_1^2}{\hat{\sigma}_1^2} = \frac{\mathbf{y}^\top \mathbf{P}_{\mathbf{B}^\perp} \mathbf{P}_{\mathbf{G}} \mathbf{P}_{\mathbf{B}^\perp} \mathbf{y}}{\mathbf{y}^\top \mathbf{P}_{\mathbf{B}^\perp} \mathbf{P}_{\mathbf{G}^\perp} \mathbf{P}_{\mathbf{B}^\perp} \mathbf{y}} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma, \tag{2.19}$$

where $\mathbf{G} = \mathbf{P}_{\mathbf{B}^\perp}\mathbf{H}$, $\mathbf{P}_{\mathbf{D}} = \mathbf{D}(\mathbf{D}^\top\mathbf{D})^{-1}\mathbf{D}^\top$ is the projection matrix and the projector $\mathbf{P}_{\mathbf{B}^\perp}$ projects onto the subspace $\langle \mathbf{B} \rangle^\perp$. With exact looking at the numerator and denominator, we notice the energy of estimated target signal and noise vector are using in a division to check their strengths for final decision.

2.4.3.2 Kelly's GLRT

MSD considers the case that the covariance matrix is exactly known. For the first time Kelly in [50] addressed the circular zero mean complex Gaussian noise with unknown covariance for detection of a rank one signal in the absence of any interference ($\boldsymbol{\phi} = \mathbf{0}$). In order to estimate the covariance matrix, it was assumed K secondary signal free data $\{\mathbf{n}_k\}_{k=1}^K$ are available which exactly share the same covariance matrix with the test data ($\mathbf{y}$). In the case of fMRI data analysis, it means K time series recorded from K voxels are available where we are sure that the measured data does not represent any types of neural activity. In radar application, it is equivalent to use the neighbor range cells to estimate the covariance matrix [53]. Depending on the source of $\mathbf{y}$, it can have one of the two following densities

$$\mathcal{H}_0 : \begin{cases} \mathbf{y} \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}) \\ \mathbf{n}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}), \ k = 1, \cdots, K, \end{cases} \quad \text{vs} \quad \mathcal{H}_1 : \begin{cases} \mathbf{y} \sim \mathcal{CN}(\mathbf{h}\theta, \mathbf{R}) \\ \mathbf{n}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}), \ k = 1, \cdots, K, \end{cases} \tag{2.20}$$

Density function of a zero mean circular complex random variable $\mathbf{y}$ is given by

$$f(\mathbf{y}) = \frac{1}{\pi^N \det(\mathbf{R})} \exp\{-\mathbf{y}^\dagger \mathbf{R}^{-1} \mathbf{y}\}. \tag{2.21}$$

Using this density function and forming the joint likelihood function with K signal free data and maximizing with respect to unknown parameters make the GLRT in following form

$$\ell(\mathbf{y}) = \frac{f(\mathbf{y}, \mathbf{n}_1, \cdots, \mathbf{n}_K; \hat{\theta}_1, \hat{\mathbf{R}}_1)}{f(\mathbf{y}, \mathbf{n}_1, \cdots, \mathbf{n}_K; \hat{\mathbf{R}}_0)} = \frac{f_1(\mathbf{y}) \prod_{k=1}^{K} f_1(\mathbf{n}_k)}{f_0(\mathbf{y}) \prod_{k=1}^{K} f_0(\mathbf{n}_k)}. \tag{2.22}$$

After applying some tedious simplifications using linear algebra, it is not hard to show that the Kelly's GLRT is given by

$$\ell_{GLRT}(\mathbf{y}) \quad = \quad \frac{|\mathbf{h}^\dagger \mathbf{S}^{-1} \mathbf{y}|^2}{(\mathbf{h}^\dagger \mathbf{S}^{-1} \mathbf{h})[1 + \frac{1}{K} \mathbf{y}^\dagger \mathbf{S}^{-1} \mathbf{y}]} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma, \tag{2.23}$$

where $\mathbf{S} = \sum_{k=1}^{K} \mathbf{n}_k \mathbf{n}_k^\dagger$ is the sample covariance matrix and γ is the appropriate threshold value. Nevertheless, this detector is designed to detect a rank one signal in a homogeneous environment. Homogeneity refers to the case that the test data ($\mathbf{y}$) and secondary data $\{\mathbf{n}_k\}_{k=1}^{K}$ both have exactly the same covariance structure. To address both issues, ASD was developed later on [51].

2.4.3.3 Adaptive Subspace Detector

To develop Kelly's GLRT for the non-homogeneous noise environment and multimodal signal, the detection of subspace signal in partially homogeneous noise was proposed in [51]. The term partially homogeneous refers to the case that the secondary data share almost the same covariance structure but can be different up to an scalar value ($\sigma^2\mathbf{R}$ instead of $\mathbf{R}$). However, in this approach, the interference was ignored. In this hypothesis testing, the test data $\mathbf{y}$ is following one of two densities

$$\mathcal{H}_0 : \begin{cases} \mathbf{y} \sim \mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{R}) \\ \mathbf{n}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}), \ k = 1, \cdots, K, \end{cases} \quad \text{vs} \quad \mathcal{H}_1 : \begin{cases} \mathbf{y} \sim \mathcal{CN}(\mathbf{H}\boldsymbol{\theta}, \sigma^2\mathbf{R}) \\ \mathbf{n}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}), \ k = 1, \cdots, K. \end{cases} \tag{2.24}$$

In this case, compared to Kelly's GLRT, in the process of maximizing the joint likelihood function, an extra parameter σ^2 needs to be estimated while $\boldsymbol{\theta}$ is a vector not a scalar. Thus, the ASD can be derived using

$$\ell(\mathbf{y}) = \frac{f(\mathbf{y}, \mathbf{n}_1, \cdots, \mathbf{n}_K; \hat{\boldsymbol{\theta}}_1, \hat{\sigma}_1^2, \hat{\mathbf{R}}_1)}{f(\mathbf{y}, \mathbf{n}_1, \cdots, \mathbf{n}_K; \hat{\sigma}_0^2, \hat{\mathbf{R}}_0)} = \frac{f_1(\mathbf{y}) \prod_{k=1}^{K} f_1(\mathbf{n}_k)}{f_0(\mathbf{y}) \prod_{k=1}^{K} f_0(\mathbf{n}_k)}. \tag{2.25}$$

After doing some linear algebra it is not hard to show

$$\ell_{ASD}(\mathbf{y}) = \frac{\mathbf{y}^\dagger \mathbf{S}^{-\frac{1}{2}} \mathbf{P}_{\tilde{\mathbf{H}}} \mathbf{S}^{-\frac{1}{2}} \mathbf{y}}{\mathbf{y}^\dagger \mathbf{S}^{-1} \mathbf{y}} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma, \tag{2.26}$$

where $\tilde{\mathbf{H}} = \mathbf{S}^{-\frac{1}{2}}\mathbf{H}$ and again $\mathbf{S} = \frac{1}{K}\sum_{k=1}^{K} \mathbf{n}_k \mathbf{n}_k^\dagger$ is the sample covariance matrix. In design steps of both Kelly's GLRT and ASD, it was assumed that parameters are unknown and at the same time, using both test and secondary data, the likelihood function is maximized. This makes the detector complex. Therefore, AMF was introduced to reduce this complexity.

2.4.3.4 Adaptive Matched Filter

In contrast to Kelly's GLRT and ASD where all parameters are maximized using the joint likelihood, in this method for the case of

$$\mathcal{H}_0 : \begin{cases} \mathbf{y} \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}) \\ \mathbf{n}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}), \ k = 1, \cdots, K, \end{cases} \quad \text{vs} \quad \mathcal{H}_1 : \begin{cases} \mathbf{y} \sim \mathcal{CN}(\mathbf{h}\theta, \mathbf{R}) \\ \mathbf{n}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}), \ k = 1, \cdots, K, \end{cases} \tag{2.27}$$

in design step, it was assumed first that the covariance matrix in known [52]. Then the GLRT was calculated with maximizing the joint likelihood with respect to θ. Finally at the end, the covariance matrix was replaced with the estimated sample covariance using secondary data. Therefore, it is easy to show

$$\ell_{AMF}(\mathbf{y}) \quad = \quad \frac{|\mathbf{h}^\dagger \mathbf{S}^{-1} \mathbf{y}|^2}{(\mathbf{h}^\dagger \mathbf{S}^{-1} \mathbf{h})} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma. \tag{2.28}$$

Now, it can be seen that the structure of the detector is more compact compared to other detectors. This can make the detection task fast and more adapted with environments where a fast decision making algorithm is needed.

2.4.4 Wald Test

Due to lack of UMP test in many applications, it is always interesting to investigate and develop tests which can beat GLRT in some problems or can achieve the same performance with less parameters. In a Wald test, instead of basing the test on likelihood values, we are interested to directly look at the parameter values. The first way of achieving

this distribution is the Wald test statistic. Let us define $\log L(\boldsymbol{\omega};\mathbf{Y}) = \sum_{i=1}^{N} \log f(\mathbf{y}_i;\boldsymbol{\omega})$, then the Fisher information matrix is defined as

$$[\mathbf{I}(\boldsymbol{\omega})]_{i,j} = -\mathbf{E}\left[\frac{\partial^2 \log L}{\partial \boldsymbol{\omega}_i \partial \boldsymbol{\omega}_j}\right] \tag{2.29}$$

and can be partitioned to

$$\mathbf{I}(\boldsymbol{\omega}) = \begin{bmatrix} \mathbf{I}_{\boldsymbol{\theta\theta}} & \mathbf{I}_{\boldsymbol{\theta\lambda}} \\ \mathbf{I}_{\boldsymbol{\lambda\theta}} & \mathbf{I}_{\boldsymbol{\lambda\lambda}} \end{bmatrix}. \tag{2.30}$$

Based on the results in [54], it can be seen that the asymptotic covariance of the parameter vector $\boldsymbol{\theta}$ is $\left(\mathbf{I}_{\boldsymbol{\theta\theta}} - \mathbf{I}_{\boldsymbol{\theta\lambda}}\mathbf{I}_{\boldsymbol{\lambda\lambda}}^{-1}\mathbf{I}_{\boldsymbol{\lambda\theta}}\right)^{-1}$. Finally, knowing this, the Wald test can be formulated as

$$\begin{aligned} T_W(\mathbf{y}) &= (\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_0)^\top \left\{ \left[\mathbf{I}^{-1}(\hat{\boldsymbol{\omega}}_1)\right]_{\boldsymbol{\theta\theta}} \right\}^{-1} (\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_0) \\ &= (\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_0)^\top \left(\mathbf{I}_{\boldsymbol{\theta\theta}} - \mathbf{I}_{\boldsymbol{\theta\lambda}}\mathbf{I}_{\boldsymbol{\lambda\lambda}}^{-1}\mathbf{I}_{\boldsymbol{\lambda\theta}}\right)(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_0) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma, \end{aligned} \tag{2.31}$$

where the parameters are whitened asymptotically then used in quadratic form to compute its norm value. $\hat{\boldsymbol{\omega}}_1 = [\hat{\boldsymbol{\theta}}_1^\top, \hat{\boldsymbol{\lambda}}_1^\top]^\top$ is the parameter vector estimated under $\mathcal{H}_1$. In contrast to GLRT which is based on the estimated parameters in both hypotheses, the Wald test just considers the alternative case. In $\mathcal{H}_0$ case, due to convergence of estimated $\boldsymbol{\theta}$ to $\boldsymbol{\theta}_0$, we expect to see small test values and naturally using an appropriate γ, we reject $\mathcal{H}_1$ and vise versa. It is shown in literature that sometimes this test can lead to a better performance than GLRT. Also in some problem models, the Wald test is equivalent to other tests for example AMF [55]. To see the procedure of deriving the test, we suggest to see the detectors which have been proposed in [56, 57].

2.4.5 Rao or Score Test

Another test statistics that have been used a lot in literature is Rao or Score test [48]. Instead of looking at the likelihood or parameter values, the derivative or score function of log likelihood is used in the test. In [54], it has been shown that the score function $\partial \log L(\boldsymbol{\omega};\mathbf{Y})/\partial\boldsymbol{\theta}$ asymptotically has the covariance of $\left(\mathbf{I}_{\boldsymbol{\theta\theta}} - \mathbf{I}_{\boldsymbol{\theta\lambda}}\mathbf{I}_{\boldsymbol{\lambda\lambda}}^{-1}\mathbf{I}_{\boldsymbol{\lambda\theta}}\right)$. In Rao test, the reference is the estimated parameters in $\mathcal{H}_0$. In $\mathcal{H}_0$ we expect to get the score function close to $\mathbf{0}$ due to optimality of the MLEs. Therefore, if we see a whitened score function with higher norm value, it means that most likely the alternative hypothesis is

correct. According to above discussion the Rao test is given by

$$\begin{aligned} T_R(\mathbf{y}) &= \left.\frac{\partial \log L}{\partial \boldsymbol{\theta}}\right|^{\top}_{\boldsymbol{\omega}=\hat{\boldsymbol{\omega}}_0} \left[\mathbf{I}^{-1}(\hat{\boldsymbol{\omega}}_0)\right]_{\boldsymbol{\theta}\boldsymbol{\theta}} \left.\frac{\partial \log L}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\omega}=\hat{\boldsymbol{\omega}}_0} \\ &= \left.\frac{\partial \log L}{\partial \boldsymbol{\theta}}\right|^{\top}_{\boldsymbol{\omega}=\hat{\boldsymbol{\omega}}_0} \left(\mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\theta}} - \mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\lambda}}\mathbf{I}^{-1}_{\boldsymbol{\lambda}\boldsymbol{\lambda}}\mathbf{I}_{\boldsymbol{\lambda}\boldsymbol{\theta}}\right)^{-1} \left.\frac{\partial \log L}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\omega}=\hat{\boldsymbol{\omega}}_0} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma, \end{aligned}$$

$\hat{\boldsymbol{\omega}}_0 = [\mathbf{0}^{\top}, \hat{\boldsymbol{\lambda}}_0^{\top}]^{\top}$ is the estimated parameter vector in $\mathcal{H}_0$. For example in the case of signal detection we suggest to see [58] for the procedure that the test can be derived. For subspace detection in complex Gaussian noise, in absence of any interference, i.e.,

$$\mathcal{H}_0 : \begin{cases} \mathbf{y} \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}) \\ \mathbf{n}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}), \ k = 1, \cdots, K, \end{cases} \quad \text{vs} \quad \mathcal{H}_1 : \begin{cases} \mathbf{y} \sim \mathcal{CN}(\mathbf{H}\boldsymbol{\theta}, \mathbf{R}) \\ \mathbf{n}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}), \ k = 1, \cdots, K, \end{cases} \tag{2.32}$$

the Rao test is given by [59]

$$T_R(\mathbf{y}) = \frac{\tilde{\mathbf{y}}^{\dagger}\mathbf{P}_{\tilde{\mathbf{H}}}\tilde{\mathbf{y}}}{(1+\tilde{\mathbf{y}}^{\dagger}\tilde{\mathbf{y}})(1+\tilde{\mathbf{y}}^{\dagger}\mathbf{P}_{\tilde{\mathbf{H}}^{\perp}}\tilde{\mathbf{y}})} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma, \tag{2.33}$$

where $\tilde{\mathbf{H}} = \hat{\mathbf{E}}^{-\frac{1}{2}}\mathbf{H}$, $\tilde{\mathbf{y}} = \hat{\mathbf{E}}^{-\frac{1}{2}}\mathbf{y}$ and $\hat{\mathbf{E}} = \sum_{k=1}^{K} \mathbf{n}_k\mathbf{n}_k^{\dagger} = K\mathbf{S}$.

2.4.6 Constrained Detection

In all above discussed methods, the parameters are replaced with their MLEs, due to its consistency and efficiency. Asymptotically MLEs attains the Cramér–Rao lower bound (CRLO) [60] which shows the minimum possible variance that an unbiased estimator can achieve. This is true when we are in large sample regime. Nevertheless, injecting information (using our knowledge about the physical behavior of underlaying process) to the problem can significantly improve the final performance in a finite sample regime. One example of using this prior information is adding constraints on the structure of the covariance matrices in multivariate signal analysis. Theses constraints include persymmetric covariances in array signal processing [61], low rank covariance [62], etc. These structures can decrease the number of required parameters and push the methods toward their asymptotic regime to achieve better performances for a finite sample case.

2.5 Robust Detectors

Most statistical signal processing methods rely on strong assumptions like Gaussianity of the noise [63] which may not be true for some real world applications. Even in a highly clean data, some proportion of outliers exist and the range of 1 to 10% contamination is a typical deviation level in data acquiring process [9]. These deviations can cause the performance of these methods to drastically degrade or to completely break down [40]. For example, the occurrence of impulsive noise has been observed in outdoor mobile communication channels, due to switching transients in power lines or automobile ignition [64]. This noise has been also observed in radar and sonar systems as a consequence of natural or artificial electromagnetic and acoustic interference [65, 66], and in indoor wireless communication channels, caused by microwave ovens and devices with electromechanical switches [40, 67]. Finally, it has been also observed that in measurements captured from different regions of the human brain, such as in MRI, non-Gaussian noise and interference exist [68].

In robust statistics, we are trying to address these issues by focusing on the majority of data and propose methods which can identify outliers. Robust statistics consists of two distinct but related categories of robust estimation and robust detection [9]. In this thesis, we cover both but with more focus on the latter. Robust detection offers two approaches of robust non-parametric and robust parametric methods to solve shortcomings of classical parametric methods. In non-parametric methods which mostly are based on minimax concept, a set of uncertainty is defined around the nominal model under both hypotheses. Then, the detector is designed for the worst case scenario to guarantee the performance in severe outliers. Different types of measure have been proposed in literature to define the neighborhood around the models that each measure leads to different robust approach. The common step in all these non parametric approaches is using the minimax concept. However, these methods assume the nominal model is perfectly known which might not be correct. In second type of robust detection approaches (robust parametric approaches) either it is assumed that the classical assumptions still can be a good model for representing data (with minor modification), or uses a more complex model with a larger number of parameters to approximate the entire data including outliers. After robust parameter estimation, the classical basic test frameworks (GLRT, Wald and Rao) are used to develop robust detectors.

In next sections, we present a brief overviews of both non-parametric and parametric robust detectors.

2.5.1 Non-Parametric Robust Detection

In non-parametric detection method, it is assumed that the nominal models in both hypotheses (f_0, f_1) are known. It was shown that the likelihood ratio test is UMP test when the distributions are known. Therefore, given the test statistics $T(\mathbf{Y}) = \prod_{i=1}^{N} f_1(\mathbf{y}_i)/f_0(\mathbf{y}_i)$, the optimal detector is given by

$$\delta_1^* = \begin{cases} 1, & T(\mathbf{Y}) > \gamma \\ \in [0,1], & T(\mathbf{Y}) = \gamma \\ 0. & T(\mathbf{Y}) < \gamma \end{cases} \tag{2.34}$$

However, this detector is not robust against outliers. Just assume a single severe outlier exist. Due to direct relation of $T(\mathbf{Y})$ to all single observations, this single outlier can push the test statistics value into unwanted region and make the result inverse. To overcome this problem, in robust detection, it is assumed that the true models under two hypotheses (g_0, g_1) belong to the defined sets in the neighborhood of nominal models, i.e., $g_0 \in \mathcal{F}_0$ and $g_1 \in \mathcal{F}_1$. The level of robustness and also the level of performance totally depend on the ranges of the defined sets. Tighter sets give the same performance as the optimal detector but gets sensitive to unwanted outliers. Usually in this method, it is required to be safe for all possible distributions within the sets. So, instead of using an online adaptive approach to find the best match within the sets with acquired data, a method that works for worst case scenario is desired. This concept is referred to as the minimax criteria. Mathematically, in minimax we are going to design a detector such that

$$\hat{q}_0, \hat{q}_1, \hat{\delta}_1 = \arg\min_{\delta \in \mathbf{\Delta}} \arg \max_{q_0 \times q_1 \in \mathcal{F}_0 \times \mathcal{F}_1} \mathbf{E}[Cost]. \tag{2.35}$$

In other words, we are looking for two distributions that maximizes the average cost value. Theses distributions are known as least favorable distributions (LFDs). Finding the LFDs is the critical step of this criterion. Because after finding the LFDs, the minimizer is the likelihood ratio test given LFDs. Our main focus in this thesis is designing a robust detector against deviations in noise distribution. However, the minimax can also be used for designing robust detectors against deviations in prior

probabilities $P_{\mathcal{H}_0}$ and $P_{\mathcal{H}_1}$ [37], interference subspace [69] and etc. Considering the $P_E(\boldsymbol{\delta}, f_0, f_1) = P_{\mathcal{H}_0} P_{FA}(\boldsymbol{\delta}, f_0) + P_{\mathcal{H}_1} P_M(\boldsymbol{\delta}, f_1)$ as the cost function, the solution of minimax criterion satisfies following inequalities

$$P_E(\hat{\boldsymbol{\delta}}, q_0, q_1) \leq P_E(\hat{\boldsymbol{\delta}}, \hat{q}_0, \hat{q}_1) \leq P_E(\boldsymbol{\delta}, \hat{q}_0, \hat{q}_1). \tag{2.36}$$

This condition is called as saddle value condition. This condition is not always met and its existence needs to be investigated. Von Neuman for the first time presented a formal definition for the existence of a saddle point [70]. For a robust binary hypothesis test it exists when

$$\min_{\delta \in \boldsymbol{\Delta}} \max_{q_0 \times q_1 \in \mathcal{F}_0 \times \mathcal{F}_1} P_E(\boldsymbol{\delta}, q_0, q_1) = \max_{q_0 \times q_1 \in \mathcal{F}_0 \times \mathcal{F}_1} \min_{\delta \in \boldsymbol{\Delta}} P_E(\boldsymbol{\delta}, q_0, q_1), \tag{2.37}$$

where P_E is needed to be bilinear in all its arguments. Shiftman showed that for the existence of saddle value, it is sufficient that P_E be convex in $\boldsymbol{\delta}$ and concave in q_0 and q_1 [71]. Beside the property of LFDs which maximizes the error probability, they are also the minimizers of the $f-$divergence between distributions from two hypotheses. In other words, if a pair of distributions (q_0, q_1) minimizes

$$D_f(q_1(\mathbf{y}) \| q_0(\mathbf{y})) = \int f\left(\frac{q_1(\mathbf{y})}{q_0(\mathbf{y})}\right) q_0(\mathbf{y}) d\mathbf{y}, \tag{2.38}$$

for all $q_0 \times q_1 \in \mathcal{F}_0 \times \mathcal{F}_1$, and all twice differentiable convex functions $f : \mathbb{R}_+ \to \mathbb{R}$, then the joint distributions (q_0^N, q_1^N) are the LFDs for all cost functions [72].
Typically, three well known uncertainty sets have been introduced in literature: 1. $\epsilon-$contamination model 2. divergence ball model 3. density band model which are described in the following sections.

2.5.1.1 $\epsilon-$Contamination Model

The earliest work in robust detection has been done by Huber [45] after observing sensitivity of LRT to bad data samples. In design steps, the uncertainty set was introduced by an $\epsilon-$contamination model. Indeed, the neighborhood around the nominal model defined by

$$\mathcal{F}_i = \{q_i | q_i = (1 - \epsilon_i) f_i + \epsilon_i h_i\}, \ i = 0, 1, \tag{2.39}$$

where $0 \leq \epsilon_i < 1$ and h_i can be any arbitrary density. Assuming that $\mathcal{F}_0$ and $\mathcal{F}_1$ are distinct, the LFDs are given by

$$\hat{q}_0 = \begin{cases} (1-\epsilon_0)f_0, & f_1/f_0 < c_u \\ (1/c_u)(1-\epsilon_0)f_1 & f_1/f_0 \geq c_u \end{cases}, \quad \hat{q}_1 = \begin{cases} (1-\epsilon_1)f_1, & f_1/f_0 > c_l \\ (c_l)(1-\epsilon_1)f_0 & f_1/f_0 \leq c_l \end{cases} \tag{2.40}$$

where c_l and c_u can be uniquely defined. Finally the robust detector can be formed by $\hat{q}_1/\hat{q}_0$ which forms a clipped version of nominal LRT, i.e. for the case of $\epsilon_0 = \epsilon_1 = \epsilon$ the test is given by

$$T_{Huber}(\mathbf{Y}) = \prod_{i=1}^{N} \max\Big(c_l, \min\big(c_u, f_1(\mathbf{y}_i)/f_0(\mathbf{y}_i)\big)\Big). \tag{2.41}$$

This test suggests to replace high and small values of LRT by two constants of c_u and c_l to suppress the impact of unwanted samples.

Huber's $\epsilon-$contamination model extensively was used in other robust detectors where extra assumptions were added to the problem to make a more powerful test. For example in [73], the authors proposed two detectors for nominal Gaussian density, the correlator-limiter detector and the limiter-correlator detector, which were obtained as the result of two different formulations of the problem. The similar idea was also used to detect stochastic signals in a contaminated observations [74]. Other generalized detectors using contamination model can be found in [75–77].

2.5.1.2 Divergence Ball Model

Another measure which its specific cases were extensively used to represent the neighborhood around the nominal model is $f-$divergence ball model. $f-$divergence measures the difference between two given functions (in our problem: density functions). Indeed, the uncertainty set can be defined by

$$\mathcal{F}_i = \{q_i | D_f(q_i \| f_i) \leq \epsilon_i\}, \ i = 0, 1, \tag{2.42}$$

where $D_f(.\|.)$ is the $f-$divergence given by

$$D_f(q_i(\mathbf{y}) \| f_i(\mathbf{y})) = \int f\left(\frac{q_i(\mathbf{y})}{f_i(\mathbf{y})}\right) f_i(\mathbf{y}) d\mathbf{y}, \tag{2.43}$$

where f is a convex function and $f(1) = 0$. In this model uncertainty directly is defined on the shape of the densities and different $f : \mathbb{R}_+ \to \mathbb{R}$ functions can lead to different uncertainty sets with different properties. For example in [39, 78], Levy used the KL-divergence as a measure of tolerance between to densities. The work proposed in [79] generalized Levy's work by a general parametric divergence family of $\alpha-$divergence. Different measures and their corresponding detectors are discussed in details by [80].

2.5.1.3 Density Band Model

In the last uncertainty model, the densities in both hypotheses are bounded from above and below. The neighborhood is defined by [81]

$$\mathcal{F}_i = \{q_i | q_{l_i} \leq q_i \leq q_{u_i}\}, \ i = 0, 1, \tag{2.44}$$

where q_u is the upper density bound and q_l is the lower density bound. This model can also be achieved by putting constraints on contamination model [82], and the same solution can be achieved by constructing a proper convex function in $f-$divergence [83]. The band model is also used for non-parametric detectors [84].

2.5.2 Parametric Robust Detection

Experience shows that the reliability of MLEs like sample mean, sample covariance or least square, and the related designed GLR, Wald and Rao tests can be influenced by outliers. To address this issue in parametric form, robust statistics with pioneering works of Tukey [85], Huber [43] and Hampel [86] developed methods to reduce the influence of misspecified models. Before that, it was common to use outlier detection and rejection techniques that sometimes could even remove good samples which was not desired [10]. In robust parametric detection the goal is to estimate parameters in a way which after placing them in the test frameworks, it offers a stable false alarm rate and power level. Some recent works addressed this problem using more complex models with more number of parameters [87, 88] which are not in our interest. We prefer to use a simple model which can represent the behavior of most data but meanwhile the procedure be able to

TABLE 2.1: Suggested models by robust statistics for inference based on contamination level (ϵ) and prior information about outlier density (h) [10].

Task	h	$\epsilon = 1$	$0 \leq \epsilon \leq 1$	$\epsilon << 1$	$\epsilon = 0$
Robust Inference	Arbitrary h	**?**	**?**	f	f
	Specified h	h	g	f	f

tolerate slight deviations in modelling. Considering the $\epsilon-$contamination model of

$$g = (1 - \epsilon)f + \epsilon h, \tag{2.45}$$

in this thesis, based on the table 2.1 [10], we address the cases that ϵ is sufficiently small and we can directly work with the nominal model f.

Before introducing some robust estimators, we briefly discuss which characteristics of robust estimators are desired.

2.5.2.1 Criterion and Metrics in Robust Estimation

Generally in design steps of robust estimators, we are looking for satisfying following aims [9]:

1. **Near Optimality:** The estimator should have reasonably well (close to optimal) performance in nominal model. In other words, we are not interested to get significant bad performance in nominal case at the price of achieving robustness.
2. **Qualitative Robustness:** Small deviations in models should not have a large impact on the estimator.
3. **Quantitative Robustness:** Somewhat larger deviations from the model should not cause a catastrophe [89].

Now, assuming a few robust estimators are provided, one may ask how can we quantify, if our estimator fulfills these aims? To be able to compare different robust estimators, some

tools and measures are required. Following measures cover the mentioned requirements in a quantitative way:

1. **Relative Efficiency:** This shows the ratio of the optimal estimator's variance in nominal model to robust estimator's variance. The value close to 1, refers to an estimator which has almost the same performance as the optimal estimator and decreasing this ratio is in trade off with robustness level.

2. **The Breakdown Point (BP):** The breakdown point measures the robustness properties of an estimator in the global sense [10] and is the maximum fraction of data that is allowed to be contaminated with any arbitrary values, while still the estimator stays stable. This metric is actually measuring the worst case scenario and representing the quantitative robustness. For example, in a location estimation task, the sample mean has the breakdown point of 0% meaning even a single bad observation can completely lead the estimator to the point ∞ [40]. It is also easy to show the median and $\alpha-$trimmed estimators have the breakdown points of 50% and α%, respectively.

3. **The Influence Function (IF):** In exactly the opposite side of BP, based on Hampel's definition [86], the IF shows the local robustness of an estimator and is defined as the asymptotic changes in estimator when ϵ% of data gets contaminated (mainly with point mass distribution) divided by ϵ when $\epsilon \to 0$. This is the generalization of sensitivity curve (SC) and shows the sensitivity of an estimator to existing small deviations. It is always desired to have continuous and bounded IF function to guarantee that small changes in data always have a small and limited impact on the estimator. Again, for the location estimation task, the sample mean has an unbounded IF and median estimator has bounded but not continuous which make both estimators locally sensitive [10]. As it is clear, this metric measures the qualitative robustness of a given estimator.

4. **The Maximum Bias Curve (MBC):** This measure gives information on the bias affected by specific amount of contamination (ϵ) [40] where starts from small ϵ with the value of $\epsilon \sup \|IF(z;\boldsymbol{\omega})\|$ where z is the location of outliers using point mass density. This curve finally ends at the BP with the value of ∞ [40].

2.5.2.2 M-Estimators

M-estimators are a general family of estimators which generalizes the classical approach of MLEs [40, 90]. It is known that the MLEs can be derived by following maximization problem

$$\hat{\boldsymbol{\omega}}_{MLE} = \arg\max \sum_{i=1}^{N} \log f(\mathbf{y}_i; \boldsymbol{\omega}), \tag{2.46}$$

where $f(.)$ is the likelihood function. By taking the derivative, the MLEs are also the solution of

$$\sum_{i=1}^{N} \mathbf{s}(\mathbf{y}_i; \boldsymbol{\omega}) = \mathbf{0}, \tag{2.47}$$

where $\mathbf{s}(\mathbf{y}; \boldsymbol{\omega}) = \partial \log f(\mathbf{y}; \boldsymbol{\omega}) / \partial \boldsymbol{\omega}$. It was shown that the IF of MELs is directly related to the $\mathbf{s}(\mathbf{y}; \boldsymbol{\omega})$; this makes MLEs non robust due to their unbounded $\mathbf{s}$ functions [10]. M-estimators are offering loss functions that their derivatives, might possess desired properties of robust estimators (boundedness and continuity). Huber [43] proposed a class of M-estimators that generalizes the MLEs in a natural way by

$$\hat{\boldsymbol{\omega}} = \arg\min \sum_{i=1}^{N} \rho(\mathbf{y}_i; \boldsymbol{\omega}), \tag{2.48}$$

where $\rho(.)$ is a general loss function which the MLEs can be recovered by choosing $\rho(.) = -\log f(.)$. An alternative approach is to find the solution of

$$\sum_{i=1}^{N} \boldsymbol{\Psi}(\mathbf{y}_i, \boldsymbol{\omega}) = \mathbf{0}, \tag{2.49}$$

where $\boldsymbol{\Psi} = \partial \rho / \partial \boldsymbol{\omega}$. The IF of a M-estimator is given by [10]

$$IF(z; \hat{\boldsymbol{\omega}}, F) = \mathbf{M}^{-1} \boldsymbol{\Psi}(z; \hat{\boldsymbol{\omega}}), \tag{2.50}$$

where

$$\mathbf{M} = -\int \frac{\partial}{\partial \boldsymbol{\omega}} \boldsymbol{\Psi}(\mathbf{x}; \boldsymbol{\omega}) dF(\mathbf{x}). \tag{2.51}$$

It is clear that by bounding $\boldsymbol{\Psi}$ the IF will be bounded as well. Relatively simple M-estimators, called weighted MLE (WMLE) are defined as the solution of [10]

$$\hat{\boldsymbol{\omega}} = \sum_{i=1}^{N} w(\mathbf{y}_i; \boldsymbol{\omega}) \mathbf{s}(\mathbf{y}_i; \boldsymbol{\omega}) - a(\boldsymbol{\omega}) = \mathbf{0}, \tag{2.52}$$

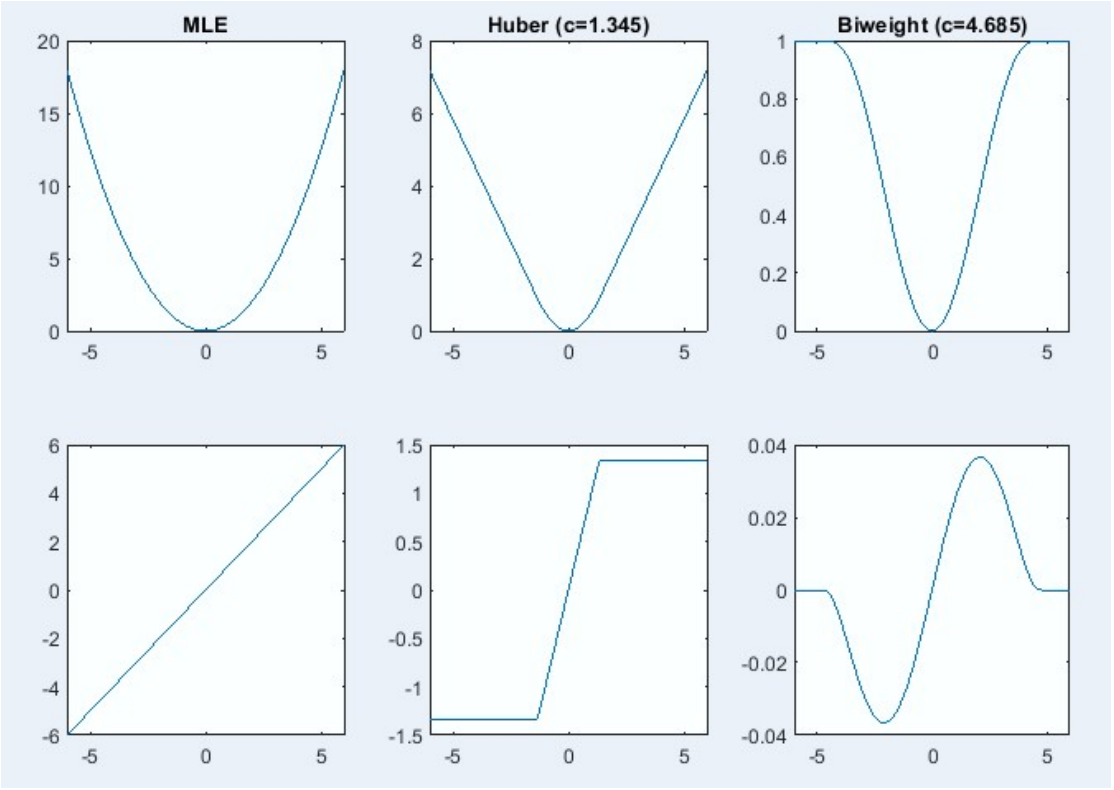

FIGURE 2.3: ρ functions (top) and $\boldsymbol{\Psi}$ functions (bottom) for the MLE, Huber and biweight estimators [10].

where $a(\boldsymbol{\omega})$ is a correction term. The idea is to limit unbounded score functions by multiplying a small weight value w. A popular form of this type of estimator is Huber's residual based weight. For example in a univariate case it is given by

$$w(r) = \min\left(1, c/|r|\right), \tag{2.53}$$

which intuitively tries to suppress high residual values and keep unchanged those residual values that are less than the hyperparameter c. The Huber $\boldsymbol{\Psi}$ function is given by

$$\boldsymbol{\Psi}(r) = \min\left(c, \max(-c, r)\right), \tag{2.54}$$

and its corresponding ρ function is

$$\rho(r) = \begin{cases} \frac{1}{2}r^2, & |r| \leq c \\ c|r| - \frac{1}{2}c^2, & |r| > c \end{cases} \tag{2.55}$$

which is a combination of quadratic loss function and l_1-norm. In the case of $c \to \infty$, the loss function is exactly the quadratic loss which can gives the MLEs for Gaussian

nominal density. On the other hand, in the case of $c \to 0$, the loss function is the l_1-norm which its robustness property was shown in literature. Figure 2.3 shows the shape of ρ function and their corresponding $\mathbf{\Psi}$ function when $c = 1.345$. As it is shown, the score function is bounded and continuous and the loss function is convex. This convexity makes the solution unique and using classical optimization algorithm, after a few steps the solution can be obtained. Nevertheless, it is more interested to have a $\mathbf{\Psi}$ function that tends to zero for large residuals to make the estimator less sensitive to heavy tail outliers. The Huber's estimator does not provide this feature and as the result has a small BP. One other famous estimator called biweight estimator proposed by Beaton and Tukey [91] increases the BP by its redescending $\mathbf{\Psi}$ function. The corresponding ρ function is given by

$$\rho(r) = \begin{cases} 3(\frac{r}{c})^2 - 3(\frac{r}{c})^4 + (\frac{r}{c})^6, |r| \leq c \\ 1, \qquad\qquad\qquad\quad |r| > c \end{cases}. \tag{2.56}$$

Figure 2.3 shows the ρ and $\mathbf{\Psi}$ functions when $c = 4.685$. As is shown, this loss function is not convex, meaning that using this loss function in M-estimation framework might have more than one local minimum. To avoid converging to a bad local minimum, the proper choice of initial point of the proposed iterative method is crucial. One suggestion is to start from a robust estimator with low efficiency but high BP [42]. After convergence, the efficiency will be improved based on the shape of ρ function and it's likely to converge to a suitable local minimum. Several robust estimators have been proposed in literature with different properties and different inspirations.

2.5.2.3 Developed Robust Tests

In the case of consistent M-estimation, it can be shown that $\hat{\boldsymbol{\omega}} \to \mathcal{N}(\boldsymbol{\omega}^*, \mathbf{V})$ where $\boldsymbol{\omega}^*$ is the true parameter value and $\mathbf{V}$ is the asymptotic variance given by

$$\mathbf{V} = \mathbf{M}^{-1}\mathbf{Q}\mathbf{M}^{-1}, \tag{2.57}$$

and

$$\mathbf{Q} = \int \mathbf{\Psi}(\mathbf{x};\boldsymbol{\omega})\mathbf{\Psi}^\top(\mathbf{x};\boldsymbol{\omega})dF(\mathbf{x}). \tag{2.58}$$

Using this fact, the robust version of LRT is given by [10]

$$LRT_\rho = 2\sum_{i=1}^{N}\left[\rho(\mathbf{y}_i;\hat{\boldsymbol{\omega}}_1) - \rho(\mathbf{y}_i;\hat{\boldsymbol{\omega}}_0)\right], \tag{2.59}$$

which is a generalization of classical GLRT. Similarly the Wald and Rao tests are given by

$$W_\rho \;=\; N(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_0)^\top \left[\mathbf{V}^{-1}(\hat{\boldsymbol{\omega}}_1)\right]_{\boldsymbol{\theta}\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_0), \tag{2.60}$$

and

$$R_\rho = N\mathbf{z}\mathbf{C}^{-1}\mathbf{z}, \tag{2.61}$$

where $\mathbf{z} = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial\rho}{\partial\boldsymbol{\theta}}(\mathbf{y}_i;\hat{\boldsymbol{\omega}}_0)$ and $\mathbf{C}$ is the asymptotic covariance of $\mathbf{z}$. It is known that both level and power of above tests are stable if the IF of corresponding function that test is defied based on it is bounded. This IF itself is bounded if the IF of estimator is bounded [10]. In other words, a given test is robust, only if the underlaying defined estimator is robust.

2.6 Conclusion

In this chapter, we have done an extensive review on the existing methods in both ML based and robust based estimation and detection. The materials and tools discussed there will be used in future chapters to analyze and compare the performance of our proposed methods.

Chapter 3

Robust Subspace Detectors Based on $\alpha-$Divergence with Application to Detection in Imaging

Robust variants of Wald, Rao and likelihood ratio (LR) tests for the detection of a signal subspace in a signal interference subspace corrupted by contaminated Gaussian noise are proposed in this chapter. They are derived using the $\alpha-$divergence, and the trade-off between the robustness and the power (the probability of detection) of the tests is adjustable using a single hyperparameter α. It is shown that when $\alpha \to 1$, these tests are equivalent to their well known classical counterparts. For example the robust LR test coincides with the LR test or the matched subspace detector (MSD) [49]. Asymptotic results are provided to support the proposed tests and robustness to outliers is obtained using values of $\alpha < 1$. Numerical experiments illustrating the performance of these tests on simulated, real functional magnetic resonance imaging (fMRI), hyperspectral and synthetic aperture radar (SAR) data are also presented.

3.1 Introduction

Detection of subspace signals in given corrupted observations, has wide applications in communications[92], radar [11], classification [93], medical imaging [16, 17] and pattern recognition [94]. In both Bayesian [39] and Neyman and Pearson [38] senses, the likelihood ratio test (LRT) is the uniformly most powerful (UMP) test and it can be used to obtain the maximum detection rate for a given false alarm rate when the distributions of data under both hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ are perfectly known. Unfortunately, the UMP test does not exist in many signal processing applications due to the uncertainties in modeling and parameter estimation approaches [40, 95]. Therefore, how to find suboptimal tests remains a topic of investigation [56].

The generalized likelihood ratio test (GLRT) is one suboptimal detector that has been widely investigated in different scenarios [6, 49–52, 96]. It is based on using maximum likelihood estimate's (MLE's) of the unknown parameters in the LRT. Furthermore, some other well known tests such as the Rao test [58, 59], the Wald test [55, 56] and their combinations [57, 97] have been investigated in the literature. In the Rao and Wald tests, parameters estimated under either the null hypothesis ($\mathcal{H}_0$) or the alternative hypothesis ($\mathcal{H}_1$) are used. For those of readers who are interested, we also suggest to check out the recent published works in this field and their references [98–101]. Typically, the design of these detectors incorporates key assumptions which can often make the tests sensitive to deviations from these assumptions. For example, Gaussianity of the noise is one very commonly used assumption. However, in many practical situations this assumption does not fully hold and contaminations such as impulsive noise have been observed in a variety of important real applications such as outdoor mobile communication channels, radar, sonar, medical imaging and indoor wireless communication channels [40]. Such contamination can significantly deteriorate the optimal performance of signal detection methods based on Gaussian noise assumption.

In this chapter, our main goal is to modify these existing tests such that robustness against small deviations in the noise model can be achieved. This problem has been studied in [45] where a clipped version of the LRT for the ϵ-contaminated density model was proposed. This test works based on the minimax idea and tries to find the least favorable densities. It then uses the LRT to maximize the power (probability of detection). This approach was adopted to derive a number of robust tests [73, 75, 82] by using different densities as the nominal model. The minimax approach was also used in

[78] where a robust test based on the Kullback-Leibler (KL) divergence to measure the uncertainty between the nominal and the least favorable distribution is proposed. This measure is a natural distance between two density functions, it is therefore a natural way to measure the existing contamination. The KL divergence was also used in the empirical likelihood ratio test [84] to propose a novel non-parametric test that handles uncertainties in the distributions achieving reasonable performance.

Recently, the authors in [79] used the $\alpha-$divergence as the similarity measure instead of KL divergence in the minimax framework to derive a test which is a nonlinear variant of the LRT. This measure is more general than the KL divergence and can be tuned by a single parameter α. It is equivalent to certain well known distances as special cases. For example when $\alpha = 2$, it is equivalent to $\mathcal{X}^2$ distance [102]. Due to the link between the KL and the MLE, it makes sense to use the $\alpha-$divergence instead of the KL divergence, to produce new robust detectors. A similar idea of using a density power divergence in a Wald test framework was adopted in [103–107].

To the best of our knowledge, $\alpha-$divergence has not been used so far in robust detection in the way adopted in this chapter. The proposed detectors are based on a robust estimation approach which is a member of the general family of M-estimators [90]. Here, we use the $\alpha-$divergence in the Wald, Rao and LRT frameworks, where instead of using MLE which leads us toward the classical detectors in the i.i.d Gaussian noise case, we use the robust $\alpha-$divergence based estimators to achieve robustness against contaminated Gaussian noise. The amount of robustness is adjustable using a single parameter α at the expense of the detector power under the nominal model. The main idea underlying the derivation of the proposed detectors is to limit the effects of large amplitude residuals by assigning them small weights. These weights are obtained from an exponential function of the residuals and α. When $\alpha \to 1$, the proposed methods are equivalent to the well known maximum likelihood (ML) based tests and all the weights are equal to 1, i.e., there is no outliers rejection and the estimator is the same as the MLE. For other choices of $\alpha < 1$, the proposed estimator will be sensitive to high amplitude residuals and the associated selectivity power, will reject outliers to give more accurate estimation of the true parameters. This idea first was introduced in [2] and it is similar to [108] but different in the form of weight function. Their method minimizes the p-norm of residuals with respect to unknown parameters and the final estimator is a weighted least square. Furthermore, the choice of the function used for the weights is justified

in the tests proposed in this chapter. The main contributions of this chapter may be summarized as follows:

- We consider the linear subspace model with interference and contaminated Gaussian noise for a multi-rank subspace target signal to model the observations. Then we make a connection between the ML and KL divergence to give insight into how the $\alpha-$divergence can be used for estimation by using the $\alpha-$logarithm function instead of the log-likelihood function.
- New Robust versions of well known Wald, Rao and LR tests with a tunable parameter are proposed. These tests include their classical form as a special case and offer robustness against contaminated Gaussian noise.
- The asymptotic analysis of the proposed tests is provided and their distributions under the null hypothesis are derived.
- The influence function (IF) of the estimator and the tests are discussed to show the local robustness.

The remainder of this chapter is organized as follows. Section 3.2 briefly describes the problem. Section 3.3 contains background on the GLRT, Wald and Rao tests, ML and α-divergence. In section 3.4 the robust $\alpha-$divergence based tests are introduced and in section 3.5 robustness properties of the methods are presented. The simulation and experimental results are presented in section 3.6 and 3.7 and finally the chapter is concluded in section 3.8.

3.2 Problem Formulation

The detection problem that we are studying in this chapter is described as follows. We are given N samples from a real and scalar time series $y_i, i = 1, ..., N$ that is represented by a column vector $\mathbf{y}$. We assume $\mathbf{y}$ obeys the general linear subspace model (GLM)[49]

$$\mathbf{y} = \underbrace{\mathbf{H}\boldsymbol{\theta}}_{\mathbf{x}} + \underbrace{\mathbf{B}\boldsymbol{\phi}}_{\mathbf{i}} + \boldsymbol{\xi}, \tag{3.1}$$

where $\mathbf{y} \in \mathbb{R}^N$ is the measurement, $\mathbf{H} \in \mathbb{R}^{N\times p}$ is a known matrix whose columns span the subspace containing the target signal $\mathbf{x}$ and $\mathbf{h}_i \in \mathbb{R}^p$ is the i^{th} row of the matrix $\mathbf{H}$. $\boldsymbol{\xi}$ is a contaminated Gaussian noise where ξ_i follows the contaminated model $(1-\epsilon)\mathcal{N}(0,\sigma^2)+\epsilon h$ where h is an arbitrary unknown density which models the outliers and σ^2 is the unknown variance of the nominal model. In this model, $\mathbf{i} = \mathbf{B}\boldsymbol{\phi} \in \mathbb{R}^N$ is the interference component, $\mathbf{b}_i \in \mathbb{R}^t$ is the i^{th} row of the matrix $\mathbf{B} \in \mathbb{R}^{N\times t}$. The parameter vectors $\boldsymbol{\theta} \in \mathbb{R}^p$ and $\boldsymbol{\phi} \in \mathbb{R}^t$ are the unknown coordinates for $\mathbf{H}$ and $\mathbf{B}$. Note that the subspaces $\langle\mathbf{H}\rangle$ and $\langle\mathbf{B}\rangle$ are assumed to be linearly independent [49]. Defining $\mathbf{C} = [\mathbf{H},\mathbf{B}]$ and $\mathbf{c}_i$ as the i^{th} row of $\mathbf{C}$, (3.1) takes the form

$$\mathbf{y} = \mathbf{C}\boldsymbol{\beta} + \boldsymbol{\xi},$$

where $\boldsymbol{\beta}^\top = [\boldsymbol{\theta}^\top, \boldsymbol{\phi}^\top]^\top$. The full vector collecting the unknown parameters $\boldsymbol{\omega}$ is defined as

$$\boldsymbol{\omega} = [\boldsymbol{\beta}^\top, \sigma^2]^\top = [\boldsymbol{\theta}^\top, \boldsymbol{\phi}^\top, \sigma^2]^\top = [\boldsymbol{\theta}^\top, \boldsymbol{\lambda}^\top]^\top, \tag{3.2}$$

and the detection problem can be described as distinguishing between the two following hypotheses

$$\mathcal{H}_0 : \boldsymbol{\theta} = \mathbf{0}, \boldsymbol{\lambda}_0, \quad \text{vs} \quad \mathcal{H}_1 : \boldsymbol{\theta} \neq \mathbf{0}, \boldsymbol{\lambda}_1, \tag{3.3}$$

where $\boldsymbol{\lambda}$ contains all of the nuisance parameters.

3.3 Background

In this section, the well known classical detectors or tests (GLRT, Wald test and Rao test) are first reviewed, then the connection between the ML estimation approach and estimation by minimizing the $\alpha-$divergence is described.

3.3.1 Generalized Likelihood Ratio Test

Due to optimality of the likelihood ratio test, the GLRT can be seen as the first alternative when the exact model is not available. This generalization of the LRT is obtained

by replacing the parameters with their MLEs and is expressed as

$$T_G(\mathbf{y}) = \frac{2}{N}[\sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_1}\sum_{i=1}^{N}\log\left(f(y_i,\boldsymbol{\omega})\right) - \sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_0}\sum_{i=1}^{N}\log\left(f(y_i,\boldsymbol{\omega})\right)] \geq \gamma, \tag{3.4}$$

where $\log(f(y_i,\boldsymbol{\omega}))$ is the log-likelihood function and γ is the corresponding threshold for a given probability of false alarm (P_{FA}). $\boldsymbol{\Omega}_i$, $i=0,1$, defines the possible parameter set for each hypothesis. Using a Gaussian assumption on the noise, the GLRT takes the form

$$\begin{aligned} T_{G1}(\mathbf{y}) &= \frac{2}{N}\sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_1}\sum_{i=1}^{N}\left\{\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}(y_i-\mathbf{c}_i\boldsymbol{\beta})^2\right\} \\ &\quad -\frac{2}{N}\sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_0}\sum_{i=1}^{N}\left\{\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}(y_i-\mathbf{b}_i\boldsymbol{\phi})^2\right\} \\ &= 2\left[\log\left(\frac{1}{\sqrt{2\pi\hat{\sigma}_1^2}}\right) - \log\left(\frac{1}{\sqrt{2\pi\hat{\sigma}_0^2}}\right)\right] \geq \gamma, \end{aligned}$$

where the parameters are replaced by their MLE values. With some simplifications and using the monotone increasing function $T_{G2}(\mathbf{y}) = \left(\exp\left\{\frac{T_{G1}(\mathbf{y})}{2}\right\}\right)^2 - 1$, we get the well known matched subspace detector (MSD) defined as

$$T_{G2}(\mathbf{y}) = \frac{\hat{\sigma}_0^2-\hat{\sigma}_1^2}{\hat{\sigma}_1^2} = \frac{\mathbf{y}^\top\mathbf{P}_{\mathbf{B}^\perp}\mathbf{P}_{\mathbf{G}}\mathbf{P}_{\mathbf{B}^\perp}\mathbf{y}}{\mathbf{y}^\top\mathbf{P}_{\mathbf{B}^\perp}\mathbf{P}_{\mathbf{G}^\perp}\mathbf{P}_{\mathbf{B}^\perp}\mathbf{y}} \geq \gamma' = \exp\left\{\frac{\gamma}{2}\right\}^2 - 1, \tag{3.5}$$

where $\mathbf{G} = \mathbf{P}_{\mathbf{B}^\perp}\mathbf{H}$, $\mathbf{P}_{\mathbf{D}} = \mathbf{D}(\mathbf{D}^\top\mathbf{D})^{-1}\mathbf{D}^\top$ is the projection matrix and the projector $\mathbf{P}_{\mathbf{B}^\perp}$ projects onto the subspace $\langle\mathbf{B}\rangle^\perp$.

3.3.2 Wald and Rao Tests

GLRT deals with estimated parameters in both $\mathcal{H}_0$ and $\mathcal{H}_1$ which makes it computationally expensive. One way to reduce this complexity is to work with only one set of the parameters, either $\boldsymbol{\omega}\in\boldsymbol{\Omega}_1$ or $\boldsymbol{\omega}\in\boldsymbol{\Omega}_0$. This is the idea that is adopted in Wald and Rao tests which are explicitly defined as

$$T_W(\mathbf{y}) = \hat{\boldsymbol{\theta}}_1^\top\left\{\left[\mathbf{I}^{-1}(\hat{\boldsymbol{\omega}}_1)\right]_{\boldsymbol{\theta}\boldsymbol{\theta}}\right\}^{-1}\hat{\boldsymbol{\theta}}_1, \tag{3.6}$$

where $\hat{\boldsymbol{\omega}}_1 = [\hat{\boldsymbol{\theta}}_1^\top, \hat{\boldsymbol{\phi}}_1^\top, \hat{\sigma}_1^2]^\top$ is the parameter vector defined in (3.2) estimated under $\mathcal{H}_1$ and

$$T_R(\mathbf{y}) = \frac{\partial \log(f_1)}{\partial \boldsymbol{\theta}}\bigg|^\top_{\boldsymbol{\omega}=\hat{\boldsymbol{\omega}}_0} \left[\mathbf{I}^{-1}(\hat{\boldsymbol{\omega}}_0)\right]_{\boldsymbol{\theta}\boldsymbol{\theta}} \frac{\partial \log(f_1)}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\omega}=\hat{\boldsymbol{\omega}}_0} \tag{3.7}$$

where f_1 is the joint probability density under $\mathcal{H}_1$, $\hat{\boldsymbol{\omega}}_0 = [\mathbf{0}^\top, \hat{\boldsymbol{\phi}}_0^\top, \hat{\sigma}_0^2]^\top$ is the parameter vector defined in (3.2) estimated under $\mathcal{H}_0$ and $\mathbf{I}(\boldsymbol{\omega})$ is the associated Fisher information matrix

$$[\mathbf{I}(\boldsymbol{\omega})]_{ij} = -\mathbf{E}\left[\frac{\partial^2 \log(f_1)}{\partial \boldsymbol{\omega}_i \partial \boldsymbol{\omega}_j}\right], \tag{3.8}$$

which can be partitioned as

$$\mathbf{I}(\boldsymbol{\omega}) = \begin{bmatrix} \mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\theta}} & \mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\lambda}} \\ \mathbf{I}_{\boldsymbol{\lambda}\boldsymbol{\theta}} & \mathbf{I}_{\boldsymbol{\lambda}\boldsymbol{\lambda}} \end{bmatrix}. \tag{3.9}$$

$\left[\mathbf{I}^{-1}(\hat{\boldsymbol{\omega}}_0)\right]_{\boldsymbol{\theta}\boldsymbol{\theta}}$ in the Rao test is given by

$$\left[\mathbf{I}^{-1}(\boldsymbol{\omega})\right]_{\boldsymbol{\theta}\boldsymbol{\theta}} = \left(\mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\theta}} - \mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\lambda}}\mathbf{I}_{\boldsymbol{\lambda}\boldsymbol{\lambda}}^{-1}\mathbf{I}_{\boldsymbol{\lambda}\boldsymbol{\theta}}\right)^{-1}, \tag{3.10}$$

evaluated at $\hat{\boldsymbol{\omega}}_0$. Note that all three defined classical tests are for the case of $\epsilon = 0$, where the exact model of density function is available and corresponds to the nominal model. Specific compact forms of Wald and Rao tests for model (3.1) are defined in remarks 1 and 2.

3.3.3 Maximum Likelihood to Minimum $\alpha-$Divergence

Given i.i.d. data samples $\mathbf{y}_1, ..., \mathbf{y}_N$, where $\mathbf{y}_i$ is sampled from a probability distribution $F(\mathbf{y}, \boldsymbol{\omega}^*)$ with $\boldsymbol{\omega}^* \in \boldsymbol{\Omega} \subset \mathbb{R}^m$ and $\{F(\mathbf{y}, \boldsymbol{\omega}), \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$ is a family of approximating probability distributions indexed by $\boldsymbol{\omega}$. The ML estimate is the one that maximizes the log-likelihood function or equivalently for a large number of data points, minimizes the KL divergence between the true and the parametric approximate densities and is given by

$$\begin{aligned} \hat{\boldsymbol{\omega}}_{ML} &= \arg\max_{\boldsymbol{\omega}} \frac{1}{N}\sum_{i=1}^{N} \log\left(f(\mathbf{y}_i, \boldsymbol{\omega})\right) \\ &= \arg\min_{\boldsymbol{\omega}} KL\left(f(\mathbf{y}, \boldsymbol{\omega}^*), f(\mathbf{y}, \boldsymbol{\omega})\right), \end{aligned} \tag{3.11}$$

where $f(\mathbf{y}, \boldsymbol{\omega}^*)$ is the true density function. Now consider another situation where the data comes from another distribution which is in the neighborhood of $f(\mathbf{y}, \boldsymbol{\omega}^*)$ to account for the presence of outliers. The KL divergence is sensitive to outliers and a slight untraceable deviation in the true distribution can significantly affect the estimated parameters and increase the error between the true and the estimated values. As an alternative to improve such errors, a general parametric form of divergence can be employed which in a specific case can be reduced to the KL divergence. Such measures of similarity exist in the class of divergences referred to as α-divergence [109][102] and can be derived from both general families of $f-$divergence and Bregman divergence [110, 111]. For two probability density functions $g(\mathbf{y}, \boldsymbol{\omega}^*)$ defined as $(1-\epsilon)f(\mathbf{y}, \boldsymbol{\omega}^*) + \epsilon h(\mathbf{y})$ (where $f(\mathbf{y}, \boldsymbol{\omega}^*)$ is the true nominal model) and $f(\mathbf{y}, \boldsymbol{\omega})$, the α-divergence is defined as [102]

$$D_\alpha \left(g(\mathbf{y}, \boldsymbol{\omega}^*) \parallel f(\mathbf{y}, \boldsymbol{\omega}) \right) = \frac{1}{\alpha(\alpha-1)} \left[\int g(\mathbf{y}, \boldsymbol{\omega}^*)^\alpha f(\mathbf{y}, \boldsymbol{\omega})^{1-\alpha} d\mathbf{y} - 1 \right]. \tag{3.12}$$

This generalization defines a class of divergences indexed by a single parameter α. When $\alpha \to 1$, it reduces to the KL divergence and the method of ML. This can be shown by deriving D_α in terms of the $\alpha-$logarithm function [102, 112]

$$\log_\alpha (x) = \frac{x^{(1-\alpha)} - 1}{1-\alpha}, \tag{3.13}$$

where $\log_1 (x) = \log (x)$. Compared to the KL divergence, this divergence provides protection against contaminated Gaussian densities which happens in many signal processing applications [40].

It has been shown that minimizing the $\alpha-$divergence D_α with respect to $\boldsymbol{\omega}$ is equivalent to the following maximization with the empirical distribution [112]

$$\hat{\boldsymbol{\omega}} \quad = \quad \arg\max_{\omega} \frac{1}{N} \sum_{i=1}^{N} \log_\alpha \left\{ f(\mathbf{y}_i, \boldsymbol{\omega}) \right\}. \tag{3.14}$$

Therefore as an alternative we can use the α-logarithmic likelihood function $\log_\alpha \{f(\mathbf{y}_i, \boldsymbol{\omega})\}$ instead of the traditional log-likelihood function for estimation purpose. This logarithmic function has the advantage of being less sensitive to outliers because it penalizes small likelihood values to a lower extent. This is illustrated in figure 3.1. As a result, the estimation is concentrated on the bulk of the data ignoring the effect of outliers which is the main goal of robust statistics.

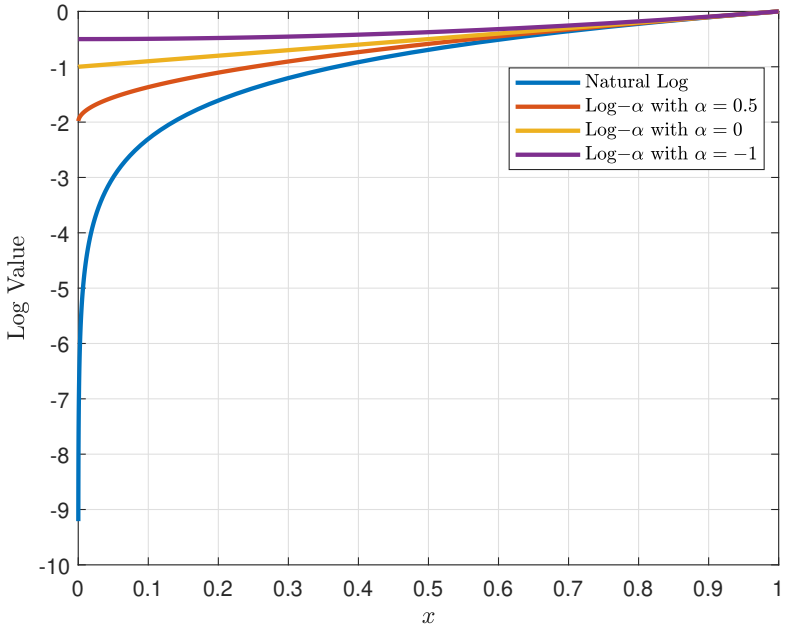

FIGURE 3.1: $\alpha-$logarithm(x) for different values of α.

3.4 Proposed Tests

3.4.1 Robust Wald Test

In the case of model (3.1), (3.14) takes the form

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \arg\max_{\beta} \frac{1}{N}\sum_{i=1}^{N} \log_\alpha \{f(y_i, \boldsymbol{\beta})\} \\ &= \arg\max_{\beta} \frac{1}{N}\sum_{i=1}^{N} \ell_\alpha (v_i) \\ &= \arg\max_{\boldsymbol{\beta}} L_\alpha(\mathbf{y}, \boldsymbol{\beta}). \end{aligned} \tag{3.15}$$

for $v_i \in \mathbb{R}$, where $v_i = \frac{\xi_i}{\sigma}$ and $\xi_i = y_i - \mathbf{c}_i\boldsymbol{\beta}$ with ℓ_α is defined as

$$\ell_\alpha(v) = \frac{1}{\tau}\left(\exp\left(\frac{-\tau v^2}{2}\right) - 1\right). \tag{3.16}$$

The expression for ℓ_α is obtained by considering a centered Gaussian nominal density [112]

$$f(y_i, \boldsymbol{\beta}, \sigma) \propto \exp\left(-\frac{(y_i - \mathbf{c}_i\boldsymbol{\beta}))^2}{2\sigma^2}\right), \tag{3.17}$$

in (3.13) and taking $\tau = 1 - \alpha$. If $\tau \to 0$, we have $\ell_\alpha(v) \to -v^2/2$, and the familiar quadratic loss is recovered. This quadratic loss is however very sensitive to outliers and offers no protection against anomalous observations frequently encountered in signal detection applications. The case $\tau > 0$, corresponds to a robust estimation procedure that tends to down weight errors that are far from the nominal density, thus removing outliers. The algorithm that maximizes the criterion (3.15) is an iterative algorithm which depends on an initial estimate. We suggest starting from some available robust estimators such as minimizers of ℓ_1 or Huber loss functions [42].The score function is then given by

$$\begin{aligned} \mathbf{d}(\mathbf{y}, \boldsymbol{\beta}) &= \nabla_{\boldsymbol{\beta}} L_\alpha \\ &= \frac{1}{N\sigma^2} \sum_{i=1}^{N} w_i(y_i, \boldsymbol{\beta}, \sigma)(y_i - \mathbf{c}_i\boldsymbol{\beta})\mathbf{c}_i^\top, \end{aligned} \tag{3.18}$$

where $w_i(y_i, \boldsymbol{\beta}, \sigma) = \exp\left(-\frac{\tau}{2\sigma^2}\left(y_i - \mathbf{c}_i\boldsymbol{\beta}\right)\right)^2\Big)$ and the Hessian is given by

$$\begin{aligned} \mathbf{T}(\mathbf{y}, \boldsymbol{\beta}) &= \nabla_{\boldsymbol{\beta}} \mathbf{d}(\mathbf{y}, \boldsymbol{\beta}) \\ &= \frac{-1}{N\sigma^2} \sum_{i=1}^{N} w_i(y_i, \boldsymbol{\beta}, \sigma) \left(1 - \frac{\tau}{\sigma^2}\left(y_i - \mathbf{c}_i\boldsymbol{\beta}\right)^2\right) \mathbf{c}_i^\top \mathbf{c}_i. \end{aligned} \tag{3.19}$$

Henceforth σ is assumed known or estimated using a consistent robust estimator; for example using the weighted sample empirical estimator

$$\hat{\sigma}^2 = \sum_{i=1}^{N} \tilde{w}_i(y_i, \hat{\boldsymbol{\beta}}, \sigma) \left(y_i - \mathbf{c}_i\hat{\boldsymbol{\beta}}\right)^2,$$

where $\tilde{w}_i(y_i, \hat{\boldsymbol{\beta}}, \sigma) = w_i(y_i, \hat{\boldsymbol{\beta}}, \sigma) / \sum_{i=1}^{N} w_i(y_i, \hat{\boldsymbol{\beta}}, \sigma)$ is derived from

$$\nabla_\sigma \tilde{L}_\alpha \left(y_i, \hat{\boldsymbol{\beta}}, \sigma\right),$$

with

$$\tilde{L}_\alpha(y_i, \boldsymbol{\beta}, \sigma) = \frac{1}{N\tau} \sum_{i=1}^{N} \left(\frac{1}{\sigma^\tau} \exp\left(-\frac{\tau}{2\sigma^2}\left(y_i - \mathbf{c}_i\boldsymbol{\beta}\right)\right)^2\Big) - 1\right).$$

This last expression is obtained by considering the constant term in expression (3.17) when (3.13) is used.

Define $\boldsymbol{\beta}^* = \arg\max_\beta \mathbf{E}\left[\ell_\alpha(y, \boldsymbol{\beta}, \sigma)\right]$, then based on the results in [113], if $\mathbf{C}$ has full column rank and $\hat{\boldsymbol{\beta}}$ is unique for each N, then $\hat{\boldsymbol{\beta}} \to_p \boldsymbol{\beta}^*$ as $N \to \infty$. Let σ^2 be known

or estimated consistently then,

$$\sqrt{N}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^{*}\right)\rightarrow_{d}\mathcal{N}(0,\mathbf{K}^{-1}\mathbf{L}\mathbf{K}^{-1})\text{ as }N\rightarrow\infty,$$

where $\mathbf{K}=\mathbf{E}\left[\mathbf{T}(y,\boldsymbol{\beta})|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}\right]$ and $\mathbf{L}=\lim_{N\rightarrow\infty}\text{Var}\left(\sqrt{N}\mathbf{d}(y,\hat{\boldsymbol{\beta}})\right)$.
To decide between $\mathcal{H}_0:\boldsymbol{\theta}=\mathbf{0}$ (null hypothesis) and $\mathcal{H}_1:\boldsymbol{\theta}\neq\mathbf{0}$ (alternative hypothesis), in this section, we use the Wald test approach. Note that under the assumptions above and defining $\mathbf{V}=\mathbf{K}^{-1}\mathbf{L}\mathbf{K}^{-1}$, the proposed robust Wald type test is defined by

$$T_{W_r}(\mathbf{y})=N\hat{\boldsymbol{\theta}}^{\top}\mathbf{R}^{-1}\hat{\boldsymbol{\theta}}. \tag{3.20}$$

where $\mathbf{R}$ is the corresponding covariance matrix of $\hat{\boldsymbol{\theta}}$. If we write the covariance of $\boldsymbol{\beta}$ as

$$\mathbf{V}=\begin{bmatrix}\mathbf{V}_{\boldsymbol{\theta}\boldsymbol{\theta}} & \mathbf{V}_{\boldsymbol{\theta}\boldsymbol{\phi}}\\ \mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\theta}} & \mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}\end{bmatrix},$$

then based on the matrix inversion lemma we have the inverse covariance of $\hat{\boldsymbol{\theta}}$ given by

$$\mathbf{R}^{-1}=\left(\mathbf{V}_{\boldsymbol{\theta}\boldsymbol{\theta}}-\mathbf{V}_{\boldsymbol{\theta}\boldsymbol{\phi}}\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}^{-1}\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\theta}}\right)^{-1}.$$

Under the null hypothesis, based on the central limit theorem, asymptotically we have

$$T_{W_r}(\mathbf{y})=N\hat{\boldsymbol{\theta}}^{\top}\mathbf{R}^{-1}\hat{\boldsymbol{\theta}}\rightarrow_{d}\chi_{p}^{2}. \tag{3.21}$$

and under the contiguous alternative ($\mathcal{H}_1:\boldsymbol{\theta}=\boldsymbol{\theta}^{*}=\boldsymbol{\delta}/\sqrt{N}$ where $\boldsymbol{\delta}$ is any vector in $\mathbb{R}^p$), $T_{W_r}\rightarrow_{d}\chi_{p}^{2}(\lambda)$ where $\lambda=N\boldsymbol{\theta}^{*\top}\mathbf{R}^{-1}\boldsymbol{\theta}^{*}$, so the test with an appropriate threshold value can reject $\mathcal{H}_0$.
Note that the above test is a special case of the general form $\mathcal{H}_0:\mathbf{A}\boldsymbol{\theta}=\mathbf{r}$ where $\mathbf{A}$ is a $k\times p$ (with $k\leq p$) full rank matrix and $\mathbf{r}$ is a $k\times 1$ vector. The test becomes

$$T_{W_r}(\mathbf{y})=N\left(\mathbf{A}\hat{\boldsymbol{\theta}}-\mathbf{r}\right)^{\top}\left(\mathbf{A}\mathbf{R}\mathbf{A}^{\top}\right)^{-1}\left(\mathbf{A}\hat{\boldsymbol{\theta}}-\mathbf{r}\right). \tag{3.22}$$

Under $\mathcal{H}_0$, $T_{W_r}\rightarrow_{d}\chi_{k}^{2}$ and when $\mathcal{H}_0$ is false T_{W_r} will become larger. Therefore $\mathcal{H}_0$ is rejected when T_{W_r} is larger than the appropriate quantile of chi-square distribution with k degrees of freedom.
Asymptotically, the proposed robust test is constant false alarm (CFAR). Based on the

asymptotic distribution of the test, the false alarm value only depends on the threshold value and k and is independent of σ.

Remark 1: When $\alpha \to 1$, $\mathbf{A} = \mathbf{I}_p$, $\mathbf{r} = \mathbf{0}$, the proposed test (3.22) reduces to the classical Wald test:

$$T_W(\mathbf{y}) = \frac{\hat{\boldsymbol{\theta}}_{ML}^{\top}\left(\mathbf{H}^{\top}\mathbf{P}_{\mathbf{B}^{\perp}}\mathbf{H}\right)\hat{\boldsymbol{\theta}}_{ML}}{\hat{\sigma}_1^2}.$$

3.4.2 Robust Rao Test

In this test, starting from (3.15)

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \arg\max_{\beta} \frac{1}{N}\sum_{i=1}^{N} \log_{\alpha}\left\{f(y_i,\boldsymbol{\beta})\right\} \\ &= \arg\max_{\beta} \frac{1}{N}\sum_{i=1}^{N} \ell_{\alpha}(y_i,\boldsymbol{\beta}) \\ &= \arg\max_{\boldsymbol{\beta}} L_{\alpha}(\mathbf{y},\boldsymbol{\beta}),\end{aligned} \tag{3.23}$$

as an alternative we can use the α-logarithmic likelihood function $L_{\alpha}(\mathbf{y},\boldsymbol{\beta})$ instead of the traditional log-likelihood function to propose a robust detector. The proposed variant of the Rao test is based on the following proposition.

Proposition 1: Define $\boldsymbol{\beta}^* = \arg\max_{\boldsymbol{\beta}} \mathbf{E}\left[\log_{\alpha}\left\{f(\mathbf{y},\boldsymbol{\beta})\right\}\right]$ and assume asymptotically $\hat{\boldsymbol{\beta}} \to_p \boldsymbol{\beta}^*$ as $N \to \infty$, then under certain regularity conditions asymptotically $\left.\frac{\partial L_{\alpha}(\mathbf{y},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} \to \mathcal{N}(\mathbf{0},\mathbf{F}(\hat{\boldsymbol{\beta}}))$, where $\mathbf{F}(\boldsymbol{\beta}) = \mathbf{E}\left[\frac{1}{N}\frac{\partial \ell_{\alpha}(\mathbf{y},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\frac{\partial \ell_{\alpha}(\mathbf{y},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}^{\top}\right]$ is the newly defined $\alpha-$information matrix.

Proof: See Appendix A.

Using proposition 1 and by considering the likelihood of the i^{th} observation as f_i

$$f_i(y_i,\boldsymbol{\beta},\sigma) \propto \exp\left(-\frac{(y_i-\mathbf{c}_i\boldsymbol{\beta})^2}{2\sigma^2}\right), \tag{3.24}$$

The α-information matrix $\mathbf{F}(\boldsymbol{\beta})$ can be expressed as

$$[\mathbf{F}(\boldsymbol{\beta})]_{ij} = \mathbf{E}\left[\frac{1}{N}\frac{\partial \ell_{\alpha}(y,\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_i}\frac{\partial \ell_{\alpha}(y,\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_j}\right], \tag{3.25}$$

which is obtained by replacing the log-likelihood function with $l_\alpha(y, \boldsymbol{\beta})$. If $\alpha \to 1$, this information matrix reduces to the well known Fisher information matrix. The α-information matrix can be defined as

$$\mathbf{F}(\boldsymbol{\beta}) = \begin{bmatrix} \mathbf{F}_{\boldsymbol{\theta}\boldsymbol{\theta}} & \mathbf{F}_{\boldsymbol{\theta}\boldsymbol{\phi}} \\ \mathbf{F}_{\boldsymbol{\phi}\boldsymbol{\theta}} & \mathbf{F}_{\boldsymbol{\phi}\boldsymbol{\phi}} \end{bmatrix}. \tag{3.26}$$

Then the proposed Rao test variant is

$$T_{R_r}(\mathbf{y}) = \frac{\partial L_\alpha(\mathbf{y}, \boldsymbol{\beta})}{\partial \boldsymbol{\theta}}\bigg|^{\top}_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_0} \left[\mathbf{F}^{-1}(\hat{\boldsymbol{\beta}}_0)\right]_{\boldsymbol{\theta}\boldsymbol{\theta}} \frac{\partial L_\alpha(\mathbf{y}, \boldsymbol{\beta})}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_0}, \tag{3.27}$$

where $\hat{\boldsymbol{\beta}}_0$ is the estimation of $\boldsymbol{\beta}$ under $\mathcal{H}_0$ and $\left[\mathbf{F}^{-1}(\hat{\boldsymbol{\beta}}_0)\right]_{\boldsymbol{\theta}\boldsymbol{\theta}}$ given by

$$\left[\mathbf{F}^{-1}(\hat{\boldsymbol{\beta}})\right]_{\boldsymbol{\theta}\boldsymbol{\theta}} = \left(\mathbf{F}_{\boldsymbol{\theta}\boldsymbol{\theta}} - \mathbf{F}_{\boldsymbol{\theta}\boldsymbol{\phi}}\mathbf{F}^{-1}_{\boldsymbol{\phi}\boldsymbol{\phi}}\mathbf{F}_{\boldsymbol{\phi}\boldsymbol{\theta}}\right)^{-1}, \tag{3.28}$$

evaluated at $\hat{\boldsymbol{\beta}}_0$. Using (3.23) and (3.1), we have

$$\frac{\partial L_\alpha(\mathbf{y}, \boldsymbol{\beta})}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_0} = \frac{\mathbf{u}}{N\hat{\sigma}_0^2}, \tag{3.29}$$

where $\mathbf{u} = \sum_{i=1}^{N} w_i(y_i, \hat{\boldsymbol{\phi}}_0, \hat{\sigma}_0)\left(y_i - \mathbf{b}_i\hat{\boldsymbol{\phi}}_0\right)\mathbf{h}_i^{\top}$, $\tau = 1 - \alpha$ and $w_i(y_i, \hat{\boldsymbol{\phi}}_0, \hat{\sigma}_0) = \exp\left\{-\frac{\tau}{2\hat{\sigma}_0^2}(y_i - \mathbf{b}_i\hat{\boldsymbol{\phi}}_0)^2\right\}$ is a weight corresponding to each observation which in the large outlier case tends to zero for outliers to reduce their effect and tends to one for the remaining observations. Using (3.25), empirically we have

$$\mathbf{F}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\hat{\boldsymbol{\beta}}_0) = \frac{1}{N^2\hat{\sigma}_0^4}\sum_{i=1}^{N} w_i^2(y_i, \hat{\boldsymbol{\phi}}_0, \hat{\sigma}_0)(y_i - \mathbf{b}_i\hat{\boldsymbol{\phi}}_0)^2\mathbf{h}_i^{\top}\mathbf{h}_i,$$

$$\mathbf{F}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\hat{\boldsymbol{\beta}}_0) = \frac{1}{N^2\hat{\sigma}_0^4}\sum_{i=1}^{N} w_i^2(y_i, \hat{\boldsymbol{\phi}}_0, \hat{\sigma}_0)(y_i - \mathbf{b}_i\hat{\boldsymbol{\phi}}_0)^2\mathbf{b}_i^{\top}\mathbf{b}_i,$$

$$\mathbf{F}_{\boldsymbol{\theta}\boldsymbol{\phi}}(\hat{\boldsymbol{\beta}}_0) = \frac{1}{N^2\hat{\sigma}_0^4}\sum_{i=1}^{N} w_i^2(y_i, \hat{\boldsymbol{\phi}}_0, \hat{\sigma}_0)(y_i - \mathbf{b}_i\hat{\boldsymbol{\phi}}_0)^2\mathbf{h}_i^{\top}\mathbf{b}_i,$$

$$\mathbf{F}_{\boldsymbol{\phi}\boldsymbol{\theta}}(\hat{\boldsymbol{\beta}}_0) = \frac{1}{N^2\hat{\sigma}_0^4}\sum_{i=1}^{N} w_i^2(y_i, \hat{\boldsymbol{\phi}}_0, \hat{\sigma}_0)(y_i - \mathbf{b}_i\hat{\boldsymbol{\phi}}_0)^2\mathbf{b}_i^{\top}\mathbf{h}_i,$$

Then using (3.28), we have

$$\left[\mathbf{F}^{-1}(\hat{\boldsymbol{\beta}}_0)\right]_{\boldsymbol{\theta\theta}} = \left(\mathbf{F}_{\boldsymbol{\theta\theta}}(\hat{\boldsymbol{\beta}}_0) - \mathbf{F}_{\boldsymbol{\theta\phi}}(\hat{\boldsymbol{\beta}}_0)\mathbf{F}_{\boldsymbol{\phi\phi}}(\hat{\boldsymbol{\beta}}_0)^{-1}\mathbf{F}_{\boldsymbol{\phi\theta}}(\hat{\boldsymbol{\beta}}_0)\right)^{-1}.$$

Finally, substituting the above expressions into the proposed Rao test variant we obtain our robust Rao based subspace detector

$$T_{R_r}(\mathbf{y}) = \frac{\mathbf{u}^\top \left(\mathbf{F}_{\boldsymbol{\theta\theta}}(\hat{\boldsymbol{\beta}}_0) - \mathbf{F}_{\boldsymbol{\theta\phi}}(\hat{\boldsymbol{\beta}}_0)\mathbf{F}_{\boldsymbol{\phi\phi}}(\hat{\boldsymbol{\beta}}_0)^{-1}\mathbf{F}_{\boldsymbol{\phi\theta}}(\hat{\boldsymbol{\beta}}_0)\right)^{-1}\mathbf{u}}{N^2\hat{\sigma}_0^4},$$

which is a function of parameters under $\mathcal{H}_0$. Under the null hypothesis it is easy to show that when $N \to \infty$, $T_{R_r} \to_d \chi^2_p$.

Remark 2: As a special case when $\alpha \to 1$, the proposed robust Rao test reduces to the classical Rao test:

$$T_R(\mathbf{y}) = \frac{\mathbf{y}^\top \mathbf{P}_{\mathbf{B}^\perp}\mathbf{H}\left(\mathbf{H}^\top \mathbf{P}_{\mathbf{B}^\perp}\mathbf{H}\right)^{-1}\mathbf{H}^\top \mathbf{P}_{\mathbf{B}^\perp}\mathbf{y}}{\hat{\sigma}_0^2}. \tag{3.30}$$

3.4.3 Robust Likelihood Ratio Test

The proposed robust LRT can be formed within the GLRT framework. It depends on the $\alpha-$logarithm function of likelihood and has the following structure:

$$T_{G_r}(\mathbf{y}) = \frac{2}{N}\Big[\sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_1}\sum_{i=1}^{N}\log_\alpha\left(f(y_i,\boldsymbol{\omega})\right) - \sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_0}\sum_{i=1}^{N}\log_\alpha\left(f(y_i,\boldsymbol{\omega})\right)\Big] \geq \eta. \tag{3.31}$$

Solving the above optimization problems does not provide a closed form solution as found with the MLE. Each part of (3.31), requires an iterative approach for the solution of the maximization problem. The solution of the first optimization problem requires the score function $\mathbf{s}$, which under the Gaussian nominal model is given by

$$\begin{aligned}\mathbf{s}_{\boldsymbol{\beta}}\left(\mathbf{y}|\mathcal{H}_1\right) &= \nabla_{\boldsymbol{\beta}}\sum_{i=1}^{N}\log_\alpha\left(f(y_i,\boldsymbol{\omega})\right)\\ &= \frac{1}{(2\pi)^{\tau/2}\sigma^{2+\tau}}\sum_{i=1}^{N}w_{i,1}(y_i,\boldsymbol{\omega})(y_i-\mathbf{c}_i\boldsymbol{\beta})\mathbf{c}_i^\top.\end{aligned}$$

Solving $\mathbf{s}_{\boldsymbol{\beta}}(\mathbf{y}|\mathcal{H}_1) = \mathbf{0}$, gives the solution for $\boldsymbol{\beta}$ which is a weighted least square

$$\hat{\boldsymbol{\beta}}_1 = \left(\mathbf{C}^\top \mathbf{W}_1 \mathbf{C}\right)^{-1} \mathbf{C}^\top \mathbf{W}_1 \mathbf{y}, \tag{3.32}$$

where the weight matrix is iteratively computed using

$$\mathbf{W}_1 = \begin{pmatrix} w_{1,1} & 0 & \dots & 0 \\ 0 & w_{2,1} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w_{N,1} \end{pmatrix}, \tag{3.33}$$

and

$$w_{i,1} = \exp\{-\frac{\tau}{2\sigma^2}(y_i - \mathbf{c}_i\boldsymbol{\beta})^2\}, \tag{3.34}$$

is the weight of i^{th} observation under $\mathcal{H}_1$. This weight itself depends on both $\boldsymbol{\beta}$ and σ^2 which are unknown. Thus, in order to estimate the parameters, we need to alternate between (3.32) and (3.34). This estimation procedure depends on the initial point and does not guarantee to converge to the global minimum. To avoid bad local minimum, we can use the existing robust estimators as the initial values. At each step, the estimated values of the previous step will be used until convergence. In the large outlier case, the weights corresponding to outliers tend to zero while for non-outliers they tend to one. When $\alpha \to 1$, $\mathbf{W}_1 \to \mathbf{I}_N$ and the weighted least squares estimator of (3.32) converts to the classical least squares estimator. To estimate σ^2, the score function can be written as

$$\begin{aligned} \mathbf{s}_{\sigma^2}(\mathbf{y}|\mathcal{H}_1) &= \nabla_{\sigma^2} \sum_{i=1}^{N} \log_\alpha\left(f(y_i, \boldsymbol{\omega})\right) \\ &= \frac{1}{2(2\pi)^{\tau/2}\sigma^{4+\tau}} \sum_{i=1}^{N} w_{i,1}(y_i, \boldsymbol{\omega}) \left[(y_i - \mathbf{c}_i\boldsymbol{\beta})^2 - \sigma^2\right]. \end{aligned}$$

Therefore the estimator of variance under $\mathcal{H}_1$ is

$$\hat{\sigma}_1^2 = \frac{\sum_{i=1}^{N} w_{i,1}(y_i - \mathbf{c}_i\boldsymbol{\beta})^2}{\sum_{i=1}^{N} w_{i,1}}. \tag{3.35}$$

To estimate the maximizer of the second term in (3.31) under the null hypothesis ($\mathcal{H}_0$), we use the above estimators where $\boldsymbol{\theta} = \mathbf{0}$ and $\mathbf{C}$ is replaced with $\mathbf{B}$, for example

$\hat{\boldsymbol{\phi}}_0 = \hat{\boldsymbol{\beta}}_1|_{\mathbf{C}=\mathbf{B}}$. Using these estimated values in (3.31), we have

$$\begin{aligned} T_{G_r}(\mathbf{y}) &= \frac{2}{N}\Big[\sum_{i=1}^{N}\log_\alpha\left(f(y_i,\hat{\boldsymbol{\omega}}_1)\right) - \sum_{i=1}^{N}\log_\alpha\left(f(y_i,\hat{\boldsymbol{\omega}}_0)\right)\Big] \\ &= \frac{2}{N}\Big[\sum_{i=1}^{N}\frac{w_{i,1}}{\tau(2\pi\hat{\sigma}_1^2)^{\tau/2}} - \sum_{i=1}^{N}\frac{w_{i,0}}{\tau(2\pi\hat{\sigma}_0^2)^{\tau/2}}\Big] \geq \eta. \end{aligned} \tag{3.36}$$

As a special case when $\alpha \to 1$ the MSD can be derived using (3.36). Now we derive the asymptotic distribution of the test under the null hypothesis. Even though, it is well known that classical forms of GLRT, Rao and Wald tests are asymptotically equivalent, this statement is not true for their robust forms derived here. In contrast to the robust Rao and Wald tests that are asymptotically chi-square distributed under the null hypothesis, the proposed robust LRT has a weighted chi-square distribution which makes it different.

Proposition 2: Define $\boldsymbol{\omega}^* = \arg\max_{\boldsymbol{\omega}} \mathbf{E}\left[\log_\alpha\left\{f(y,\boldsymbol{\omega})\right\}\right]$ and $\mathbf{Q} = \mathbf{E}\left[\nabla_{\boldsymbol{\omega}} g_N \nabla_{\boldsymbol{\omega}} g_N^\top\right]$ where $g_N(\boldsymbol{\omega}) = \sum_{i=1}^{N}\log_\alpha\left\{f(y_i,\boldsymbol{\omega})\right\}$. Assume asymptotically $\hat{\boldsymbol{\omega}} \to_p \boldsymbol{\omega}^*$ as $N \to \infty$, then under $\mathcal{H}_0$ and certain regularity conditions, asymptotically $NT_{G_r}(\mathbf{y}) \simeq N\hat{\boldsymbol{\theta}}_1^\top(\mathbf{M}_{\theta\theta} - \mathbf{M}_{\theta\lambda}\mathbf{M}_{\lambda\lambda}^{-1}\mathbf{M}_{\lambda\theta})\hat{\boldsymbol{\theta}}_1$ and is distributed as $\sum_{i=1}^{p}\gamma_i N_i^2$, where $\mathbf{M} = -\mathbf{E}\left[\frac{\partial^2 g_N(\boldsymbol{\omega})}{\partial\boldsymbol{\omega}\partial\boldsymbol{\omega}^\top}\right]$ is a semi positive definite matrix, $N1,\cdots,N_p$ are independent standard normal variables and finally γ_i are the p positive eigenvalues of the matrix $\mathbf{Q}[\mathbf{M}^{-1} - (\mathbf{M}^*)^+]$, where

$$(\mathbf{M}^*)^+ = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{\lambda\lambda}^{-1} \end{bmatrix}.$$

Proof: See Appendix B.

Using the asymptotic distribution of the proposed test under $\mathcal{H}_0$, the threshold value based on the desired false alarm rate can be computed in the receiver side. A summary of proposed detector's expressions together with classical counterparts are presented in table 3.1.

3.5 Robustness Properties

The Robustness of an estimator or test can be analyzed in terms of local and global robustness. Hampel [114] introduced the IF as an approach to investigate the asymptotic

TABLE 3.1: Summarizing table of classical and robust tests for the model (3.1).

	Wald Test	Rao Test	GLRT
Classical	$T_W(\mathbf{y}) = \dfrac{\hat{\boldsymbol{\theta}}_{ML}^{\top}\left(\mathbf{H}^{\top}\mathbf{P}_{\mathbf{B}^{\perp}}\mathbf{H}\right)\hat{\boldsymbol{\theta}}_{ML}}{\hat{\sigma}_1^2}$	$T_R(\mathbf{y}) = \dfrac{\mathbf{y}^{\top}\mathbf{P}_{\mathbf{B}^{\perp}}\mathbf{H}\left(\mathbf{H}^{\top}\mathbf{P}_{\mathbf{B}^{\perp}}\mathbf{H}\right)^{-1}\mathbf{H}^{\top}\mathbf{P}_{\mathbf{B}^{\perp}}\mathbf{y}}{\hat{\sigma}_0^2}$	$T_G(\mathbf{y}) = \dfrac{\mathbf{y}^{\top}\mathbf{P}_{\mathbf{B}^{\perp}}\mathbf{P}_{\mathbf{G}}\mathbf{P}_{\mathbf{B}^{\perp}}\mathbf{y}}{\mathbf{y}^{\top}\mathbf{P}_{\mathbf{B}^{\perp}}\mathbf{P}_{\mathbf{G}^{\perp}}\mathbf{P}_{\mathbf{B}^{\perp}}\mathbf{y}}$
Robust	$T_{W_r}(\mathbf{y}) = N\hat{\boldsymbol{\theta}}^{\top}\mathbf{R}^{-1}\hat{\boldsymbol{\theta}}$	$T_{R_r}(\mathbf{y}) = \dfrac{\mathbf{u}^{\top}\left(\mathbf{F}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\hat{\boldsymbol{\beta}}_0) - \mathbf{F}_{\boldsymbol{\theta}\boldsymbol{\phi}}(\hat{\boldsymbol{\beta}}_0)\mathbf{F}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\hat{\boldsymbol{\beta}}_0)^{-1}\mathbf{F}_{\boldsymbol{\phi}\boldsymbol{\theta}}(\hat{\boldsymbol{\beta}}_0)\right)^{-1}\mathbf{u}}{N^2\hat{\sigma}_0^4}$	$T_{G_r}(\mathbf{y}) = \dfrac{2}{N}\left[\sum_{i=1}^{N}\dfrac{w_{i,1}}{\tau(2\pi\hat{\sigma}_1^2)^{\tau/2}} - \sum_{i=1}^{N}\dfrac{w_{i,0}}{\tau(2\pi\hat{\sigma}_0^2)^{\tau/2}}\right]$

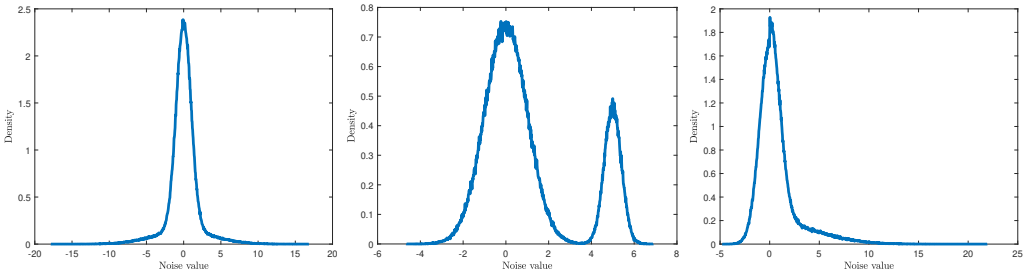

FIGURE 3.2: Sample contaminated Gaussian densities that are used in the simulation section with 20% contamination. Heavy tail noise density (left), point mass noise density (middle) and asymmetric noise density (right).

sensitivity of an estimator to small departures from its nominal model and it was shown that the IF is directly related to the bias due to this departure. To measure the global robustness, the breakdown point (BP) has been used in the literature to approximate the maximum fraction of outliers that a robust estimator can handle.

3.5.0.1 The Influence Function

The IF of an estimator or a test represents the sensitivity of the estimator or the test to a slight departure from its nominal distribution. In the literature, it is considered as a measure of local robustness because it is based on analyzing small deviations from the nominal model. Based on the results in [42], using the contamination model $(1-\epsilon)\mathcal{N}(0,\sigma^2)+\epsilon\delta_{z_0}$, $0\leq\epsilon\leq 1/2$ where δ_{z_0} is the Dirac's Delta distribution function at the point z_0, the IF of the proposed estimator $\hat{\boldsymbol{\beta}}$ with known σ has the form

$$IF(z_0;\boldsymbol{\beta}) = b^{-1}z_0\exp\{-\tau\frac{z_0^2}{2\sigma^2}\}\mathbf{U}_c^{-1}\mathbf{c}_0^\top,$$

where $\mathbf{U}_c=\mathbf{C}^\top\mathbf{C}$ and b is a constant term that does not depend on z_0. When $\tau\to 0$, the IF is monotone increasing meaning that large outliers have a large impact on the performance of the estimators which is equivalent to MLE. For any $\tau>0$, the shape of $IF(z_0;\boldsymbol{\beta})$ is redescending and for $z_0\to\infty$, $IF(z_0;\boldsymbol{\beta})\to 0$; thus the estimator is not sensitive to the large outliers. In this case, $IF(z_0;\boldsymbol{\beta})$ is bounded and continuous which means that the bias does not diverge to infinity for severe contamination and small outliers have a small effect on the estimator.

Robustness of a test has two characteristics. The first one is stability in level (P_{FA}) where the instability in this case can be created by deviations in distribution of the null hypothesis. The second is instability in power that appears when there is deviation

of the noise model in the alternative hypothesis. Based on the results in [115], both level and power of the proposed tests depend on the influence function of the functional defining the test (for example $\mathbf{U}_{LRT} = (\mathbf{M}_{\theta\theta} - \mathbf{M}_{\theta\lambda}\mathbf{M}_{\lambda\lambda}^{-1}\mathbf{M}_{\lambda\theta})^{\frac{1}{2}}\hat{\boldsymbol{\theta}}_1$). For example for the level of contaminated model (P_{FA_ϵ}) in robust likelihood ratio test, we have

$$P_{FA_\epsilon} \simeq P_{FA_0} + \epsilon^2\mu\| \int IF(x;\mathbf{U}_{LRT})dH(x)\|^2, \tag{3.37}$$

where μ is a scale factor and P_{FA_0} is the nominal level. Therefore, bounded IF of the estimators guarantees the robustness of the tests.

3.5.0.2 The Breakdown Point

The proposed estimator belongs to the redescending M-estimators family and its finite sample BP for our defined univariate scenario is given by $\frac{0.5(N-p-t-1)}{N}$ [42]. When $N \to \infty$ this value tends to 0.5 which means the estimator can handle up to 50% outliers in the data with selecting an appropriate initial point.

3.6 Simulation Results

In this section, the proposed robust tests (3.20), (3.27) and (3.36) are compared with the MSD [49], Wald test and robust version of Rao test using Huber estimation approach [116] with 95% efficiency in three different contaminated Gaussian densities: heavy tail, point mass and asymmetric which are illustrated in figure 3.2. For all simulations $N = 200$ and $p = t = 2$. For both $\mathbf{H}$ and $\mathbf{B}$, we use cosine function as $h_{i,j} = \cos(\frac{2\pi ij}{N})$ and $b_{i,j} = \cos(\frac{14\pi i(j-1)}{N})$. The signal to noise ratio (SNR) which shows the average power of signal to the power of the noise in decibel is defined by

$$\text{SNR} = 10\log_{10}\frac{\|\mathbf{H}\boldsymbol{\theta}\|_2^2/N}{\sigma^2_{nominal}},$$

and is set to -10 dB.

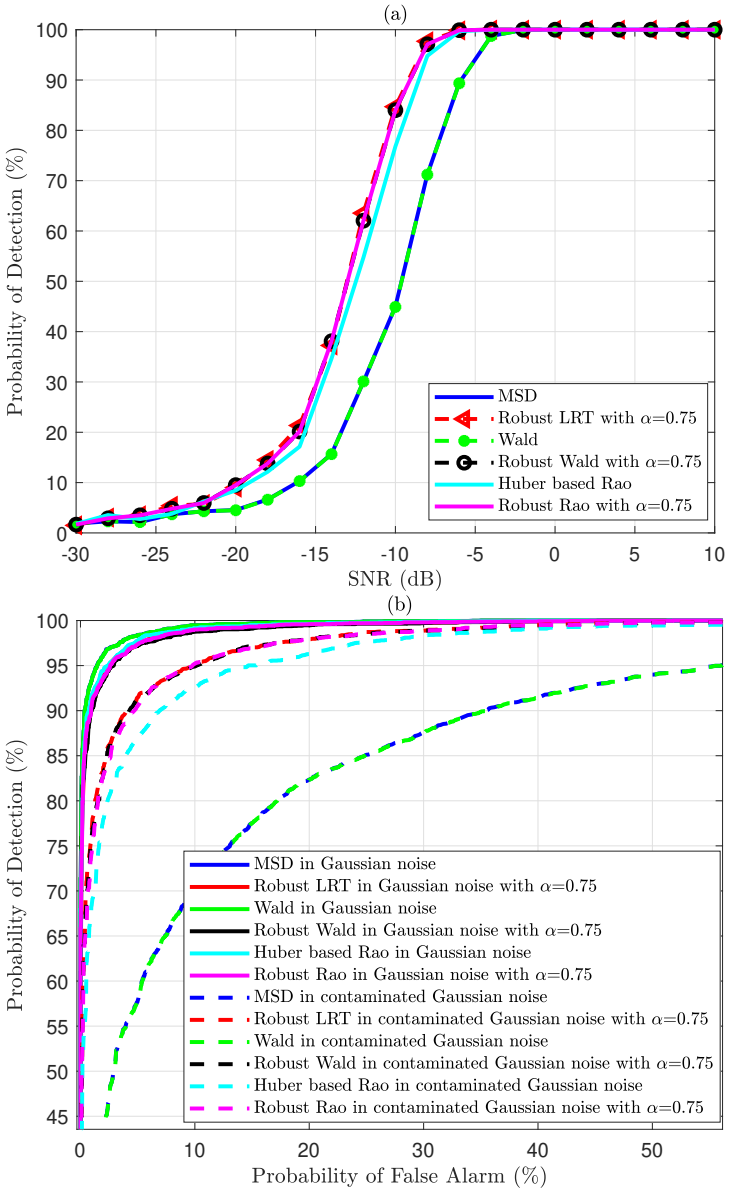

FIGURE 3.3: Performance of proposed tests compared to the MSD, Wald, Huber based Rao in heavy tail noise scenario, (a)$\epsilon = 15\%$, $P_{Fa} = 2\%$ and $\alpha = 0.75$, (b) $\epsilon = 15\%$, SNR $= -10dB$ and $\alpha = 0.75$.

3.6.1 Heavy Tail Noise

To generate heavy tail noise density g, we use the following contamination model:

$$g(n) : (1-\epsilon)\mathcal{N}(0,\sigma_1^2) + \epsilon\mathcal{N}(0,\sigma_2^2), \tag{3.38}$$

where usually $\sigma_2^2 \gg \sigma_1^2$ and the second term makes the tail of density. In this density, ϵ determines the level of contamination. 8000 noise samples are generated by this density using Monte Carlo when $\epsilon = 15\%$, $\sigma_1^2 = 1$ and $\sigma_2^2 = 16$. Target signal is embedded in

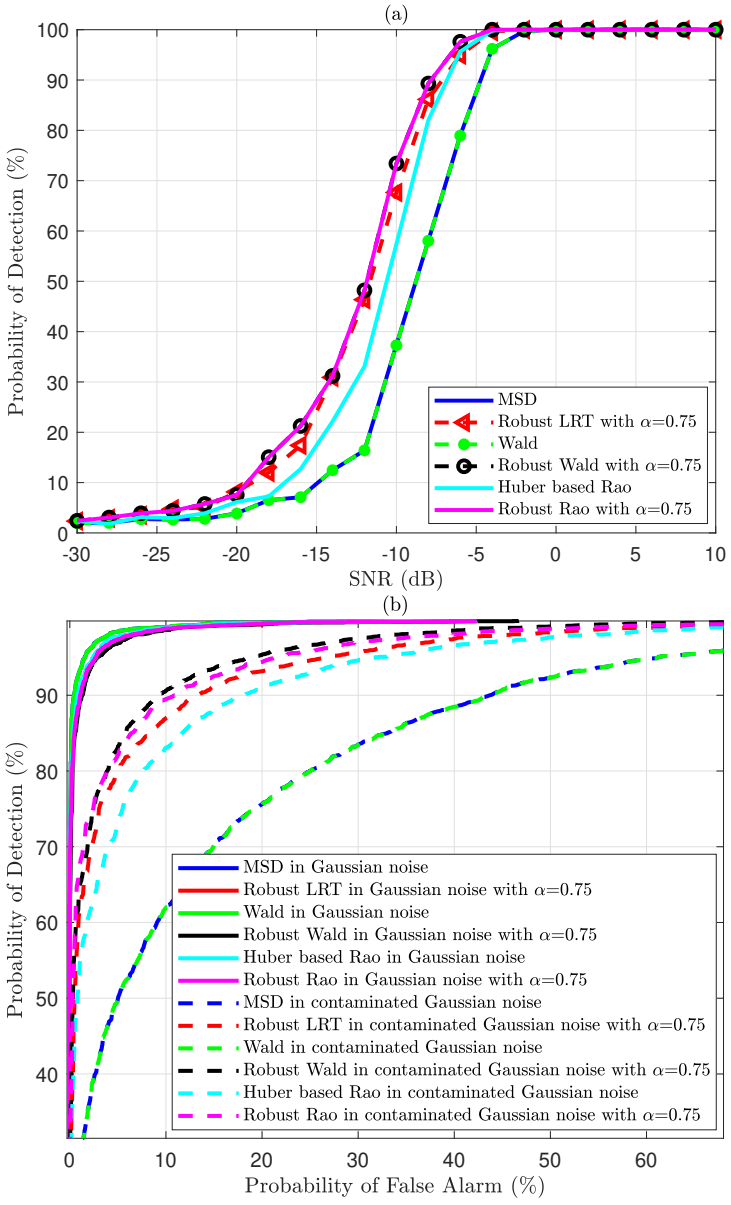

FIGURE 3.4: Performance of proposed tests compared to the MSD, Wald, Huber based Rao in point mass noise scenario, (a)$\epsilon = 15\%$, $P_{Fa} = 2\%$ and $\alpha = 0.75$, (b) $\epsilon = 15\%$, SNR $= -10dB$ and $\alpha = 0.75$.

the noise + interference with probability 1/2 to have a balanced detection problem. The density of this noise is illustrated in figure 3.2. Figure 3.3 shows the receiver operating characteristic (ROC) curve of the proposed tests in comparison to the classical tests in both Gaussian and heavy tail densities when $\alpha = 0.75$. It is shown that with increasing SNR, all tests give more power, but the proposed tests uniformly have a better detection rate. In general, the classical tests have better power in Gaussian case due to using efficient estimates of the parameters but significantly lose their power when adding contamination. On the other hand, the proposed tests instead of having a better power

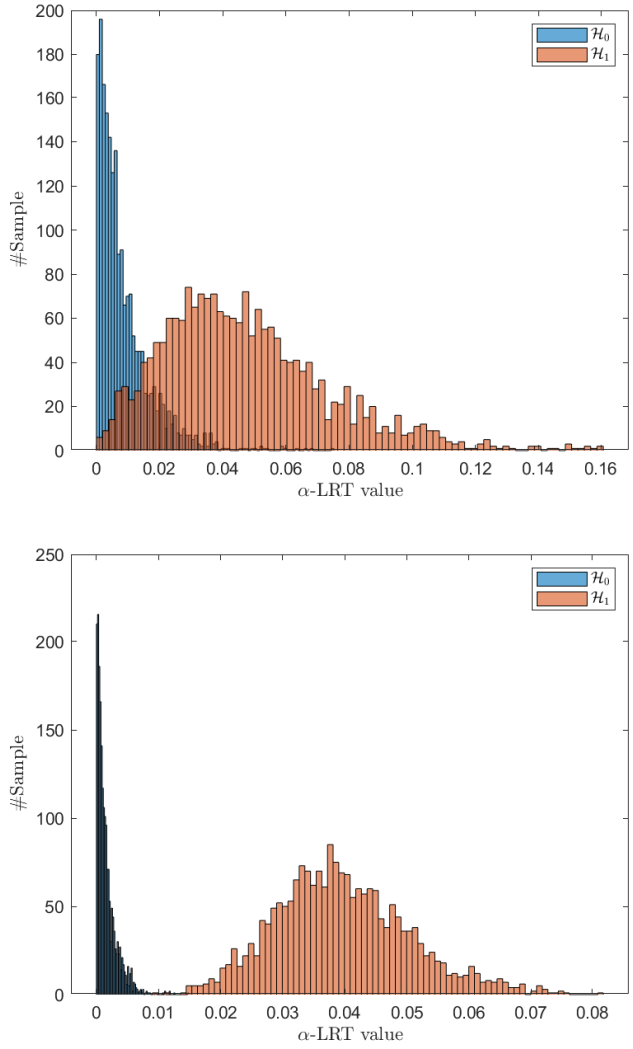

FIGURE 3.5: Empirical distribution of proposed robust LRT when $N = 200$ (top) and $N = 1000$ (bottom). Blue samples show the distribution of the test under $\mathcal{H}_0$ and the red part shows the distribution of the test under $\mathcal{H}_1$.

in the Gaussian case, keep the power unchanged when adding contamination by assigning smaller weights to the outliers when estimating the parameter. Compared to the Huber estimator, the proposed estimation approach assigns an exponentially decreasing weights to the high residuals which means it can effectively reduce the contribution of severe outliers in the estimation step while Huber estimator still can be affected by those outliers. The trade-off between the power in Gaussian case and robustness to outliers can be tuned by the value of α. $\alpha \to 1$ gives exactly the same performance as the classical methods, and with decreasing α but close to 1, the power in the nominal case is reduced to increase robustness. More remarks on tuning α are provided in the next subsections.

3.6.2 Point Mass Noise

In point mass noise scenario, the outliers come from another Gaussian density which is centered at μ_2 with significantly smaller variance. To generate point mass noise density g, we use the following contamination model

$$g(n) : (1-\epsilon)\mathcal{N}(0,\sigma_1^2) + \epsilon\mathcal{N}(\mu_2,\sigma_2^2). \tag{3.39}$$

Similar to the previous scenario, 8000 noise samples are generated using this introduced density with $\epsilon = 15\%$, $\sigma_1^2 = 1$, $\mu_2 = 5$ and $\sigma_2^2 = 0.16$ and have been added to the interference. The signal of interest is embedded in half of them and then the result is considered as the input of the tests. Figure 3.4 shows the ROC of the tests. Results show that the proposed approaches can significantly account for outliers in the observations and have more accurate estimation compared to ML based methods. It is the consequence of relaxing some power in Gaussian noise. For example in $P_{Fa} = 10\%$, with adding contamination, MSD loses approximately 40% of its power while proposed tests just approximately lose 7% of its power while relaxing only 3% of its power in Gaussian noise. This robustness can even be increased by reducing α. Figure 3.5 shows the histogram of the proposed robust LRT under both $\mathcal{H}_0$ and $\mathcal{H}_1$ as a function of N. It's shown that when increasing N, the separability of two hypotheses will increase and the distribution of the test under $\mathcal{H}_0$ tends to the asymptotic distribution which was discussed above.

3.6.3 Asymmetric Noise

To generate asymmetric noise density g, we use the following contamination model:

$$g(n) : (1-\epsilon)\mathcal{N}(0,\sigma_1^2) + \epsilon|\mathcal{N}(\mu_2,\sigma_2^2)|, \tag{3.40}$$

where the second term is a density which always generates a positive random variable. Similar to previous scenarios, 8000 observations are generated which half of them include the signal of interest. parameters of density are set to $\epsilon = 15\%$, $\sigma_1^2 = 1$, $\mu_2 = 2$ and $\sigma_2^2 = 16$. ROC of the detectors are illustrated in figure 3.6. Again it is shown that the proposed detectors give us better power in different SNR values and for different false

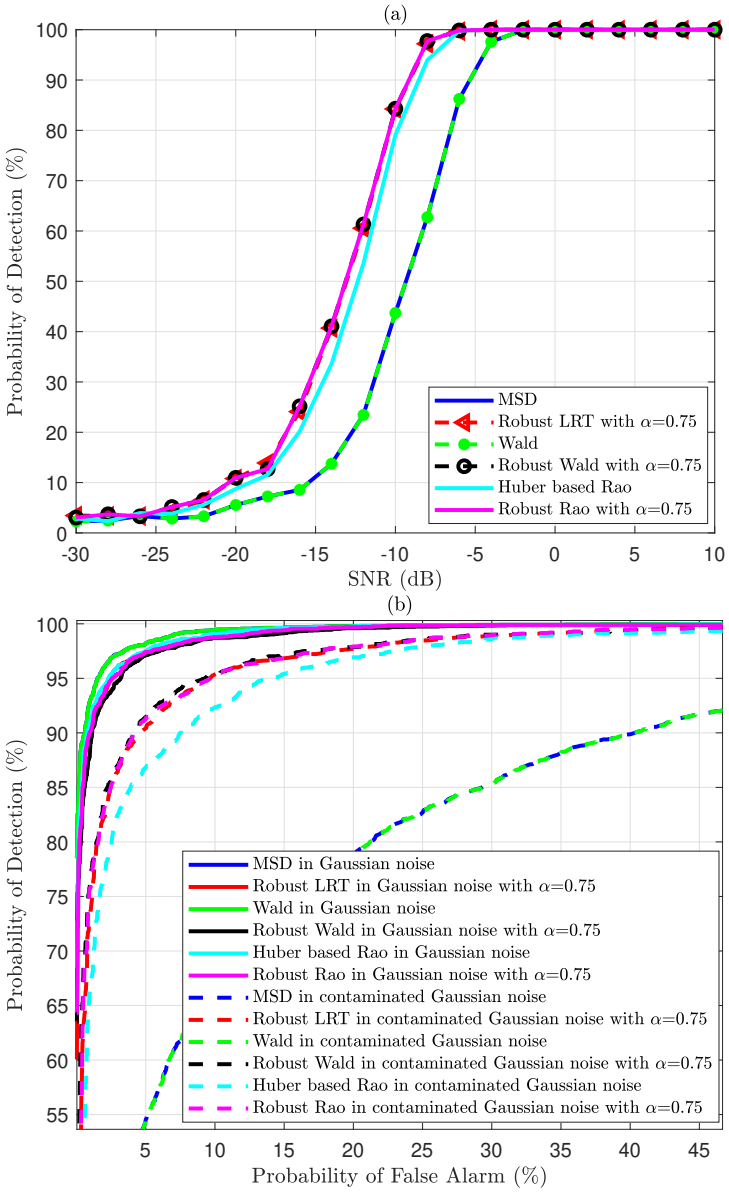

FIGURE 3.6: Performance of proposed tests compared to the MSD, Wald, Huber based Rao in asymmetric noise scenario, (a)$\epsilon = 15\%$, $P_{Fa} = 2\%$ and $\alpha = 0.75$, (b) $\epsilon = 15\%$, SNR $= -10dB$ and $\alpha = 0.75$.

alarm rates. This verifies that the proposed detectors are robust against outliers. As the specific case where we set $\alpha \to 1$, the result is exactly the same as the classical detectors.

3.6.4 Tuning Parameter α

In general, the values of α in robust estimation should be taken such that $\alpha < 1$, otherwise the methods will be more sensitive to outliers due to assigning high weights to high residuals. On the other hand, with the choice of $\alpha \to -\infty$ detectors get more strict and reject all of the observations which means no data is left to be used in estimation.

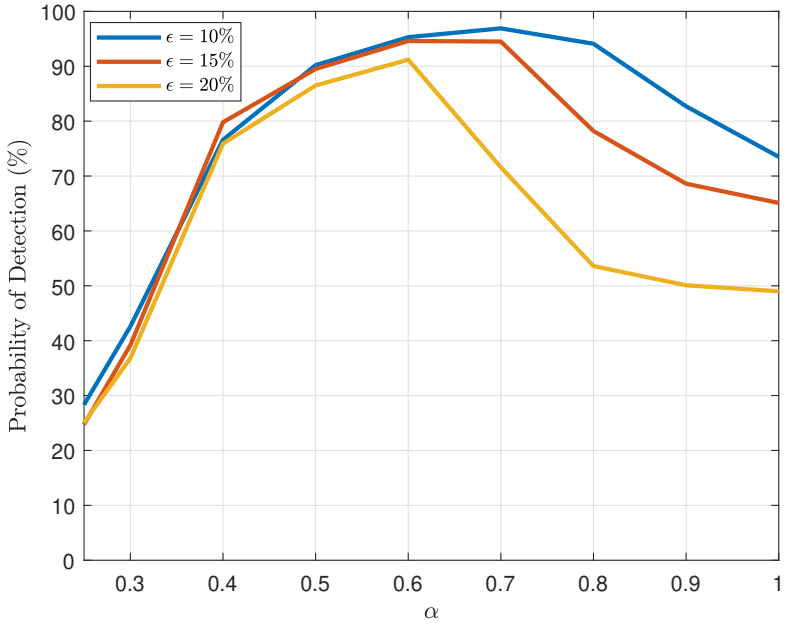

FIGURE 3.7: Performance of Proposed robust LRT in different contamination levels of point mass noise scenario as a function of α, when the probability of false alarm is fixed to 10% and SNR $= -10dB$. It is shown that there is an optimal α where after or before that point the power will be decreased due to outliers and limited samples, respectively.

Therefore, there is a proper interval that is upper bounded by 1 and the lower bound depends on the amount and the type of outliers. We can choose the α within this interval based on the amount of robustness and power that we need. Figure 3.7 shows the power of robust LRT in simulated point mass noise with different levels of contamination. When gradually the value of α is decreased, the power in all levels will increase. There is an optimal point for each level of contamination. Passing that point, the detector starts to reject more than enough observations and as a consequence, the power will be decreased. Therefore, tuning this parameter has a major impact on the result. This value can be determined by methods such as cross-validation or an outliers detection statistic [42]. It can also be determined by the user, based on the level of outliers in a specific application using knowledge about the underlying process. For example, in functional magnetic resonance imaging (fMRI) data analysis it is accepted that in low SNR values, noise is almost Rician distributed [35].

Considering $\epsilon_m\%$ as the maximum fraction of outliers, we need an estimator to at most, reduce the effect of $\epsilon_m\%$ of data. Therefore we can start from $\alpha \simeq 1$, compute the absolute value of the residuals and sort them to find the minimum value among the biggest $\epsilon_m\%$ of residuals and name it n_l. Then solving the equation $w = 1/2 = \exp\{-\frac{1-\alpha_l}{\sigma^2} n_l^2\}$

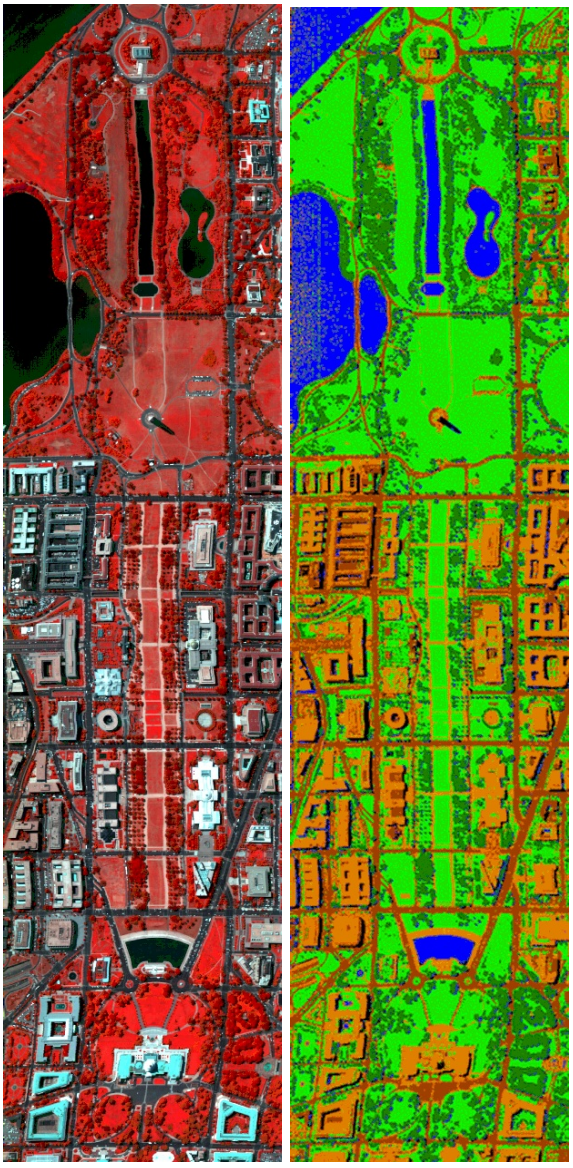

FIGURE 3.8: HYDICE data (left) and the ground truth (right).

we can find the corresponding α_l and assigning the weight of for example $1/2$ (user can change it) to that residual. We can repeat these steps until converge. Now α can be picked within $[\alpha_l, 1)$. This approach is a relaxed form of minimax concept that tries to be robust even for the worst-case scenario and selecting α_l, as the consequence reduces the power in the Gaussian noise. On the other hand, a similar technique of other robust estimators like Huber's estimator can be used to tune α. In the second approach, the goal is to achieve at least $a\%$ (for example $a = 90$) of efficiency of the optimal estimator in Gaussian scenario and hopefully that $(100 - a)\%$ loss, can provide enough robustness

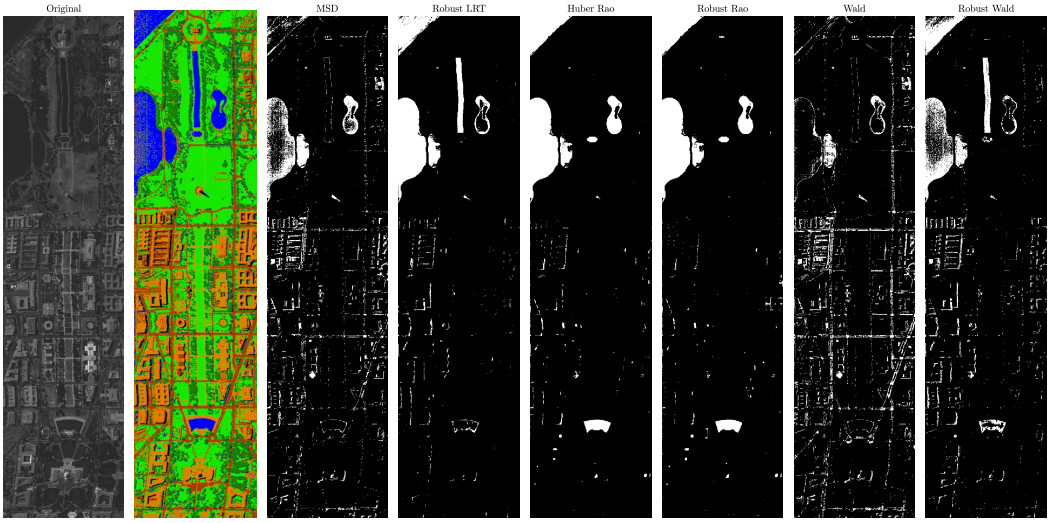

FIGURE 3.9: The results of proposed and conventional detectors on HYDICE data. From left the images show the Original data in 50^{th} band, ground truth, matched subspace detector, proposed robust likelihood ratio test, robust Huber based Rao test, proposed robust Rao test, Wald test and proposed robust Wald test.

in neighborhood distributions. The efficiency ratio can be defined as

$$\frac{\text{tr}(\mathbf{I}^{-1})}{\text{tr}(\mathbf{M}^{-1}\mathbf{Q}\mathbf{M}^{-1})}, \tag{3.41}$$

where $\mathbf{I}$ is the Fisher information matrix and the denominator is the asymptotic variance of the proposed estimator [113].

3.7 Experimental Results

In this section, the proposed tests are evaluated using three experimental datasets. In the first experiment, the proposed methods are applied to hyperspectral data and in the second experiment they are applied to the real fMRI data. At the end of this section, we show the power of the proposed robust estimator for the case of non-Gaussian nominal noise (Gamma distribution) which is useful for synthetic-aperture radar (SAR) data analysis.

3.7.1 Washington DC Mall Data

This data is consisted of 191 channels (bands) with 1280×307 pixels and was collected by the Hyperspectral digital imagery collection experiment (HYDICE) sensor over the

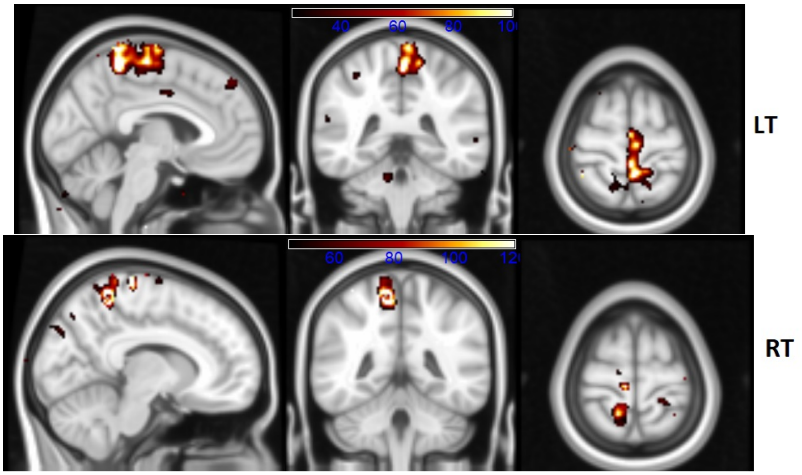

FIGURE 3.10: Activation map of the brain using Proposed robust Wald test for two different tasks of moving left toe (LT) and right toe (RT).

Washington DC MALL[1] in 1995. Figure 3.8 shows the collected data as an image and the labeled pixels as the ground truth where the labels include "grass", "water", "road" and etc. To be able to apply the detectors on the data, The "water" is chosen as the target signal and some sample pixels are picked as the training data to represent the target signal subspace. Using these training pixels we can approximate the matrix $\mathbf{H}$. To approximate the noise density, we use a normal Gaussian noise nominal model and assume that the noise variance is the same over the different bands. "Water" labeled pixels are shown in blue in the figure 3.9. We estimated the sample covariance of 1500 sample pixels and used the first 4 eigenvectors of the sample covariance as the target subspace. The value 4 is selected experimentally to have approximately 80% of the energy of the singular values and meanwhile reduce the impact of interferences (in particular, other classes). Also a rank 1 dc value vector is set as the interference to remove the dc values in observations. Figure 3.9 shows the output of detectors on the data when $\alpha = 0.8$. To represent the results, 25000 of the most possible pixels as "water" are shown. All of proposed detectors have better detection rate compared to their classical form. Most of the water pixels are detected correctly using our proposed methods while MSD and Wald tests could not detect some parts of water pixels.

[1]https://engineering.purdue.edu/ biehl/MultiSpec/hyperspectral.html

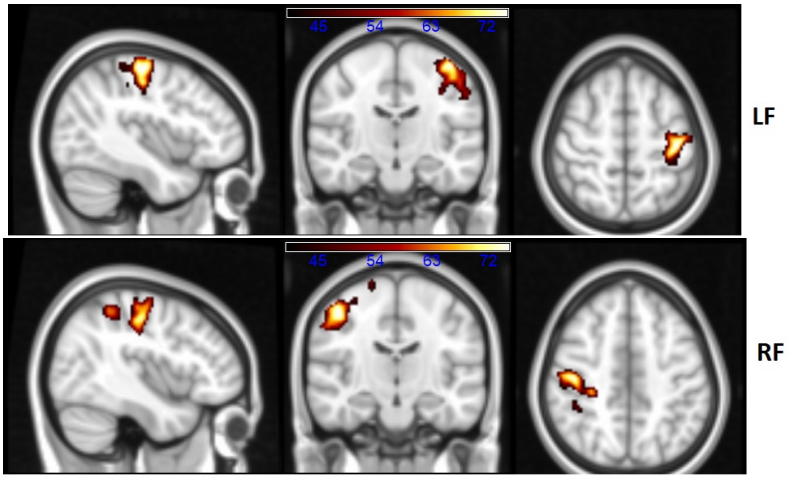

FIGURE 3.11: Activation map of the brain using Proposed Rao test for two different tasks of moving left finger (LF) and right finger (RF).

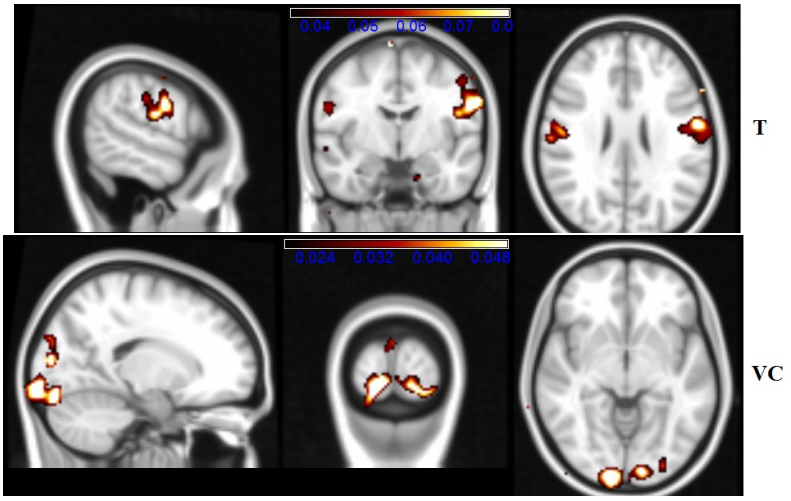

FIGURE 3.12: Activation map of the brain using Proposed robust LRT for two different tasks of moving tongue (T) and doing visual cue (VC).

3.7.2 Real fMRI Data

One important application of the signal detection approaches is in fMRI data analysis where we try to find the active areas in the brain for a specific task or in the resting state situation. In [117], MSD is used as a method for active area detection in the brain. Unlike communication problems where the receiver side usually has some accurate information about parameters, in this application, we usually are facing many of unknown parameters. Most of the time, for simplicity, researchers assume that the brain voxels are linear and time-invariant (LTI) systems. With convolving the task stimulus with hemodynamic response function (HRF) as the impulse response, the subspace of the signal can be approximated [8]. Sometimes, researchers use data-driven methods and

use assumptions like sparsity [29] [28] and independent components [118] to increase the accuracy in fMRI data analysis.

For evaluating the performance of our proposed detectors on the real fMRI data, we assume that the noise distribution is not purely Gaussian. To evaluate the proposed method, we used a single subject (ID: 100307) motor-task fMRI data set. This dataset was acquired from the Q1 release of the Human Connector Project [119]. The acquisition parameters were: 90×104 matrix, 220mm FOV, 72 slices, TR=0.72s, TE=33.1ms, flip angel=52°, BW=2290 Hz/Px, in-plane FOV= 208×180mm with 2mm isotropic voxels. This data was already preprocessed using the following steps: motion correction, temporal pre-whitening, slice time correction, global drift removal. Some other spatial masking and smoothing are done similarly to [32] and finally results are spatially normalized to the standard MNI152 template to be presented. In this experiment, subject was asked to squeeze left or right toes, tap left or right fingers, move tongue or do a visual task. At each voxel, 284 scans was done. 5 low frequency cosine functions are used as the interference subspace to remove the drift terms. In figures 3.10, 3.11 and 3.12 we see the output of the proposed detectors and the area of the brain that is detected as active during the experiment when α is 0.8. In the simulations, we used the canonical HRF as an impulse response of the brain and estimated the rank-1 signal subspace $\mathbf{H}$ by convolving the HRF with stimulus function of the data. The activation maps are completely aligned with the activation maps that the classical approaches offer. The results are also consistent with the knowledge that we have from the functionality of the brain. In these figures, the colormaps show the output of test statistics. Higher values show the higher chance of neural activation. The main advantage of the proposed methods in fMRI data analysis is their robustness against possible outliers.

3.7.3 Ship Detection using SAR Images

In this section, we briefly demonstrate the strength of $\alpha-$divergence based robust estimators for non-Gaussian nominal models. Ships or generally man-made targets detection in high resolution SAR images is one of the applications of GLRT where the noise distribution is not Gaussian. For example, in [120], the GLRT has been used to detect ships in sea clutters. Based on literature [121, 122], in this chapter we consider the case that the radar cross section (RCS) under both $\mathcal{H}_0$ and $\mathcal{H}_1$ follows two different Gamma

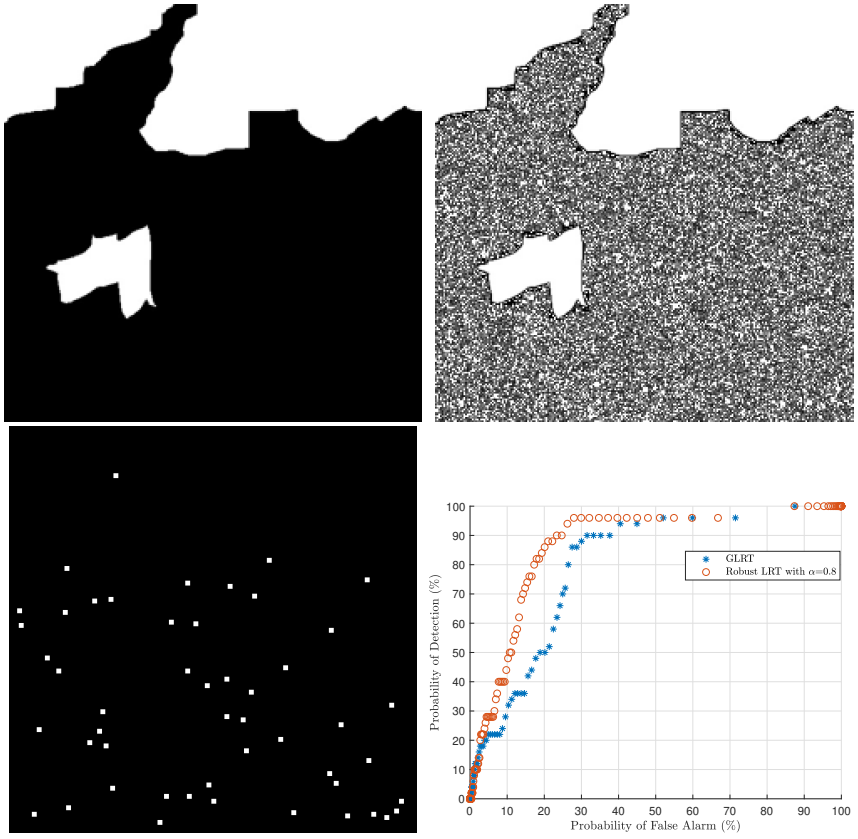

FIGURE 3.13: The results of proposed robust LRT on the simulated SAR image. The land mask (top left), the input SAR image (top right), true location of targets (bottom left), ROC (bottom right).

distributions which are formulated as

$$f(y) = \frac{1}{\Gamma(k)q^k} y^{k-1} \exp\left(-\frac{y}{q}\right), \; y \geq 0, \tag{3.42}$$

where k and q are the shape and scale parameters. Even though there is a high confidence for validity of above distributions in GLRT, but still small deviations can add significant bias to the results. To guarantee the reliability of the results, we put the ML based estimators aside and derive the $\alpha-$divergence based estimators to put in GLRT framework.

Starting form (3.14), for the scale we have

$$\hat{q} \;\; = \;\; \frac{\sum_{i=1}^{N} w_i y_i}{k \sum_{i=1}^{N} w_i}, \tag{3.43}$$

and for the shape parameter it can be estimated by solving following equation

$$\psi(k) \;\; = \;\; -\log q + \frac{\sum_{i=1}^{N} w_i \log y_i}{\sum_{i=1}^{N} w_i}, \tag{3.44}$$

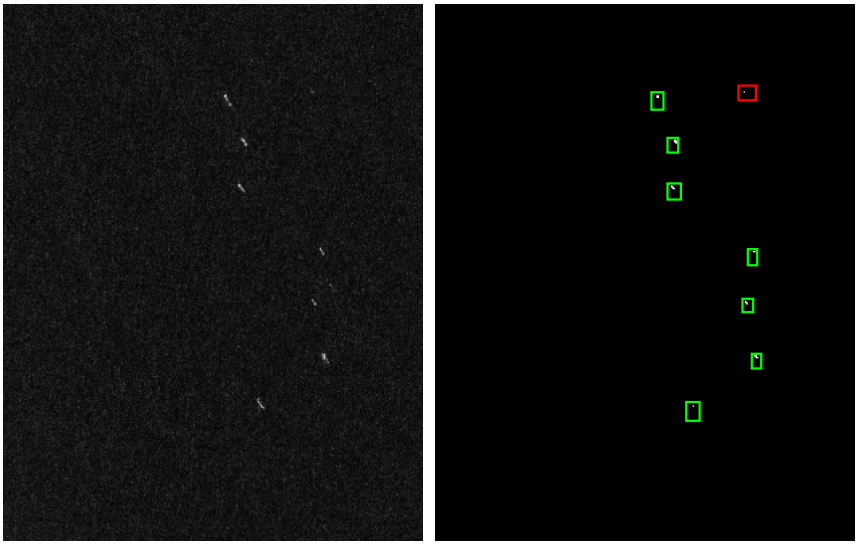

FIGURE 3.14: Input SAR image (left) and detection results (right).

where $w_i = y_i^{(1-\alpha)(k-1)} \exp\{-\frac{(1-\alpha)y_i}{q}\}$. These estimators, as we could guess based on previous sections, are recursive weighted estimators which can converge after a few steps. When $\alpha \to 1$ the ML based estimators can be recovered and for $\alpha < 1$ they offer robustness property. In order to show the superiority, a single 250×250 simulated SAR image with 15% contamination (point mass) using two nominal Gamma distributions is generated (figure 3.13), one with $(k = 4, q = 0.1)$ under $\mathcal{H}_0$ and one with $(k = 4, q = 0.3)$ under $\mathcal{H}_1$. 50 distributed targets with the size of 3×3 are randomly located in the image. The area which is uniformly white shows the land area near the sea and usually can be segmented by land masking block using other types of modalities. Setting $\alpha = 0.8$ and using a windowing approach described in [123] with the size of 9×9, the ROC curve can be extracted using 9 centered pixels as the test and outer 32 pixels as the training samples of clutter model. It can be observed that using proposed method the detection rate increases at fixed false alarm rates.

Next, we apply our proposed robust LRT to real SAR data. Figure 3.14 shows acquired image by RADARSAT-2 over the port of Tianjin (China) and the output of detector. The format of the original data is single-look complex and spatial multilooking is performed by combining 2×2 pixels in the intensity domain. The proposed method detects all targets correctly (green boxes) and the red box shows an incorrect target (false alarm).

3.8 Conclusions

In this chapter, robust Wald, Rao and likelihood ratio tests that enable detection of a subspace signal in contaminated Gaussian noise and interference are proposed. The tests or detectors are derived using the $\alpha-$divergence and are adjustable with a single parameter α. When $\alpha \to 1$, these tests are equivalent to their well known classical counterparts. This confirms their validity as the $\alpha-$divergence reduces to the KL divergence in this particular case. For other choices of $\alpha < 1$, the proposed tests achieve the robustness property by assigning small weights to the outliers. Approaches on how to adjust the tuning parameter are also provided. Asymptotic analysis of the proposed detectors is provided and their performance is illustrated using simulated, fMRI, hyperspectral and SAR datasets. As the future directions, we are considering to extend the proposed approaches to derive the subspace detectors in the case of auto-correlated noise. It would be also interesting to investigate the case where the nominal density is a Gaussian mixture model (GMM).

Chapter 4

Learning Robust and Sparse Principal Components with the $\alpha-$Divergence

In this chapter, novel robust principal component analysis (RPCA) methods are proposed to exploit the local structure of datasets. The proposed methods are derived by minimizing the α-divergence between the sample distribution and the Gaussian density model. The $\alpha-$divergence is used in different frameworks to represent variants of PCA approaches including orthogonal, non-orthogonal and sparse methods. We show that the classical PCA is a special case of our proposed methods where the $\alpha-$divergence is reduced to the Kullback-Leibler (KL) divergence. It is shown in simulations that the proposed approaches recover the underlying principal components (PCs) by down-weighting the importance of structured and unstructured outliers. Furthermore, using simulated data, it is shown that the proposed methods can be applied to fMRI signal recovery and foreground-background (FB) separation in video analysis. Results on real world problems of FB separation as well as saliency map identification are also provided.

4.1 Introduction

Principal component analysis (PCA) is one of the most popular techniques for visualizing data[124], data analysis [125] and dimension reduction[6] which makes it applicable

for a wide variety of image and video processing applications like video surveillance, face recognition, image restoration, hyperspectral image denoising, etc.[126]. Classical PCA provides orthogonal basis vectors, such that projecting data on the subspace spanned by these vectors captures the maximum possible variability of the data (keeps the maximum possible information which is defined by the variance) in the reduced dimension space.

Despite PCA's popularity, it has three main drawbacks which can reduce its utility in real world applications. First, PCA does not offer any method to determine the number of needed principal components (PCs) (i.e., the dimension of projected data). This issue has been addressed by using model order selection criteria [127, 128]. Second, it is unable to determine the contribution of each variable on the PCs due to the dense structure of loading vectors. This is a critical issue in areas such as explainable artificial intelligence (AI) and feature selection. This has been addressed by forcing the contribution of some variables to be zero and allowing a few nonzero entries exist in the loading vectors [129] or by adding sparsity constraint and relaxing orthogonality [130]. Finally, it is sensitive to outliers. The classical PCA performs well when the data follows a unimodal distribution. In the case of multi-modal distributions or the presence of outliers which are the concerns of robust statistics, PCA is fragile due to the use of a quadratic loss function over the entire dataset. In theory, even a single severe outlier can lead the method toward a highly biased solution. Practically, real world datasets include unwanted outliers which make the classical PCA unsuitable and often useless [131]. To address this issue, robust PCA (RPCA) algorithms have been proposed in the literature [132–134]. However, they need to know the number of modalities present in the data in advance to be able to deal with outliers, which is not always practical. Generally, two major strategies are adopted to resolve this problem. One is to consider the RPCA as a rank minimization problem and another is based on using a proper robust loss function. This study will focus on RPCA using the latter approach.

In the seminal works [135, 136] the problem of RPCA was revisited as a convex rank minimization problem and it was shown that this problem can be solved under certain conditions. Recently, convex and non-convex approaches have been used to recover the true subspace with the minimum possible number of arithmetic operations which ideally can be similar to the classical PCA [137–139]. However, in theses models, noise is considered as a generalized Gaussian, outliers are sparsely placed in a different matrix and

the singular vectors are reasonably spread, which might not be the case in real world applications. As an example of loss function based approaches, the ℓ_1 norm has been used as an approach to achieve robustness against heavy tail densities but this cannot guarantee the robustness against impulsive outliers [140, 141]. To guarantee the robustness of algorithms in non-Gaussian noises, the general family of M-estimators has also been used in statistics to achieve robustness. For example, Huber's robust loss function was used in [142] to develop a RPCA algorithm. Furthermore, some other robust covariance estimators are used in the literature to estimate the principal components by assigning different weights to different data points [143–145]. The main difference of these types of works with rank minimization algorithm is looking at the problem from a different perspective (suppressing outliers instead of data decomposition) and dealing with worst case scenarios which make them highly robust but not scalable to high dimensional data scenarios [126]. Other different points of view on the RPCA problem were recently proposed in [146, 147] which are based on the coherence of the data. This coherence was measured by computing the inner product between all possible pair of samples or by computing the angle between any two given samples.

This chapter is an extension of [4] where only one algorithm was presented briefly. Particularly, we propose four new highly robust algorithms that extract the principal loading vectors and their corresponding PCs by minimizing a cost function derived from the $\alpha-$divergence between the sample density (obtained using the observed data) and the nominal density model. In the first approach, we adopt the classical approach of RPCA and estimate the covariance matrix using the proposed robust loss function. In the second and third algorithms, we look at the problem differently and sequentially estimate the loading vectors without the orthogonality constraint in a robust framework. Finally, in our last method, we also take into account the interpretability of the PCs by bringing sparsity into play as a constraint when we use the robust loss function. In contrast to most of the other robust approaches, we do not need to know the number of modalities in the data. Our methods focus on the bulk of the data and consider the rest as outliers. The trade-off between the performance in uni-modal data and robustness against outliers can be adjusted by the single parameter α which defines the shape of our cost function.

The main contributions of this chapter can be summarized as follows:

- Establishing the link between the classical PCA algorithm and minimizing the

Kullback-Leibler (KL) divergence and adopting the $\alpha-$divergence as an alternative.

- Deriving the robust loading vector estimates using the proposed robust estimate of the covariance.
- Estimation of the loading vectors using the new proposed loss function derived from the $\alpha-$divergence by ignoring the orthogonality constraint in two different setups; at the vector level and at the entry level outliers.
- Proposing a new robust and sparse PCA algorithm by penalization to make the loading vectors interpretable with consistent sparsity.

The remainder of the chapter is organized as follows. In section 4.2 we provide some background on classical, sparse, robust and probabilistic PCA. In section 4.3 we present our proposed methodologies. In section 4.4 we present comparisons between the classical PCA methods and our proposed methods in various setups and applications. Finally, in section 4.5 we present some concluding remarks.

4.2 Background

4.2.1 Classical PCA

Given a dataset $\mathbf{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_N\} \in \mathbb{R}^{m \times N}$ which includes centered i.i.d. samples $\mathbf{y}_n \in \mathbb{R}^m, n = 1, \cdots, N$, the aim of PCA is to find an $r-$dimensional subspace $\mathbf{U} = \{\mathbf{u}_1, ..., \mathbf{u}_r\} \in \mathbb{R}^{m \times r}$ by solving

$$\hat{\mathbf{u}}_k = \arg\max_{\mathbf{u}} \text{Var}(\mathbf{u}^\top \mathbf{Y}) \ \textbf{s.t.} \ \mathbf{u}^\top \mathbf{u} = 1, \tag{4.1}$$

where $k = 1, ..., r$ and $\mathbf{u}_k^\top \mathbf{u}_j = 0$ for $k \neq j$. In other words, PCA finds those principal loading vectors which the maximum data variability is preserved after projecting the data. An alternative formulation for PCA is based on projection. In this case, the cost takes the form

$$\hat{\mathbf{U}} = \arg\min_{\mathbf{U}} \|\mathbf{Y} - \mathbf{U}\mathbf{U}^\top \mathbf{Y}\|_F^2 \ \textbf{s.t.} \ \mathbf{U}^\top \mathbf{U} = \mathbf{I}_r, \tag{4.2}$$

where $\mathbf{U} \in \mathbb{R}^{m \times r}, m > r$ is an orthonormal basis, $\|.\|_F$ is the Frobenius norm and $\mathbf{I}_r$ is the $r \times r$ identity matrix. The solution of both the above optimization problems is the first

r eigenvectors of the sample covariance matrix ($\frac{1}{N}\mathbf{Y}\mathbf{Y}^\top$). However, the corresponding loading vectors of the above solution are dense and sensitive to the existence of possible outliers in the data.

4.2.2 Sparse PCA

For some applications in image processing (for example hand written digit recognition), the observed variables are sparse meaning that most variables are zero and just a few nonzero values exist [148]. Thus, it is useful to investigate the development of a specific algorithm for deriving loading vectors for such scenarios. Given a centered data matrix $\mathbf{Y} \in \mathbb{R}^{m\times N}$, the sparse PCA algorithm solves the following penalized optimization:

$$\hat{\mathbf{U}}, \hat{\mathbf{V}} = \arg\min_{\mathbf{U},\mathbf{V}} \|\mathbf{Y} - \mathbf{U}\mathbf{V}\|_F^2 + \sum_{j=1}^{m} \pi_j \sqrt{r} \|\mathbf{u}^j\|_2 \ \textbf{s.t.} \ \mathbf{V}\mathbf{V}^\top = \mathbf{I}_r. \tag{4.3}$$

In the case of rank-1 approximation ($r = 1$), the above problem leads to sparse PCA introduced in [149]. In the case of $r > 1$, the above problem becomes preserved sparsity based PCA developed in [129] where $\mathbf{u}^j$ is a $1 \times r$ vector corresponding to the j^{th} row of the matrix $\mathbf{U}$. By penalizing the ℓ_2^1-norm of $\mathbf{u}^j$ it is forced to be 0 for most of the time unless it has a significant role in estimating the data matrix $\mathbf{Y}$. However, the above solution is still sensitive to outliers.

4.2.3 "Low Rank + Sparse" Decomposition

In order to reduce the sensitivity of PCA to outliers, using results in [135, 136, 150], it has been suggested to assume the following structure for data:

$$\mathbf{Y} = \mathbf{L} + \mathbf{S} + \boldsymbol{\epsilon}, \tag{4.4}$$

where $\mathbf{L}$ is a low rank matrix supposed to approximate the static behavior of the data, $\mathbf{S}$ is sparse matrix that contains the outliers and $\boldsymbol{\epsilon}$ contains Gaussian noise. To find the $\mathbf{L}$ and $\mathbf{S}$ matrices, the following convex problem can be solved:

$$\hat{\mathbf{L}}, \hat{\mathbf{S}} = \arg\min_{\mathbf{L},\mathbf{S}} \|\mathbf{L}\|_* + \pi_1 \|\mathbf{S}\|_1 + \pi_2 \|\mathbf{Y} - \mathbf{L} - \mathbf{S}\|_F. \tag{4.5}$$

Here $\|\mathbf{L}\|_*$ is the nuclear norm of $\mathbf{L}$ and measures the sum of singular values of $\mathbf{L}$. $\|\mathbf{S}\|_1$ measures the ℓ_1 norm of the vectorized $\mathbf{S}$. Even though some modifications in the above model have been proposed to increase robustness [151], this optimization problem still deals with norms that are somehow sensitive to outliers [142]. In the next section we introduce the probabilistic principal component used in this chapter together with a robust measure from information geometry to estimate the PCs.

4.2.4 Probabilistic Principal Component Model

Given the dataset $\mathbf{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_N\}$ composed of centered i.i.d. samples $\mathbf{y}_n \in \mathbb{R}^m$, probabilistic PCA (PPCA) estimates a span of the $r-$dimensional subspace that preserves the maximum data variance using the following generative model:

$$\mathbf{y}_i = \mathbf{U}\mathbf{v}_i + \boldsymbol{\xi}_i, \;\; i = 1, \cdots, N. \tag{4.6}$$

In (4.6), $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_r)$ is a low dimensional ($r < m$) Gaussian latent vector, $\mathbf{U}$ is a $m \times r$ matrix of loading vectors and $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_m)$ is a Gaussian noise term independent of $\mathbf{v}$.

In [152], it is shown that the PCs of $\mathbf{Y}$ can be retrieved using the maximum likelihood (ML) estimator of $\mathbf{U}$. Indeed if $\mathbf{A}$ is a $m \times r$ matrix of ordered principal eigenvectors of $\mathbf{Y}\mathbf{Y}^\top$ and $\boldsymbol{\Lambda}$ is the diagonal matrix of the corresponding eigenvalues, then

$$\mathbf{U}_{ML} = \mathbf{A}(\boldsymbol{\Lambda} - \sigma^2\mathbf{I}_r)^{1/2}\mathbf{R}, \tag{4.7}$$

where $\mathbf{R}$ is an arbitrary $(r \times r)$ orthogonal matrix. Since (4.7) is true for all $\sigma > 0$, the limit corresponding to the noiseless setting ($\sigma \to 0$) allows full recovery of principal components. This framework was first introduced in [153] and used in several other studies [154, 155].

4.2.5 From KL Divergence to $\alpha-$Divergence

It is well known that for a large dataset ($N \to \infty$), ML estimation is equivalent to estimation by minimizing the KL divergence:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{ML} &= \arg\max_{\boldsymbol{\beta}} \frac{1}{N} \sum_{i=1}^{N} \log\left(f(\mathbf{y}_i, \boldsymbol{\beta})\right) \\ &= \arg\min_{\boldsymbol{\beta}} KL\left(f(\mathbf{y}, \boldsymbol{\beta}^*), f(\mathbf{y}, \boldsymbol{\beta})\right), \end{aligned} \tag{4.8}$$

where $f(\mathbf{y}, \boldsymbol{\beta}^*)$ is the true unknown density of data. However, the KL divergence is sensitive to deviations from the assumed models and a slight deviation from the assumptions can cause significant changes in the estimates. Thus, as an alternative, we suggest the use of another generalized parametric divergence, the $\alpha-$divergence, to estimate the principal components. This divergence has the advantages of including KL as a special case and offers protection against contaminated noise [156–158].

4.3 Proposed Method

For two probability density functions $g(\mathbf{y}, \boldsymbol{\beta}^*)$ defined as $g(\mathbf{y}, \boldsymbol{\beta}^*) = (1-\epsilon) f(\mathbf{y}, \boldsymbol{\beta}^*) + \epsilon h$; where ϵh characterizes the outliers in the observations, and parametrized density $f(\mathbf{y}, \boldsymbol{\beta})$, the α-divergence is defined as [102]

$$D_\alpha\left(g(\mathbf{y}, \boldsymbol{\beta}^*) \parallel f(\mathbf{y}, \boldsymbol{\beta})\right) = \frac{1}{\alpha(\alpha-1)} \left[\int g(\mathbf{y}, \boldsymbol{\beta}^*)^\alpha f(\mathbf{y}, \boldsymbol{\beta})^{1-\alpha} d\mathbf{y} - 1\right], \tag{4.9}$$

where α is a tuning parameter. D_α is convex with respect to both distributions and is always greater than or equal to zero. Equality holds if and only if $g(\mathbf{y}, \boldsymbol{\beta}^*) = f(\mathbf{y}, \boldsymbol{\beta})$. For the specific case when $\alpha \to 1$, the α-divergence reduces to the KL-divergence on which the traditional PCA is based. Four different points of view in this chapter will be used to develop the proposed robust PCA algorithms.

4.3.1 RPCA using Robust Covariance Estimation

One way to estimate principal loading vectors is to apply eigenvalue decomposition to the sample covariance estimate and find the first eigenvectors. However, this estimate is sensitive to existing outliers in the data. To address this problem, we propose

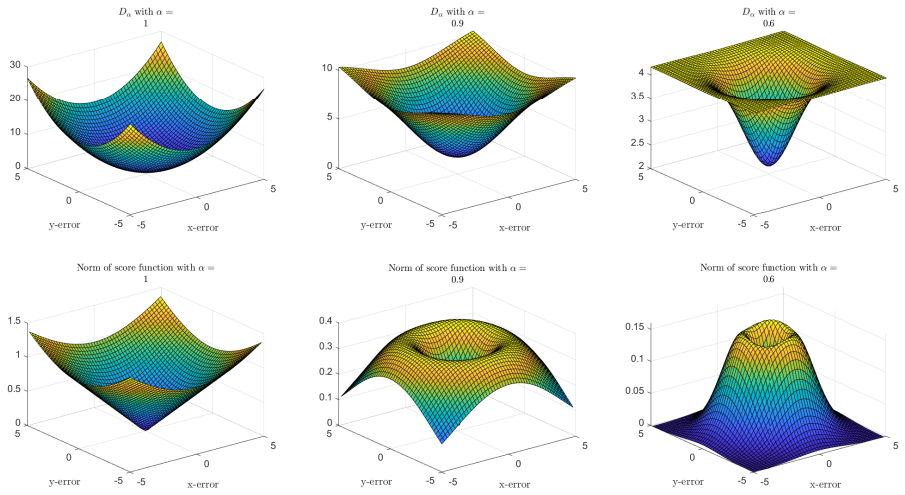

FIGURE 4.1: D_α with its score function for different values of α when $\boldsymbol{\Sigma} = \mathbf{I}_2$.

$\alpha-$divergence as a measure to derive a robust covariance estimator. To do this, we use the multivariate Gaussian density

$$\mathcal{N}(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{m/2}\det(\boldsymbol{\Sigma})^{1/2}} \times \exp\left\{ -\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right\}, \tag{4.10}$$

in (4.9) for $f(\mathbf{y}, \boldsymbol{\beta})$ where m is the dimension of $\mathbf{y}$, and $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ show the mean and the covariance of the data, respectively. The contaminated model of the data can be represented by the sample density function. Mathematically, the sample density function can be defined by

$$p(\mathbf{y}; \mathbf{y}_1 \cdots \mathbf{y}_N) = \frac{1}{N}\sum_{i=1}^{N} \delta(\mathbf{y}-\mathbf{y}_i), \tag{4.11}$$

where $\mathbf{y}$ is a random variable and δ is the Dirac delta function. Considering the sample density function $p(\mathbf{y})$ for $g(\mathbf{y}, \boldsymbol{\beta}^*)$ in (4.9), the mean and covariance of the model can be estimated using a minimization problem obtained by replacing the KL divergence in (4.8) by the $\alpha-$divergence:

$$\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}} = \arg\min_{\boldsymbol{\mu}, \boldsymbol{\Sigma}} D_\alpha \left(p(\mathbf{y}) \parallel \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\right). \tag{4.12}$$

For fixed α, solving (4.12) using an iterative approach, provides estimates of the unknown parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. However, appropriate initial estimates plays an important role in converging to a good local minimum. D_α can be written in the following form using the mentioned sample density and multivariate normal density:

$$D_\alpha\left(p(\mathbf{y}) \parallel \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\right) = \frac{1}{\alpha(1-\alpha)}\Bigg[1 - \sum_{i=1}^{N}(1/N)^\alpha\Big(\frac{1}{(2\pi)^{m/2}\det(\boldsymbol{\Sigma})^{1/2}}\Big)^{1-\alpha} \times \exp\Big\{-\frac{1-\alpha}{2}(\mathbf{y}_i-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i-\boldsymbol{\mu})\Big\}\Bigg]. \tag{4.13}$$

Figure 4.1 shows the shapes of D_α for a two dimensional error term $\mathbf{e} = \mathbf{y} - \boldsymbol{\mu}$ and different values of α and its corresponding score functions when $\boldsymbol{\Sigma} = \mathbf{I}_2$ and $N = 1$. When $\alpha \to 1$ this divergence has a quadratic form where its solution is unique and equal to the solution of classical PCA. On the other hand, decreasing α, gradually generates a cost function which is less sensitive to large error values. The norm of the score function, using the $\alpha-$divergence (directly related to the influence function (IF) of the estimators [10]) is becoming less sensitive to large residuals. The score function is both continuous and bounded. Both of these properties are desired in a robust statistic [40]. The $\alpha-$divergence provides a redescending score function that is robust against severe outliers. Compared to the proposed robust loss function of [159], this loss depends only on a single parameter, is supported by the information geometry [102] and can be easily tuned. We adopt an alternative approach, where at each step one parameter is fixed and we sequentially solve:

$$\begin{aligned}\frac{\partial D_\alpha}{\partial \boldsymbol{\mu}} &= k_1\sum_{i=1}^{N} w_i(\boldsymbol{\mu}-\mathbf{y}_i) = \mathbf{0}\\ \frac{\partial D_\alpha}{\partial \boldsymbol{\Sigma}} &= k_2\sum_{i=1}^{N} w_i(\boldsymbol{\Sigma}-(\mathbf{y}_i-\boldsymbol{\mu})(\mathbf{y}_i-\boldsymbol{\mu})^\top) = \mathbf{0}.\end{aligned} \tag{4.14}$$

k_1 and k_2 are positive constants that depend on α, N, m, and the $\det(\boldsymbol{\Sigma})$. Finally, the robust mean and the robust covariance at step t are computed using

$$\boldsymbol{\mu}^t = \frac{\sum_{i=1}^{N} w_i^t \mathbf{y}_i}{\sum_{i=1}^{N} w_i^t}, \tag{4.15}$$

$$\boldsymbol{\Sigma}^t = \frac{\sum_{i=1}^{N} w_i^t (\mathbf{y}_i-\boldsymbol{\mu}^{t-1})(\mathbf{y}_i-\boldsymbol{\mu}^{t-1})^\top}{\sum_{i=1}^{N} w_i^t}. \tag{4.16}$$

w_i^t shows the weight of $\mathbf{y}_i$ at step t in the estimation of parameters and can be computed by $w_i^t = \exp\{-\frac{1-\alpha}{2}(\mathbf{y}_i-\boldsymbol{\mu}^{t-1})^\top(\boldsymbol{\Sigma}^{\boldsymbol{t-1}})^{-1}(\mathbf{y}_i-\boldsymbol{\mu}^{t-1})\}$. Referring to model (4.6), we can see $\boldsymbol{\Sigma} = \mathbf{U}\mathbf{U}^\top + \sigma^2\mathbf{I}_m$. Then the same conclusion can be made, i.e., $\hat{\mathbf{U}} = \mathbf{A}\left(\boldsymbol{\Lambda} - \sigma^2\mathbf{I}_r\right)^{1/2}\mathbf{R}$, where $\mathbf{A}$ is obtained using the first sorted eigenvectors of $\boldsymbol{\Sigma}$.

Algorithm 1 Robust PCA with Robust Covariance (A1)

Input: Non centered dataset $\mathbf{Y}$, α, r, th
Output: $\mathbf{U}$
Procedure:
Initialize $\boldsymbol{\mu}^1$ and $\boldsymbol{\Sigma}^1$, compute the $\mathbf{w}$ and set $t = 1$
while $\|\mathbf{w}^t - \mathbf{w}^{t-1}\|_2 \geq th$ **do**
$t \leftarrow t + 1$
$w_i^t \leftarrow \exp\{-\frac{1-\alpha}{2}(\mathbf{y}_i - \boldsymbol{\mu}^{t-1})^\top (\boldsymbol{\Sigma}^{\boldsymbol{t-1}})^{-1}(\mathbf{y}_i - \boldsymbol{\mu}^{t-1})\}$
$\boldsymbol{\mu}^t \leftarrow \frac{\sum_{i=1}^N w_i^t \mathbf{y}_i}{\sum_{i=1}^N w_i^t}$
$\boldsymbol{\Sigma}^t \leftarrow \frac{\sum_{i=1}^N w_i^t (\mathbf{y}_i - \boldsymbol{\mu}^{t-1})(\mathbf{y}_i - \boldsymbol{\mu}^{t-1})^\top}{\sum_{i=1}^N w_i^t}$
end while
$\mathbf{U} \leftarrow$ r-dominant eigen($\boldsymbol{\Sigma}$)
Output $\leftarrow \mathbf{U}$

$\boldsymbol{\Lambda}$ is the corresponding eigenvalues and $\mathbf{R}$ is an arbitrary $(r \times r)$ orthogonal rotation matrix. The loading vectors $\mathbf{U}$ in the limit case $\sigma \to 0$ with $\mathbf{R} = \mathbf{I}_r$, correspond to the first r eigenvectors of the estimated covariance matrix $\boldsymbol{\Sigma}$ given in (4.16). σ^2 can be estimated using $\hat{\sigma}^2 = \frac{1}{m-r}\sum_{i=r+1}^m \lambda_i$ where λ_i is the i^{th} smallest eigenvalue of estimated $\boldsymbol{\Sigma}$. Clearly, when $\alpha \to 1$ then $w_i^t \to 1$ and the estimators will be the sample mean and the sample covariance that were derived using the ML approach for the classical PCA [152]. The step-by-step procedure of the proposed method is presented in **Algorithm 1**.

4.3.2 Direct Robust Low Rank Approximation

Assume the probabilistic PCA model $\mathbf{y}_i = \mathbf{U}\mathbf{v}_i + \boldsymbol{\xi}_i, i = 1, \cdots, N$, where $\boldsymbol{\xi}_i = \mathbf{y}_i - \mathbf{U}\mathbf{v}_i$ is the nominal zero mean Gaussian white residual vector with variance σ^2. It has been shown that minimizing the $\alpha-$divergence D_α with respect to unknown parameters $\boldsymbol{\beta}$ is equivalent to the following maximization [112]

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \frac{1}{N}\sum_{i=1}^N \log_\alpha \{f(\mathbf{y}_i, \boldsymbol{\beta})\}, \tag{4.17}$$

where $\log_\alpha(x) = \frac{x^{(1-\alpha)}-1}{1-\alpha}$. Therefore

Algorithm 2 Robust PCA with Low Rank Approx. (A2)

Input: Centered dataset $\mathbf{Y}$, α, r, th
Output: $\mathbf{U}$,$\mathbf{V}$
Procedure:
Initialize $\mathbf{U}$ and $\mathbf{V}$, compute $\mathbf{w}$ and set $t = 1$
while $\|\mathbf{w}^t - \mathbf{w}^{t-1}\|_2 \geq th$ **do**
  $t \leftarrow t + 1$
  Fix $\mathbf{V}$ and estimate $\mathbf{U}$ using (4.18)
  Normalize $\mathbf{U}$
  Fix $\mathbf{U}$ and estimate $\mathbf{V}$ using (4.19)
end while
Output $\leftarrow \mathbf{U}, \mathbf{V}$

$$\begin{aligned}
\hat{\mathbf{U}}, \hat{\mathbf{V}} &= \arg\max_{\boldsymbol{\beta}} \frac{1}{N} \sum_{i=1}^{N} \log_\alpha \{f(\mathbf{y}_i, \mathbf{U}, \mathbf{v}_i)\} \\
&= \arg\max_{\boldsymbol{\beta}} \frac{1}{N} \sum_{i=1}^{N} \frac{f(\mathbf{y}_i, \mathbf{U}, \mathbf{v}_i)^{(1-\alpha)} - 1}{1-\alpha} \\
&= \arg\max_{\boldsymbol{\beta}} \frac{1}{N} \sum_{i=1}^{N} \frac{\exp\{-\frac{1-\alpha}{2\sigma^2} \boldsymbol{\xi}_i^\top \boldsymbol{\xi}_i\} - 1}{1-\alpha} \\
&= \arg\max_{\boldsymbol{\beta}} \frac{1}{N} \sum_{i=1}^{N} \frac{\exp\{-\frac{1-\alpha}{2\sigma^2} \|\mathbf{y}_i - \mathbf{U}\mathbf{v}_i\|_2^2\} - 1}{1-\alpha},
\end{aligned}$$

where $\mathbf{v}_i$ is the i^{th} column of matrix $\mathbf{V}$. Using the block coordinate descent algorithm, $\mathbf{U}$ and $\mathbf{V}$ can be estimated iteratively using

$$\hat{\mathbf{U}} = \left(\sum_{i=1}^{N} w_i \mathbf{y}_i \mathbf{v}_i^\top \right) \left(\sum_{i=1}^{N} w_i \mathbf{v}_i \mathbf{v}_i^\top \right)^{-1} \tag{4.18}$$

followed by a normalization step and by

$$\hat{\mathbf{v}}_i = \left(\mathbf{U}^\top \mathbf{U} \right)^{-1} \mathbf{U}^\top \mathbf{y}_i, \; i = 1, \cdots, N, \tag{4.19}$$

where $w_i = \exp\{-\frac{1-\alpha}{2}\|\mathbf{y}_i - \mathbf{U}\mathbf{v}_i\|_2^2\}$. This procedure converges after only a few steps. The step-by-step procedure of the proposed method is presented in **Algorithm 2**. Note that this method assigns a single weight to the given observation vector. This means having outliers in one single entry (e.g., outliers initiated by faulty performance of a single sensor in an array of sensors) can result in loss of the entire measurement vector. In order to address this issue, starting from expression (4.9) and using the low rank approximation of the centered data matrix $\mathbf{Y}$, it is straightforward to show the first robust principal loading vector ($\mathbf{u}_1 \in \mathbb{R}^{m \times 1}$) and the corresponding principal component

Algorithm 3 Robust PCA with Low Rank Approx. (A3)

Input: Centered dataset $\mathbf{Y}$, α, r, th
Output: U,V
Procedure:
Initialize $\mathbf{U}$ and $\mathbf{V}$
for i=1 **to** r **do**
  Set $t = 1$
  Assign $\mathbf{E}_i = \mathbf{Y} - \sum_{j=1}^{i-1} \mathbf{u}_j \mathbf{v}^j$
  while $\|\mathbf{vec}(\mathbf{W}^t) - \mathbf{vec}(\mathbf{W}^{t-1})\|_2 \geq th$ **do**
    $t \leftarrow t + 1$
    Update $\mathbf{W}$ and $\mathbf{K}$ using expressions (4.23) and (4.24)
    Update $\mathbf{u}$ using (4.21)
    $\mathbf{u} \leftarrow \frac{\mathbf{u}}{\|\mathbf{u}\|_2}$
    Update $\mathbf{v}$ using (4.22)
  end while
end for
$\mathbf{U} \leftarrow [\mathbf{u}_1, \cdots, \mathbf{u}_r]$
$\mathbf{V} \leftarrow [\mathbf{v}^{1\top}, \cdots, \mathbf{v}^{r\top}]^\top$
Output $\leftarrow \mathbf{U}, \mathbf{V}$

($\mathbf{v}^1 \in \mathbb{R}^{1\times N}$) can be derived by solving following optimization problem:

$$\hat{\mathbf{u}}_1, \hat{\mathbf{v}}^1 = \arg\min_{\mathbf{u},\mathbf{v}} \sum_{j=1}^{m}\sum_{i=1}^{N} l_\alpha(y_{ji} - u(j)v(i)) \text{ s.t. } \mathbf{u}_1^\top \mathbf{u}_1 = 1, \tag{4.20}$$

where $l_\alpha(e) = \frac{1}{1-\alpha}(1 - \exp\{-\frac{1-\alpha}{2}e^2\})$. The estimation of $\mathbf{u}_i, i = 2, \cdots, r$, is achieved by replacing $\mathbf{Y}$ with $\mathbf{E}_i = \mathbf{Y} - \sum_{j=1}^{i-1} \mathbf{u}_j \mathbf{v}^j$ in (4.20). Assuming $\mathbf{v}^1$ known, the j^{th} entry of $\mathbf{u}_1$ before normalization at step t is given by

$$u_1^t(j) = (\mathbf{v}_{t-1}^1 \mathbf{W}_j^t {\mathbf{v}_{t-1}^1}^\top)^{-1} \mathbf{v}_{t-1}^1 \mathbf{W}_j^t {\mathbf{y}^j}^\top, j = 1, \cdots, m, \tag{4.21}$$

where $\mathbf{y}^j$ is the j^{th} row of the matrix $\mathbf{Y}$. Similarly, $v_t^1(i)$ is given by

$$v_t^1(i) = \frac{{\mathbf{u}_1^{t-1}}^\top \mathbf{K}_i^t \mathbf{y}_i}{{\mathbf{u}_1^{t-1}}^\top \mathbf{K}_i^t \mathbf{u}_1^{t-1}}, \ i = 1, \cdots, N, \tag{4.22}$$

where $\mathbf{y}_i$ is the i^{th} column of matrix $\mathbf{Y}$ and $\mathbf{u}_1$ is the normalized vector of (4.21). $\mathbf{W}$ and $\mathbf{K}$ are two weight tensors with dimension $N \times N \times m$ and $m \times m \times N$, respectively and indices j and i in (4.21) and (4.22) are on the third index of these tensors. The weights can be computed by

$$\mathbf{W}^t(k,k,i) = \exp\{-\frac{1-\alpha}{2}(y_{ik} - u_1^{t-1}(i)v_{t-1}^1(k))^2\}, \tag{4.23}$$

and

$$\mathbf{K}^t(i,i,k) = \exp\{-\frac{1-\alpha}{2}(y_{ik} - u_1^{t-1}(i)v_{t-1}^1(k))^2\}, \tag{4.24}$$

where $i = 1, \cdots, m$ and $k = 1, \cdots, N$ and the other entries are zero. Note that the loading vectors that are estimated from (4.20) are not necessarily orthogonal. The step by step procedure of the proposed procedure is presented in **Algorithm 3**.

4.3.3 Robust Sparse PCA

Using the above loss function in optimization problem (4.3), gives

$$\hat{\mathbf{U}}, \hat{\mathbf{V}} = \arg\min_{\mathbf{U},\mathbf{V}} l_\alpha(\mathbf{Y} - \mathbf{UV}) + \sum_{j=1}^{m} \pi_j \sqrt{r} \|\mathbf{u}^j\|_2 \ \textbf{s.t.} \ \mathbf{V}\mathbf{V}^\top = \mathbf{I}_r, \tag{4.25}$$

where

$$l_\alpha(\mathbf{Y} - \mathbf{UV}) = \sum_{j=1}^{m}\sum_{i=1}^{N} l_\alpha(y_{ji} - \mathbf{u}^j\mathbf{v}_i). \tag{4.26}$$

The first term is the robust loss function and aims to fit the model to the data without heavily penalizing the large residual and the second term aims to enforce sparsity on the loading vectors. Note that this proposed method is useful anywhere where sparse patterns exist. We present the results of this algorithm only in applications where sparsity makes sense. Similar to previous setups $\mathbf{U} \in \mathbb{R}^{m\times r}$ and $\mathbf{V} \in \mathbb{R}^{r\times N}$. The $\mathbf{u}^j \in \mathbb{R}^{1\times r}$ and $\mathbf{v}_i \in \mathbb{R}^{r\times 1}$ show the j^{th} row and i^{th} column of matrices $\mathbf{U}$ and $\mathbf{V}$, respectively. The block coordinate descent approach is used to solve the problem defined in (4.25). Assuming $\mathbf{V}$ is known, the problem is to find $\mathbf{U}$ such that

$$\hat{\mathbf{U}} = \arg\min_{\mathbf{U}} l_\alpha(\mathbf{Y} - \mathbf{UV}) + \sum_{j=1}^{m} \pi_j \sqrt{r} \|\mathbf{u}^j\|_2. \tag{4.27}$$

After some simplifications provided in Appendix C, it is easy to show that $\hat{\mathbf{u}}^j$, $j = 1, \cdots, m$ is given by

$$\begin{cases} \mathbf{y}^j\mathbf{W}\mathbf{V}^\top \left(\frac{\pi_j\sqrt{r}}{\|\mathbf{u}^j\|_2}\mathbf{I}_r + \mathbf{V}\mathbf{W}\mathbf{V}^\top\right)^{-1}, & \text{if } \|\mathbf{y}^j\mathbf{W}_0\mathbf{V}^\top\|_2 \geq \pi_j\sqrt{r} \\ \mathbf{0}, & \text{otherwise} \end{cases}$$

where $\mathbf{W}$ is a diagonal weight matrix with diagonal elements $w_i = \exp\{-\frac{1-\alpha}{2}\left(y_{ji} - \mathbf{u}^j\mathbf{v}_i\right)^2\}$ and $\mathbf{W}_0 = \mathbf{W}|_{\mathbf{u}^j=\mathbf{0}}$. Note that the above estimator is also a function of the $\|\mathbf{u}^j\|_2$ which

Algorithm 4 Robust Sparse PCA (A4).

Input: Centered dataset $\mathbf{Y}$, α, r, η, ρ, π, th
Output: $\mathbf{U}$,$\mathbf{V}$
Procedure:
Initialize $\mathbf{U}$ and $\mathbf{V}$ and set $t = 1$
while $\|\mathbf{vec}(\mathbf{U}^t) - \mathbf{vec}(\mathbf{U}^{t-1})\|_2 \geq th$ **do**
  $t \leftarrow t + 1$
  Fix $\mathbf{V}$ and estimate $\mathbf{U}$ using (4.27)
  Fix $\mathbf{U}$ and estimate $\mathbf{V}$ using (4.28)
end while
Output $\leftarrow \mathbf{U}, \mathbf{V}$

need to be updated at each step until convergence. When estimation of $\mathbf{U}$ is performed, we assume $\mathbf{U}$ is known and estimate $\mathbf{V}$ using following optimization

$$\hat{\mathbf{V}} = \arg\min_{\mathbf{V}} l_\alpha(\mathbf{Y} - \mathbf{UV}) \ \textbf{s.t.} \ \mathbf{V}\mathbf{V}^\top = \mathbf{I}_r. \tag{4.28}$$

This problem is a Stiefel Manifold constrained optimization which can be solved by following the recursive steps [160]:

- Estimate $\mathbf{V}$ at step $t+1$ using $\mathbf{V}_t$ by expression $\mathbf{V}_{t+1} = \mathbf{V}_t\mathbf{Q}$ where $\mathbf{Q} = (\mathbf{I}_N - \eta\mathbf{A})(\mathbf{I}_N + \eta\mathbf{A})^{-1}$ and $\mathbf{A} = \mathbf{V}_t^\top \nabla l_\alpha - \nabla l_\alpha^\top \mathbf{V}_t$.
- Update the Gradient, $\mathbf{A}$ and $\mathbf{Q}$ at each step.
- $t = t + 1$ and repeat above steps until convergence.

It is guaranteed that by choosing a proper η, the above procedure leads to a local minimum of (4.28) (see Appendix D). The step-by-step procedure of the proposed method is presented in **Algorithm 4**.

4.3.4 How to Choose α?

The hyperparameter α establishes compromise between the performance of the nominal model and robustness against outliers in the data. Thus, this value can be specified by the user based on the desired efficiency. Tuning this parameter can be done using exactly the same procedure of tuning the parameter of Huber's loss function. Assume relative efficiency of $p = 90\% - 95\%$ is desired in the nominal model. Therefore, we need

to find the solution of the equality

$$\frac{\text{tr}(\mathbf{I}^{-1})}{\text{tr}(\mathbf{V}(\alpha))} = p, \tag{4.29}$$

where $\mathbf{I}$ is the Fisher information matrix, $\mathbf{V}$ is the asymptotic variance of estimated parameters using the proposed robust loss function and tr(.) is the trace of a matrix. This parameter can also be determined by having prior information about the maximum fraction of outliers (worst case scenario consideration). In this case, α can be determined such that ϵ percent of data has low weight. More details are provided in the simulation results.

4.3.5 How to Choose π_j?

In this part we suggest an automatic approach for penalizing each row of the loading vectors. Intuitively, most of the time those rows which have a larger ℓ_2^1-norm, have significant effect on approximation of data and should be less penalized. Therefore, we use the following formulation for setting π_j:

$$\pi_j = \frac{\pi}{\|\mathbf{u}^j\|_2^\rho}, \ j = 1, \cdots, m, \tag{4.30}$$

where π and ρ are hyperparameters of the method. Naturally, when $\pi \to 0$, there is no sparsity constraint on the loading vectors and they can get any values. On the other hand by setting $\pi \to \infty$, all elements of the loading vectors will be forced to zero to get the highest possible sparsity (zeros everywhere). ρ can change the sensitivity of the π_j to the $\|\mathbf{u}^j\|_2$ and we set it to 1.

4.4 Simulation Results

In this section we present an empirical study on the proposed methods. To do this, we consider three different synthetic and two real datasets. In the first synthetic dataset, two and three dimensional samples are generated and contaminated with outliers to examine the proposed methods. In the second synthetic dataset, we show how the proposed methods can be used to recover the fMRI signals and finally in last synthetic dataset we investigate the the performance of the methods in the foreground-background

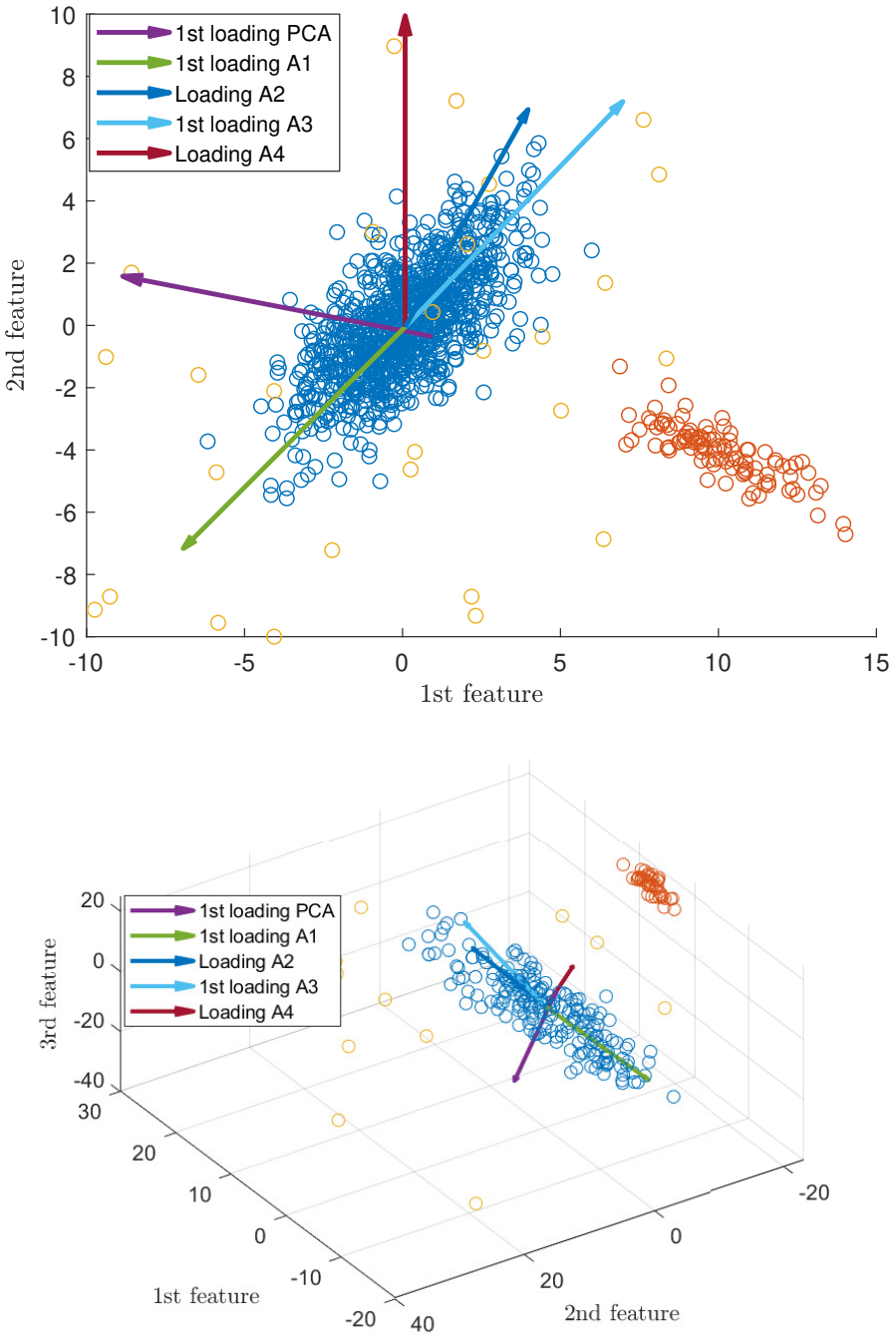

FIGURE 4.2: Scatter plot of simulated data (blue, red and yellow circles show the original data, structured and unstructured outliers, respectively) and estimated loading vectors using classical and robust PCA algorithms in 2D and 3D feature space.

(FB) separation task. To illustrate real applications, the proposed methods are also applied to video and image processing problems.

4.4.1 Dataset 1

In order to show the performance of the proposed algorithms, a total of 1130 samples generated using Monte Carlo simulations shown in figure 4.2(top). 1000 of these samples come from a multivariate Gaussian distribution with the mean vector $\boldsymbol{\mu}_1 = (0, 0)^\top$ and the covariance of $\boldsymbol{\Sigma}_1 = \begin{pmatrix} 3.0 & 2.1 \\ 2.1 & 3.0 \end{pmatrix}$ which forms the bulk of the data (blue circles) and 100 samples come from another multivariate Gaussian distribution function with the mean vector $\boldsymbol{\mu}_2 = (10, -4)^\top$ and the covariance of $\boldsymbol{\Sigma}_2 = \begin{pmatrix} 2.6 & -1.2 \\ -1.2 & 0.8 \end{pmatrix}$ to form the structured outliers (red circles). Finally, 30 random samples generated using a uniform distribution are used to model unstructured outliers (yellow circles). The term unstructured outliers refers to data points that are highly uncorrelated with other data points; this can be seen as a uniform distribution over the space. Conversely, the structured outliers represent data points that are highly correlated with each other but not correlated with other data points. They are like a small cluster in datasets. As it is shown, the proposed RPCA algorithms (with $\alpha = 0.8$) correctly recovered the underlying subspace characterizing the majority of data and are not sensitive to both types of structured and unstructured outliers. Note that the direction of the first loading vector shows the direction that achieves the maximum variance when projecting the blue circles on this direction. On the other hand, the classical PCA, due to outliers, provides a different direction that maximizes the projected variance of the entire dataset including outliers. Beyond the directions of the loading vectors, the mean value that is estimated using A1, is correctly presenting the center of blue circles (the origin of the robust loading vectors in figure 4.2). The same simulation is also carried out for a 3D feature space using fewer samples. We can see how the true direction of the original data is recovered using the proposed methods while PCA recovers an irrelevant direction. Note that in A4, except for fitting the model to the data, there is a sparity constraint. This explains why in both simulations loading vectors are in parallel with at least one of the axes.

4.4.2 Synthetic fMRI Data

In this section, we use the general linear model (GLM) to generate synthetic i.i.d. functional magnetic resonance imaging (fMRI) data $\mathbf{y}_i, i = 1, \cdots, N$ in following form

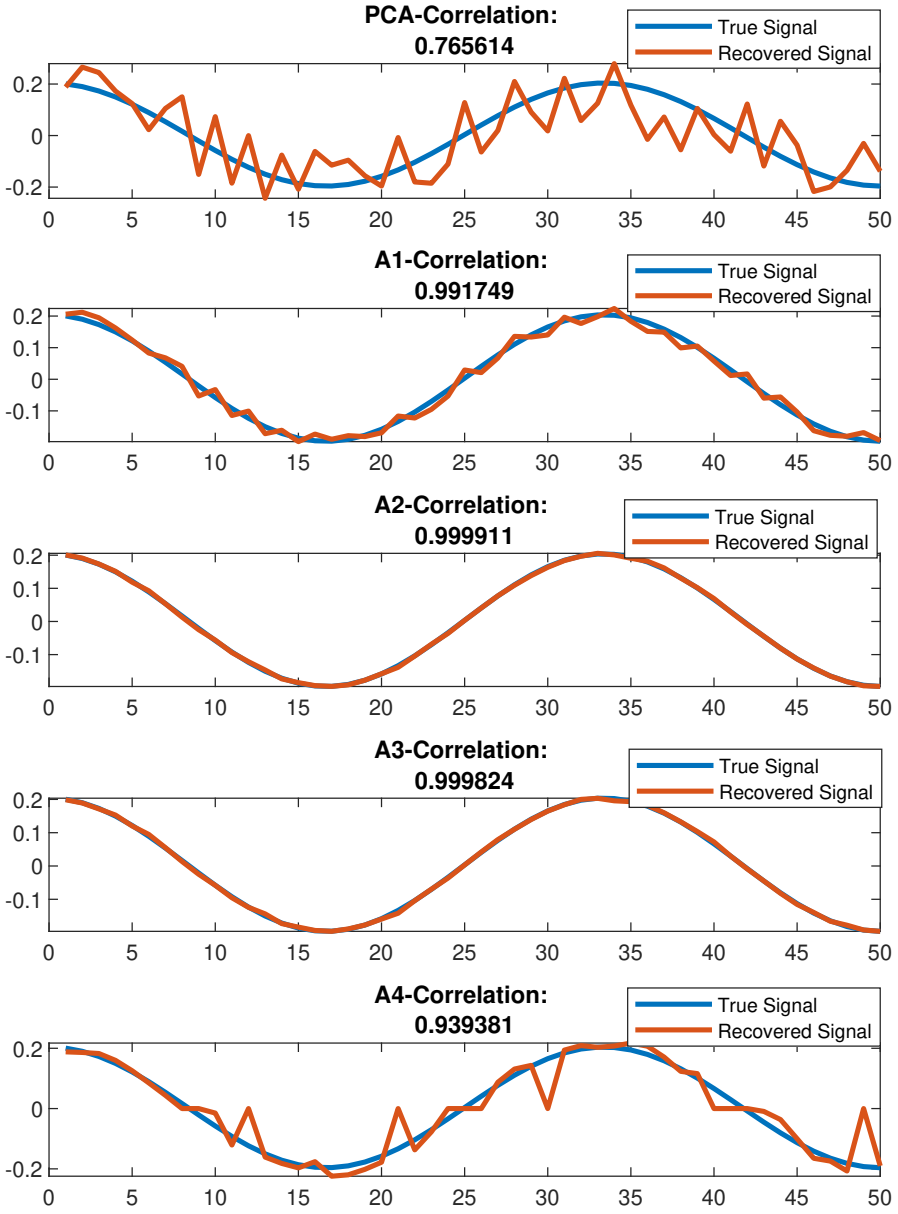

FIGURE 4.3: Recovered signals (red plots) compared to the true non sparse signal (blue plots) using classical PCA (first row) and proposed robust PCA algorithms. It can be observed that using the robust approaches, the recovered signals are more similar to the true signal (the correlation values are reported on top of each plot).

$$\mathbf{y}_i = \mathbf{H}\boldsymbol{\theta} + \mathbf{n}_i, \;\; i = 1, \cdots, N. \tag{4.31}$$

$\mathbf{H}$ is a matrix whose columns span the signal subspace and $\mathbf{n}_i$ is the noise vector which nominally comes from a Gaussian distribution but sometimes it follows an unknown outliers density h. In this model, $\boldsymbol{\theta}$ is the unknown and needs to be estimated. Typically, the main goal in fMRI data analysis is to find the active area of the brain for a specific task. To be able to apply existing methods in detection theory to the fMRI data, first, we need to exactly know the matrix $\mathbf{H}$. Therefore, we need to have an initial estimate

TABLE 4.1: Accuracy of GLRT (in %) using estimated loading vectors.

Methods	PCA	A1	A2	A3	A4
Non Sparse	81.66	**93.33**	**93.33**	**93.33**	92.5
Sparse	77.5	**92.5**	**92.5**	**92.5**	**92.5**

of $\mathbf{H}$ and the accuracy of the used estimator affects the overall detection performance. There are different approaches to approximate $\mathbf{H}$ in practice. One of the easiest approaches is convolving the hemodynamic response function (HRF) with input stimuli [8]. However, for resting state fMRI data analysis or other tasks similar to watching TV, such stimulus function does not exist. Therefore, a data driven method is needed to estimate the dominant patterns in the given time series. Generally, this problem is an ill-posed problem and extra assumptions are needed to enable us to recover the underlying subspace. These assumptions include principal component analysis (PCA) [161], independent component analysis (ICA) [27], canonical correlation analysis (CCA) [162] and sparse dictionary learning (DL) [32]. So far it was shown that the standard PCA is sensitive to model deviations. Thus, a direct use of this approach might not be effective in analyzing the fMRI data where different types of outliers exist; therefore a robust technique is required. To illustrate the applicability of our proposed methods to the recovery of columns of the matrix $\mathbf{H}$, 120 samples with $m = 50$ are generated with 15% contamination using the $\epsilon-$contamination noise model $g = (1-\epsilon)f + \epsilon h$ where ϵ determines the level of deviation from the nominal noise model f by the outliers density h.

In first simulation setups, $\mathbf{H}$ is set to a rank-1 subspace where it is made by using a cosine function. After applying the proposed methods, the correlations of recovered patterns and the true signal are computed and the most similar ones are presented in figure 4.3. As is shown, the proposed algorithms perform better in terms of the correlation between recovered and the true signal compared to classical PCA. This signal estimation leads toward a better detection rate in the next steps of analysis (see table 4.1). However, in this setup A4 is not as powerful as the other proposed methods due to non sparsity of the true signal. This behavior is clear in figure 4.3 by having flat zero

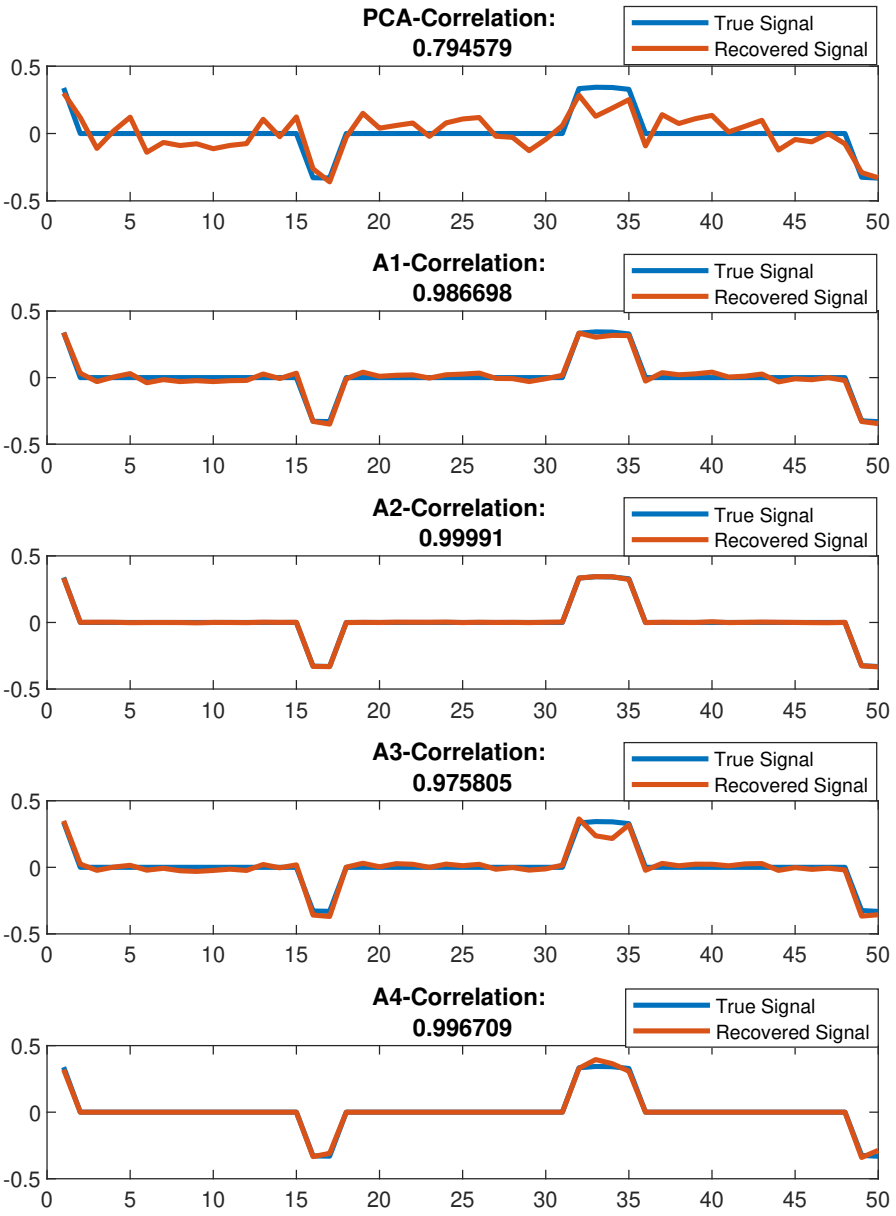

FIGURE 4.4: Recovered signals (red plots) compared to the true sparse signal (blue plots) using classical PCA (first row) and proposed robust PCA algorithms. It can be observed that using the robust approaches, the recovered signals are more similar to the true signal (the correlation values are reported on top of each plot).

values every time when the true signal is passing the zero value and also a zero forcing effect on some other points. Therefore, we repeat the simulation one more time using a sparse signal by applying a threshold on initial cosine function and forcing most of entries to be zero. The results of the algorithms are shown in figure 4.4 and it is clearly seen that under sparsity, A4 is among the best results. Our proposed methods belong to a general family of M-estimators. M-estimators, use a modified cost function, and try to minimize the effect of outliers by less penalizing high residual values. We are

trying to down weight outliers by computing an exponential function of residuals while keeping the weights of other observations close to 1 to have different contributions in the estimation.

The performance of the methods depends on the single hyperparameter α which can change the shape of the cost function. Generally, when $\alpha \to 1$ the methods are the same as classical PCA, for example, the estimated covariance in A1 will be exactly the sample covariance. For any other values of $\alpha < 1$, the methods provide robustness where their sensitivity depend on the value of $1 - \alpha$. By increasing this difference, they will reject more samples in the estimation part. The value of α can be adjusted in different approaches as described above. For example, if users are aware of heavy outliers during recording data, they should pick an α that is far away from 1. As an automatic approach, the user can define the maximum deviation from the nominal model as ϵ_m and tune α in a way that the maximum $\epsilon_m\%$ of data get weights less than a chosen value $1 \geq \beta \geq 0$. A suggestion for β can be 0.5. The value of α can be computed by using $\beta = \exp\{-\frac{1-\alpha}{2}(\mathbf{y}_j - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y}_j - \boldsymbol{\mu})\}$ where $\mathbf{y}_j$ is the closest sample among $\epsilon_m\%$ of data to the bulk.

4.4.3 Synthetic Foreground Background Separation

Before applying our methods on real data for the FB separation task we first show their performances on a synthetic video to have an understanding of the effect of object size, object speed, and contamination level on the final accuracy. In order to do this, 200 dark frames of size of 80×80 are combined with a white Gaussian noise and after placing two squared white objects of different sizes in some proportions of the frames, they are passed through a low pass filter (an example frame is shown in figure 4.5). The paths of two objects are defined randomly and in each run they are different. Accuracy is measured as

$$Accuracy = \frac{TP + TN}{\text{Total pixels}}, \tag{4.32}$$

where TP shows the number of foreground pixels that are correctly assigned to the foreground and TN is the number of background pixels that are correctly assigned to the background. Foreground is detected as any part of the pixel values that do not belong to the background (with rank= 3) and exceed the threshold value of 0.4. To be able to apply the proposed method A1 to this application, the first necessary condition is to

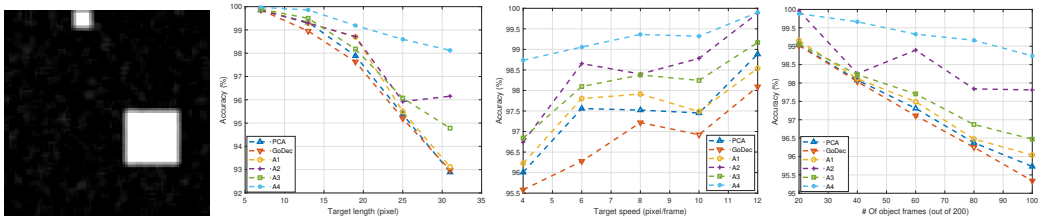

FIGURE 4.5: Comparison between different methods on the task of FB separation on synthetic data (first figure from the left shows an example frame). The experiment is done in various scenarios of different object size (second column), different object speed (third column) and different proportions of frames with objects and background (last column).

have a higher number of frames than the number of pixels which is not satisfied here. The reason is that for computing the weights of the observations, we need to compute $\mathbf{\Sigma}^{-1}$ which means the covariance matrix should be a non-singular matrix. To deal with this issue, we assume outliers have different statistical behavior compared to the rest of the data even after some linear projection. Therefore, we can meet the requirement if we reduce the dimension, using the first eigenvectors of the singular matrix computed using (4.16). The number of bases is picked such that it is less than the number of frames.

In the first experiment, we have 100 frames including the objects one with the size of 7×7 and speed of 12 Pixels/Frame and one with different sizes and the speed of 3 Pixels/Frame. The results are shown in figure 4.5(left). As is shown, with increasing object or target size, the performances of all algorithm decrease due to minor variation of interior pixels and leaking to the first loading vectors. It is also shown that in this experimental setup, the proposed methods are always better than classical approaches and sparse PCA (A4) is significantly better. Surprisingly, the conventional PCA algorithm is even better than the GoDec algorithm [163] which is designed for robust applications. To implement the GoDec algorithm the online version is used[1]. In the second experiment, we have 100 frames including the objects one with the size of 7×7 and speed of 12 Pixels/Frame and one with the size of 25×25 and different speeds. The results are shown in figure 4.5(middle). As is clear from the curves, with increasing target speed, the accuracy of all methods increases due to high dynamics that the object achieves and that makes it easily detectable from the background. Again the sparse algorithm and one of our other proposed algorithms (A2) are significantly better. In our last experiment, we evaluate the effect of the number of object frames in representing the background. In this setup, similar to previous scenarios, two objects one with the size of 7×7 and speed

[1]https://sites.google.com/site/godecomposition/matrix/artifact-1

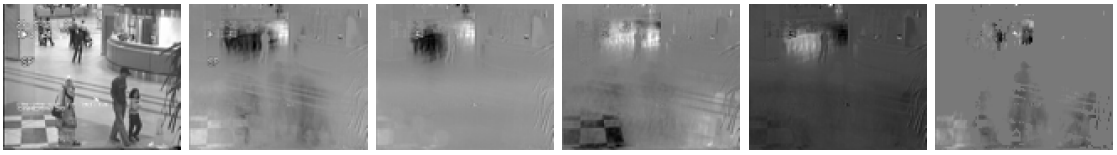

FIGURE 4.6: Comparison between the recovered loading vectors. From the left, frame number of 12, principal component using PCA, first loading vector using A1, loading vector A2, first loading vector using A3 and finally estimated loading vector using A4.

FIGURE 4.7: Foreground-background separation for the airport video (rank=3) . From the left: classical PCA, GoDec, A1, A2, A3, and A4.

of 12 Pixels/Frame and one with the size of 25×25 and the speed of 3 Pixels/Frame are put in different frame proportions. The results are shown in figure 4.5(right). As expected, with increasing the number of frames with objects inside, fewer background representing frames are available which decreases accuracy. Still our proposed methods have uniformly better accuracy compared to two other methods.

4.4.4 Foreground Background Separation

In this section, we apply our methods to the FB separation problem using the first 200 frames of video which are recorded at an airport [164]. Assuming the background changes slowly over the frames, we can use RPCA to approximate the background using a low rank matrix and consider the leftover components as the foreground or objects with higher dynamics. To decrease the run time, the original frames are reduced to an image of 87×106. The frames are vectorized first, then are concatenated to form the data matrix of $\mathbf{Y} \in \mathbb{R}^{9222\times 200}$. The parameter α is set such that 30% of data be considered as outliers with $\beta = 0.4$.

We set the number of iterations to 5, to achieve a stable solution using the classical sample mean as the initial point. Figure 4.6 shows the first eigenvector of the estimated covariance matrix using A1 and the dominant component of the frames using A2, A3 and A4 in comparison with the classical PCA. As expected, the effects of foreground objects

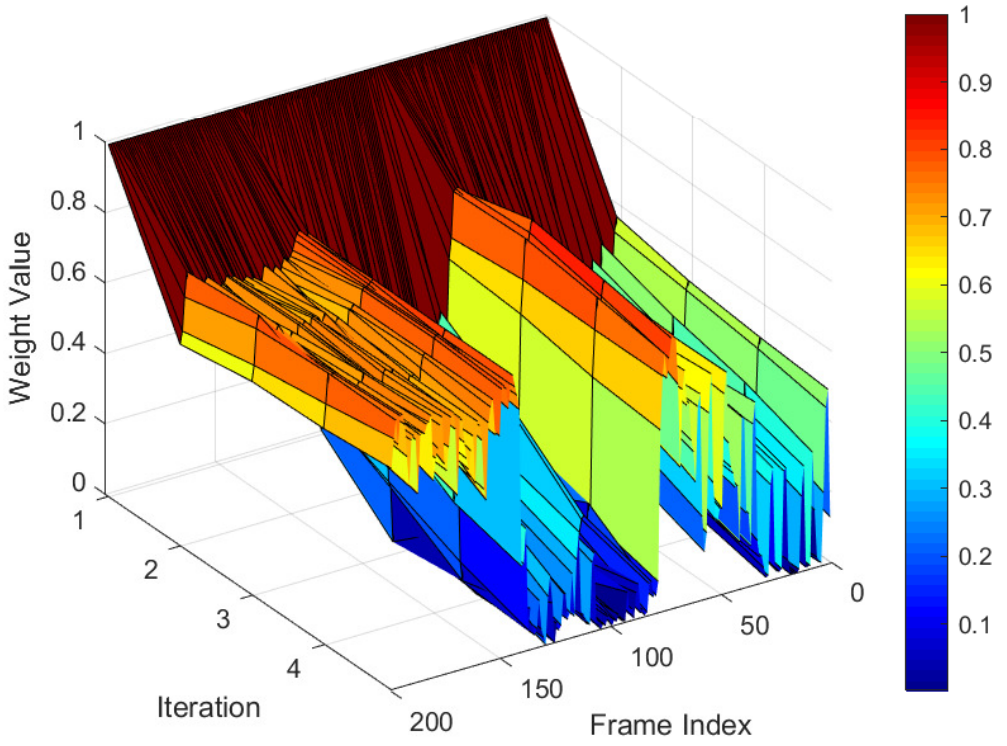

FIGURE 4.8: Dynamic of assigned weights to the frames over the iterations in A1. At iteration 1, all the weights are set to 1 to have the same contribution to estimate parameters. After some iterations, frames with objects got a small weight and frames representing background got weight close to 1.

(the man with his daughter and the woman) are obvious in loading vector of classical PCA (dark shadow effect) where the proposed methods effectively ignored those objects and tried to just represent the background. However, most components in A4 are zero due to its sparse nature.

Due to such differences in loading vectors, the proposed methods can outperform PCA in the task of FB separation. Figure 4.7 illustrates the results of the algorithms on the frame number 12, compared to the classical PCA and the GoDec methods. It can be observed that the proposed methods especially A1 and A2 can effectively separate objects from the background while in the other two classical methods, white and black shadows exist in the estimated background. This is achieved because of assigning different weights to the frames of video to estimate the dominant components (see figure 4.8).

In order to show the superiority of the proposed algorithms in different conditions, we have recorded a video from an intersection including small and big, slow and fast objects, and applied the methods. As a summary, the frames are picked form the middle of the

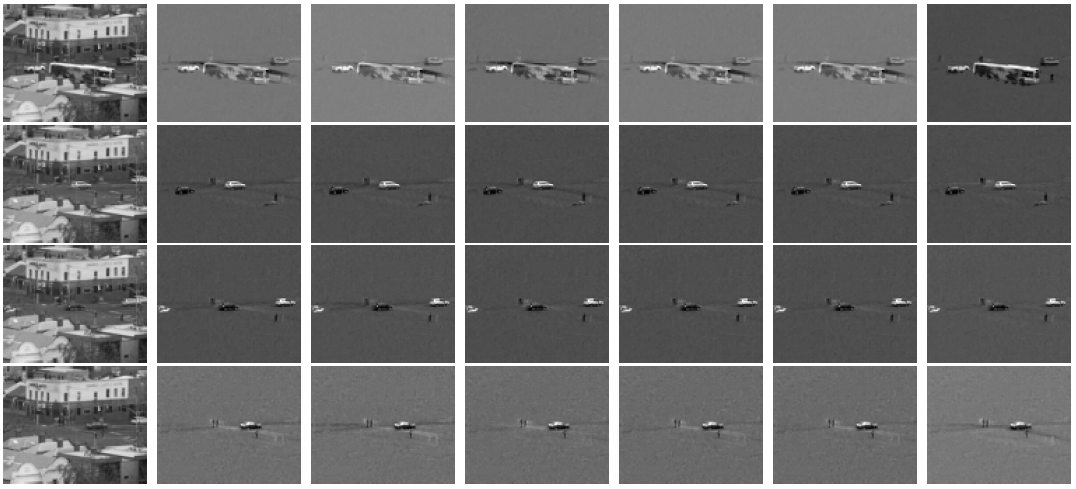

FIGURE 4.9: Comparison on detected foreground between different PCA algorithms. From the left, original frame, detected foreground using PCA, Godec, A1, A2, A3 and finally A4.

video which have more variations (frame 1000 to 1999) and resized to an images of 84×103. The frames are vectorized then are put in a matrix $\mathbf{Y} \in \mathbb{R}^{8652\times1000}$. The foregrounds for all methods are presented in figure 4.9. Those foreground with less shadow effects where just objects (pedestrians, cars and bus) are bold show a better separation which verifies the superiority of our proposed methods over classical approaches.

4.4.5 Saliency Map Identification

A saliency map shows which areas of a given image attract the attention of most humans. It can be thought of as parts of images that are not consistent with other areas. RPCA can be used to approximate the saliency map. In order to perform such a task, we partition images into non overlapped patches with size 10×10. Then we accumulate all patches and form our dataset. Using RPCA algorithms the dominant patterns of data can be extracted and the residual parts after projection represent the pixels that could not be modeled by PCs and are most probable to attract human attention. Applying the same threshold on absolute values of residuals using images provided in MSRA Salient Object Database [165] generates figure 4.10. As is shown, the robust approaches can detect even small differences in pixel values due to assigning different weights to different patches. As we can see, in general, in a local area, a given patch is either within the saliency map or the texture. Therefore, A4 is not a ood choice of algorithm for this application.

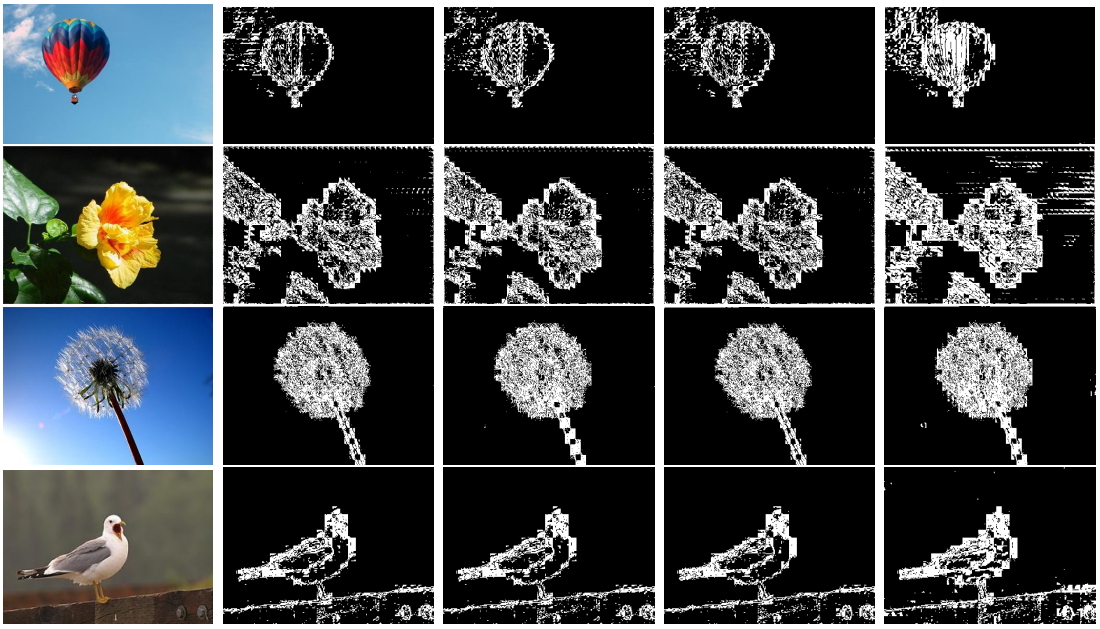

FIGURE 4.10: Saliency map identification using variant PCA methods. From the left: example images, results of PCA, A1, A2 and A3.

4.4.6 Discussion

In this chapter, four algorithms for RPCA are proposed. The first algorithm (A1) is based on robust covariance estimation and it is shown that the weights of each data point are an exponentially decreasing function of whitened residuals where the whitening is done by the covariance matrix. Therefore, a necessary condition to be able to apply this method is $N \geq m$. Thus, for the case of $N < m$, A2 and A3 are preferable algorithms to implement, unless we do some tricks, such as projection, to enable use of A1. For the case of $N >> m$, A1 and specially A2 will be better choices due to use of a simple estimation step and estimation of a limited number of weights. In this scenario, applying A3 is computationally expensive because of the number of weights which need to be computed. On the other hand A3 is robust even for a single pixel contamination level while A1 and A2 just remove outliers at the vector level. Generally, A2 and A3 do not have any constraint on loading vectors and have more degrees of freedom to find the true subspace. Finally, A4 has a sparsity constraint, makes it less generally applicable, but in the applications where sparsity has a role, A4 is useful.

4.5 Conclusion

In this chapter, new robust principal component analysis (RPCA) algorithms were developed based on the $\alpha-$divergence criterion. This divergence is a general parametric family of divergences which includes other well known divergences as special cases, especially KL divergence from which the classical PCA can be derived. It is shown that the proposed algorithms provide robustness against both structured and unstructured outliers which may happen in many signal and image processing applications. Robustness is achieved by automatically assigning small weights to the outliers. These methods belong to the general family of robust estimators called M-estimators [10]. Using simulated datasets, it was shown how the proposed methods can be used in fMRI data analysis to estimate the signal subspace before further steps are performed for activation detection. Meanwhile, the results of the proposed RPCA methods in the synthetic foreground background (FB) separation task are presented and compared. Finally, some real world applications of the proposed methods on FB separation and saliency map identification were discussed and the superiority of the algorithms over the existing works is demonstrated.

Chapter 5

Adaptive Matched Filter using Non-Target Free Training Data

The problem of detecting a subspace signal in colored Gaussian noise with unknown covariance matrix is investigated when the training data may contain samples with target signal. The target signal is assumed that it lies in a subspace spanned by columns of a known matrix. To develop the test, an ad hoc approach, similar to the classical adaptive matched filter (AMF) is used where instead of the maximum likelihood (ML) estimator of the covariance, the minimum $\alpha-$divergence based estimator is substituted in the likelihood ratio. This test just depends on the single parameter α and as a special case can be turned to the AMF. For a range of α, the proposed test has the benefits of being robust to outliers and the existence of other targets in the training data. Numerical examples illustrating that the proposed detector can achieve better detection rates in such a scenario while providing almost the same performance in a target free scenario are presented.

5.1 Introduction

Signal or subspace detection in colored Gaussian noise is of interest to researchers in radar/sonar signal processing [47] and medical imaging data analysis [16, 17, 166]. By the term of noise, we mean a broad-sense noise including clutters and interferences. For the white noise case, a general study which is named as matched subspace detectors

(MSD) was proposed in [49] considering the general linear model (GLM). In many applications, this scenario is not appropriate and a method should be developed for colored noise scenarios. Therefore, in the last decades, some works were done to cover all possible scenarios[6, 50–52, 57, 59, 167]. To deal with estimation of unknown covariance, typically it was assumed that a set of training data which is target free is available [50]. For example in the radar community, this dataset is usually collected from adjacent cells (AC) to the cell under the test (CUT).

If the training and test data share the same covariance matrix, this situation is referred as the homogeneous case and the well known generalized likelihood ratio (GLR) test [50] and adaptive matched filter [52] are based on it. Both of these detectors are constant false alarm (CFAR), but none of them is uniformly most powerful (UMP)[95]. In other cases, it was assumed that the environment is partially homogeneous. It means that the covariance matrix of the noise has the same structure in the test and training data but may differ by a scaler factor. The adaptive subspace detectors (ASD) and its variants are one of the famous detectors in this group [6, 51].

Recently, most of the attention of the works was on increasing the robustness of the detectors against possible mismatches in target signal [57, 59, 167] or finding a way to reduce the number of training data by exploiting some structures for the covariance matrix [6, 61, 97]. Even though the number of required target free samples is reduced when some structures for the covariance matrix are assumed, but still a significant number of data points is needed in radar application which is not easy to collect. One of the main reason for this limitation is the non-stationarity of the covariance of the noise in far range cells.

One of the strong assumptions on this training data is being target free. In the radar community, it means we assume in the entire ranges that we collect our data, there is only one target, that might not be always correct by having different targets in different ranges. Figure 5.1 (top) visually shows this scenario where other targets can act as clutters in AC. This situation may also happen in functional magnetic resonance imaging (fMRI) data analysis due to spatial smooth functionality of the voxels (figure 5.1, bottom). Therefore, when we are analyzing a voxel, most probably the adjacent voxels are active. Although recent works are trying a lot to improve the performance of the detectors against different uncertainties, surprisingly, to the best of our knowledge, there is no work with this concern and sample covariance was used in the tests as an accurate estimation.

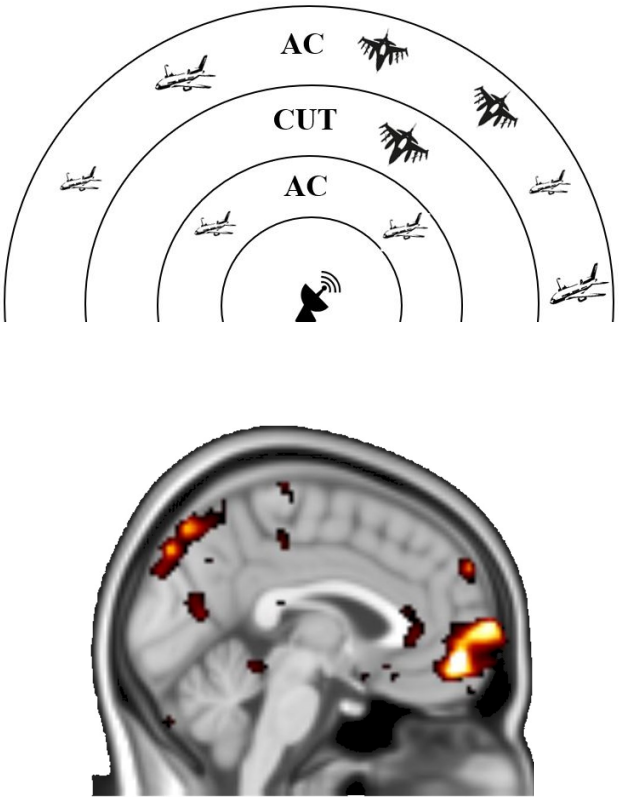

FIGURE 5.1: Two examples that show the assumption of target free observations might not be correct. 1:In radar example, the adjacent cells (AC) also include some targets which may mislead detectors in a way that it is thought that the target in the CUT is nothing more than clutters (top). 2: In fMRI data analysis which we cannot be sure that which areas of the brain, close to the voxel under the test, are target free (bottom). The red areas are the active areas during a task.

In this chapter, we consider the case that the training data may include some samples where there is a target signal inside and take them into account as outliers. Using a robust estimation approach we reduce the effect of those samples to increase the power of the detector. The $\alpha-$divergence is used to propose our new covariance estimator which belongs to the general family of M-estimators.

The rest of the chapter is organized as follows. In section 5.2 we present the background. In section 5.3 we propose our methodology. In section 5.4 we present comparisons between the classical AMF and our proposed method. Finally, in section 5.5 we present some concluding remarks.

5.2 Background

Let's assume the test data follows a general linear model

$$\mathbf{y} = \mathbf{H}\boldsymbol{\theta} + \boldsymbol{\xi}, \tag{5.1}$$

where $\mathbf{H} \in \mathbb{C}^{N\times p}$ is the known target subspace matrix. $\boldsymbol{\theta} \in \mathbb{C}^p$ is a deterministic but unknown coefficients for the columns of $\mathbf{H}$. Finally $\boldsymbol{\xi} \in \mathbb{C}^N$ denotes noise which follows a circularly complex Gaussian distribution with unknown covariance. Beside the test data, another K training data denoted by $\{\mathbf{n}_k\}_{k=1}^K$ is available, where nominally $\mathbf{n}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{R})$. The subspace detection problem now can be formulated as the following hypothesis testing problem:

$$\mathcal{H}_0 = \begin{cases} \mathbf{y} = \boldsymbol{\xi} \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}) \\ \mathbf{y}_k = \mathbf{n}_k \sim (1-\epsilon)\mathcal{CN}(\mathbf{0}, \mathbf{R}) + \epsilon\mathcal{CN}(\mathbf{H}\boldsymbol{\beta}, \mathbf{R}) \end{cases} \tag{5.2}$$

and

$$\mathcal{H}_1 = \begin{cases} \mathbf{y} = \mathbf{H}\boldsymbol{\theta} + \boldsymbol{\xi} \sim \mathcal{CN}(\mathbf{H}\boldsymbol{\theta}, \mathbf{R}) \\ \mathbf{y}_k = \mathbf{n}_k \sim (1-\epsilon)\mathcal{CN}(\mathbf{0}, \mathbf{R}) + \epsilon\mathcal{CN}(\mathbf{H}\boldsymbol{\beta}, \mathbf{R}) \end{cases} \tag{5.3}$$

where $\mathcal{H}_0$ and $\mathcal{H}_1$ are the null and alternative hypotheses, respectively and $\mathbf{H}\boldsymbol{\beta}$ models the existence of target signal in training data which may differ from $\mathbf{H}\boldsymbol{\theta}$. For this problem, considering $\epsilon = 0$, some classical methods such as AMF and GLRT were proposed and have the following structures [168]:

$$T_{AMF} = \mathbf{y}^\dagger\mathbf{S}^{-1}\mathbf{H}(\mathbf{H}^\dagger\mathbf{S}^{-1}\mathbf{H})^{-1}\mathbf{H}^\dagger\mathbf{S}^{-1}\mathbf{y}, \tag{5.4}$$

and

$$T_{GLRT} = \frac{\mathbf{y}^\dagger\mathbf{S}^{-1}\mathbf{H}(\mathbf{H}^\dagger\mathbf{S}^{-1}\mathbf{H})^{-1}\mathbf{H}^\dagger\mathbf{S}^{-1}\mathbf{y}}{1+\mathbf{y}^\dagger\mathbf{S}^{-1}\mathbf{y}}, \tag{5.5}$$

where $\mathbf{S} = \frac{1}{K}\sum_{k=1}^K \mathbf{y}_k\mathbf{y}_k^\dagger$ is the sample covariance matrix and $\dagger$ is the conjugate transpose operator. In our problem definition, $\epsilon \geq 0$ and is not necessarily equal to zero. A naive approach to deal with this contamination can be the projection on orthogonal subspace $\langle\mathbf{H}\rangle^\perp$ where the projected samples do not have any target signals anymore, but this projection makes the covariance low rank which is not our interest. Therefore, in the next section, we extend the AMF to achieve robustness to this deviation. Note that the

defined problem is a homogeneous problem where the covariance of the training and test dataset is the same.

5.3 Proposed Method

In the above described problem, the only difference with classical assumption is the distribution of the observations in training data. In this chapter, training data follows a contaminated Gaussian distribution which affects the performance of the maximum likelihood (ML) approach where in this problem is the sample covariance $\mathbf{S}$. It is well known that maximizing the likelihood function approach is equivalent to minimizing the KL divergence of two density functions. KL divergence is so sensitive to possible outliers. To deal with this sensitivity, a general parametric measure from information geometry is used where under some setups, it is robust against such outliers [169].
For two probability density functions $g(\mathbf{x}, \boldsymbol{\lambda})$ and $f(\mathbf{x}, \boldsymbol{\beta})$ parametrized by $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$, the α-divergence is defined as [102, 112, 113]

$$D_\alpha \left(g(\mathbf{x}, \boldsymbol{\lambda}) \parallel f(\mathbf{x}, \boldsymbol{\beta}) \right) = \frac{1}{\alpha(\alpha - 1)} \left[\int g(\mathbf{x}, \boldsymbol{\lambda})^\alpha f(\mathbf{x}, \boldsymbol{\beta})^{1-\alpha} d\mathbf{x} - 1 \right], \tag{5.6}$$

where α is a tuning parameter that defines the shape of the divergence. As the specific case when $\alpha \to 1$, the α-divergence turns to the KL-divergence. With changing the KL-divergence criteria to the $\alpha-$divergence, we end up with another estimator which is different from the sample covariance matrix and is robust for $\alpha < 1$. To be able to minimize this new measure, we need to have two density functions and then try to minimize the divergence with respect to unknown parameters. To represent the true distribution of the data, we use the empirical distribution which is a non-parametric approach to approximate the data distribution. The empirical density function can be defined by

$$g_e(\mathbf{x}; \mathbf{y}_1 \cdots \mathbf{y}_K) = \frac{1}{K} \sum_{k=1}^{K} \delta(\mathbf{x} - \mathbf{y}_k), \tag{5.7}$$

where $\mathbf{x}$ is the random variable and δ is the Dirac delta function. Now it is the time to use our parametric model which is the circularly complex Gaussian distribution in the divergence expression to derive the new robust estimators.

$$f(\mathbf{y}) = \frac{1}{\pi^N \det(\mathbf{R})} \exp \left\{ - \mathbf{y}^\dagger \mathbf{R}^{-1} \mathbf{y} \right\}, \tag{5.8}$$

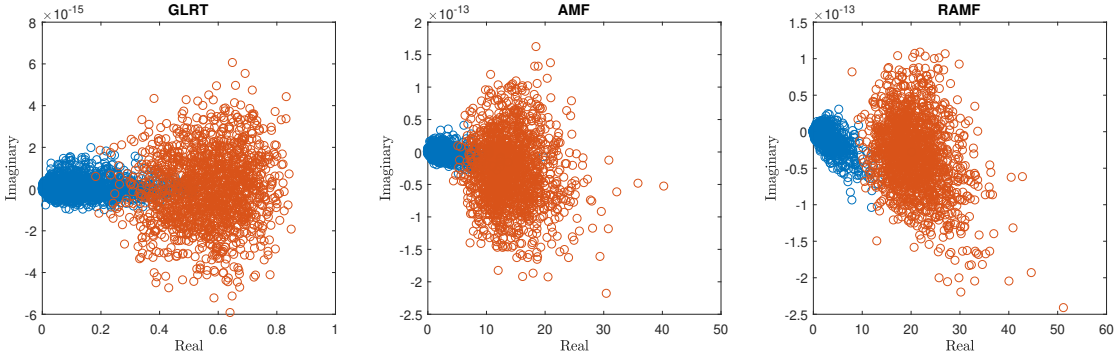

FIGURE 5.2: Scatter plot of GLRT, AMF and RAMF for null hypothesis (blue) and alternative hypothesis (red) . Ignoring the imaginary part and projection on real axis, shows the better classification of data.

Now with substitution of above densities in diverence expression we have

$$D_\alpha = \frac{1}{\alpha(1-\alpha)}\left[1 - \sum_{k=1}^{K}(1/K)^\alpha \left(\frac{1}{\pi^N \det(\mathbf{R})}\right)^{1-\alpha} \exp\left\{-(1-\alpha)\mathbf{y}_k^\dagger \mathbf{R}^{-1}\mathbf{y}_k\right\}\right]. \tag{5.9}$$

As is shown, now beside α, D_α is a measure that only depends on $\mathbf{R}$. Taking the derivative with respect to the covariance matrix and equalizing it to zero we have

$$\frac{\partial D_\alpha}{\partial \mathbf{R}} = c\sum_{k=1}^{K} w_k(-\mathbf{R} + \mathbf{y}_k\mathbf{y}_k^\dagger) = \mathbf{0}, \tag{5.10}$$

and finally

$$\hat{\mathbf{R}} = \frac{1}{\sum_{k=1}^{K} w_k}\sum_{k=1}^{K} w_k \mathbf{y}_k\mathbf{y}_k^\dagger, \tag{5.11}$$

where $w_k = \exp\{-(1-\alpha)\mathbf{y}_k^\dagger\mathbf{R}^{-1}\mathbf{y}_k\}$ and c is a positive constant value. This estimator is a weighted form of sample covariance where the weights are bounded and are a function of α and the norm of whitened observations. The weights themselves depend on the covariance estimator. Thus, this estimator is a recursive estimator which belongs to the general family of robust estimation approach namely M-estimators and converges fast to the minimizer of the $\alpha-$divergence criteria. Surprisingly, the proposed covariance estimator can be changed to the sample covariance when $\alpha \to 1$ and the weights all are equal to one.

Using the same *ad hoc* two-step approach of AMF, we can substitute the robust estimated covariance matrix $\hat{\mathbf{R}}$, into the following test:

$$T_{GLRT} = \frac{\max_{\boldsymbol{\theta}} f_1(\mathbf{y}|\mathbf{R})}{f_0(\mathbf{y}|\mathbf{R})}, \tag{5.12}$$

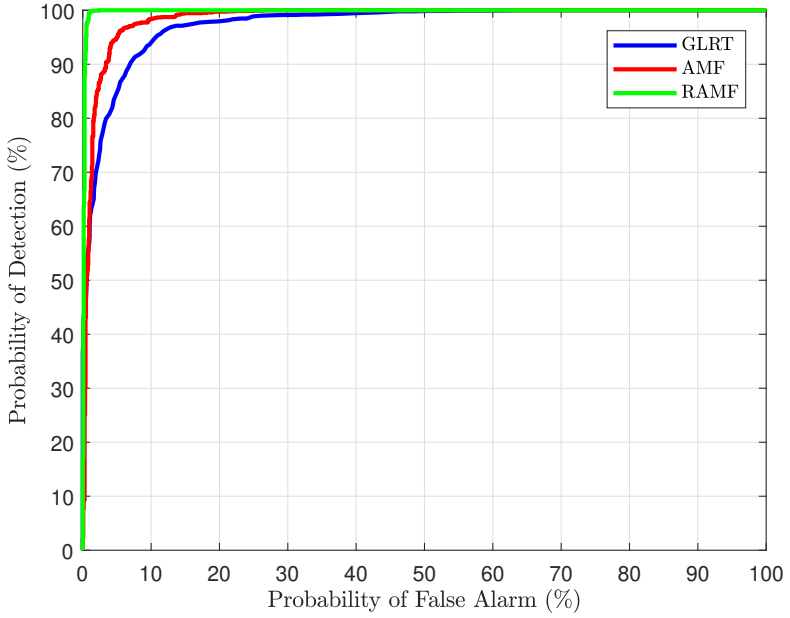

FIGURE 5.3: Performance comparison of the proposed method (RAMF) with AMF and GLRT. It can be observed that using the robust covariance estimator, the proposed test could distinguish better between two hypotheses and as the consequence, better probability of detection achieved for a fixed false alarm rate.

which ends up with the robust adaptive matched filter (RAMF):

$$T_{RAMF} = \mathbf{y}^{\dagger}\hat{\mathbf{R}}^{-1}\mathbf{H}(\mathbf{H}^{\dagger}\hat{\mathbf{R}}^{-1}\mathbf{H})^{-1}\mathbf{H}^{\dagger}\hat{\mathbf{R}}^{-1}\mathbf{y}. \tag{5.13}$$

5.4 Simulation Results

In this section, numerical simulations are conducted to illustrate the robustness performance of the proposed test. $\mathbf{R}$ is an exponentially correlated covariance matrix with one-lag correlation coefficient 0.9, i.e., the $(i,j)-$th element of $\mathbf{R}$ is set to $0.9^{|i-j|}$. During the simulations, signal to noise ratio (SNR) is defined as

$$\text{SNR} = 10\log_{10}(\boldsymbol{\theta}^{\dagger}\mathbf{H}^{\dagger}\mathbf{R}^{-1}\mathbf{H}\boldsymbol{\theta}). \tag{5.14}$$

The subspace matrix $\mathbf{H}$ is chosen to be

$$\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_p], \tag{5.15}$$

where

$$\mathbf{h}_i = 1/\sqrt{N}[1, e^{-j2\pi f_i}, \cdots, e^{-j2\pi f_i(N-1)}], \tag{5.16}$$

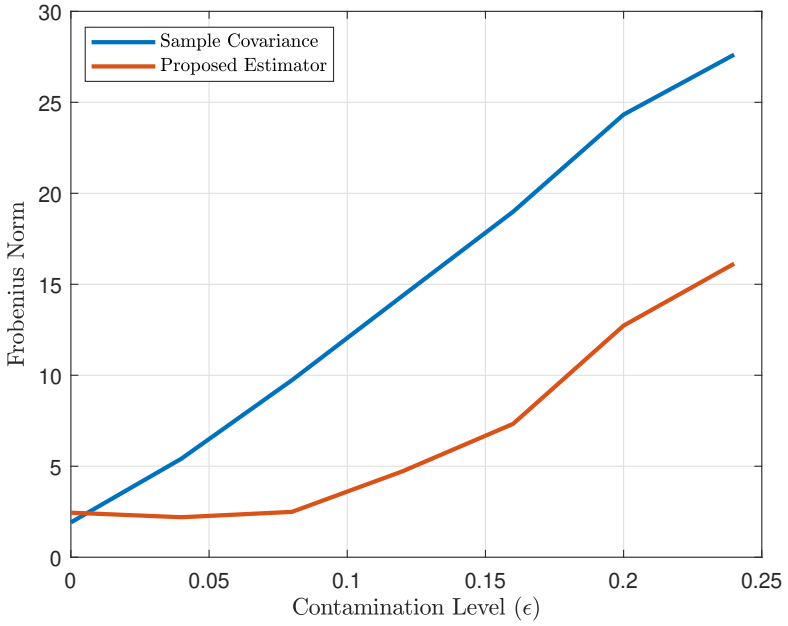

FIGURE 5.4: Frobenius norm of error between the real covariance and estimated one using the ML and $\alpha-$divergence. For the target free case $\epsilon = 0$, the sample covariance is slightly better. Adding target to the training data increases the error of the sample covariance faster than the robust estimator.

and $f_i = i \times 0.05 + 0.05$, $K = 40$, $p = 2$ and $N = 10$. $\boldsymbol{\theta}$ has the same entries and is chosen in a way to have the desired SNR value.

To show the performance results, 3000 observation vectors are generated using Monte Carlo as the test data. The target signal is added with the probability of $1/2$ to the test data with SNR= 25dB. Scaled target signal, with the same SNR of test data randomly is added to train data with the probability of $\epsilon = 0.15$ to simulate the situation that there might be a target signal in training data.

First, the GLRT, AMF and RAMF are applied to all generated test data and the results are presented in figure 5.2. As is shown, GLRT has the worst performance. The proposed method has the best separation of two hypotheses from each other due to the better estimation of the covariance matrix. The thresholding on the real part of the tests ends up with figure 5.3 that shows the probability of detection versus the probability of the false alarm. It is shown that the performance of the proposed test is significantly better when the training data is not target free.

To understand the benefit of the proposed test, the average Frobenius norm of the error between the real covariance matrix and the proposed covariance estimator over 80 trials is presented (figure 5.4) in comparison with the sample covariance matrix. It can be seen that when there is no contamination, the sample covariance is slightly better than the

proposed approach. On the other hand, increasing contaminations increases the error of sample covariance but the proposed method has much better estimation error for higher contamination levels.

There are so many approaches that can be used to tune the value of α. One of the important methods is using some prior knowledge about the maximum level of contamination and then assigning a small weight to the ϵ percent of data which are far away from the assumed model.

5.5 Conclusion

In this chapter, a classical assumption of target free training data is revisited. It is discussed that for some applications, it might not always be valid. A classical general linear model (GLM) is assumed for multi-rank signal case and the environment is assumed homogeneous which means both training and test data have the same covariance matrix. The link between the classical adaptive matched filter and the KL divergence is discussed and an alternative divergence ($\alpha-$divergence) is used to derive a robust version of the covariance estimator which can reduce the effect of target signals in the training data. Finally, the proposed test is applied to such scenario and is compared with the generalized likelihood ratio test (GLRT) and adaptive matched filter (AMF). The simulation results reveal the superiority of the proposed test over these two well known tests.

Chapter 6

Adaptive Subspace Detector in High Dimensional Space with Insufficient Training Data

Adaptive subspace detectors (ASD) generalize matched subspace detectors (MSD) by accounting for possible correlation. Both ASD and MSD are derived using the generalized likelihood ratio test (GLRT). While MSD assumes there is no correlation between observations, ASD estimates a sample covariance matrix of possibly correlated samples using signal-free observations. In this chapter, we address the performance of the ASD when the number of secondary data is insufficient and the observed signal lies in higher dimensional space. Such high dimensional spaces are frequently encountered in functional magnetic resonance imaging (fMRI) data for the analysis of brain activation detection. We propose a methodology that works based on the latent variables in a lower dimensional space. A low-rank decomposition of the sample covariance matrix is derived based on the singular value decomposition (SVD) and an adaptive basis selection method is used to decide which eigen-vectors are useful in data projection. Performing detection in the lower dimensional subspace has the benefit of reducing the number of parameters which need to be estimated. Simulation results show superiority of our proposed adaptive reduced subspace detector (ARSD) over conventional ASD in term of probability of detection.

6.1 Introduction

Signal detection in a noisy environment is an important problem that arises in many applications such as radar [11], communication [92] and medical imaging [16, 17]. Neyman and Pearson [38] have shown that when parameters of the noise distribution are perfectly known, the likelihood ratio test is the uniformly most powerful (UMP) test which maximizes the probability of detection for a given false alarm rate. However, in many applications, the likelihood ratio test is unknown and therefore a UMP test cannot be derived. Therefore, several studies have investigated alternatives for the likelihood ratio test [170]. In [49] the matched subspace detector (MSD) was proposed based on the generalized likelihood ratio test (GLRT), for which all unknown parameters are replaced by their maximum likelihood values. In their study, [49] considers the noise samples to be independent and identically distributed (i.i.d) with a white Gaussian probability density. In another study, Kelly proposed a method that considers a covariance matrix for the noise distribution [50]. This covariance matrix is estimated from K signal-free samples. As an extension, [51] proposed a method that assumes both training data and test data have almost the same covariance structure, with only a scale factor difference. This method is also based on the likelihood ratio test and is named adaptive subspace detector (ASD). The ASD can be viewed as a generalized case of the MSD in which the data and subspaces are prewhitened using the sample covariance matrix. There are also alternative tests with different points of view, such as Rao test [58] and Wald test [55] or based on model selection criteria [171, 172] that have been shown to be more suitable than the GLR under certain conditions. However, this study solely focuses on improving the performance of the GLRT.

The rest of the chapter is organized as follows. In section 6.2 we present the background and problem statement. In section 6.3 we propose our methodology. In section 6.4 we present comparisons between conventional ASD and our proposed method. Finally, in section 6.5 we present some concluding remarks.

6.2 Background

The detection problem that is addressed in our study is described as follows. We are given d samples from a real and scalar time series $\mathbf{y}(i), i = 0, 1, ..., d-1$ that is represented

by the column vector $\mathbf{y}$. This vector of observations is generated by some components based on a general linear model (GLM):

$$\mathbf{y} = \mu\mathbf{H}\boldsymbol{\theta} + \boldsymbol{\xi}, \tag{6.1}$$

where $\mathbf{H} \in \mathbb{R}^{d\times p}$ is a known matrix whose columns span the signal subspace. It is also assumed that the columns of $\mathbf{H}$ are linearly independent. μ is a scalar factor that is non-zero when the signal is present. Entries of $\boldsymbol{\theta}$ contains unknown deterministic values. The noise components are $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2\mathbf{R})$, where σ^2 and $\mathbf{R}$ are unknown and should be estimated separately from training data.

The aim of detection is to decide between the null hypothesis $\mathcal{H}_0 : \mu = 0$ and the alternative hypothesis $\mathcal{H}_1 : \mu > 0$ for a given measurement vector $\mathbf{y}$. Given a set of K, i.i.d training noise vectors $\mathbf{N} = [\mathbf{n}_1, ..., \mathbf{n}_K]$, each distributed as $\mathcal{N}(0, \mathbf{R})$, test signal vector $\mathbf{y} \in \mathbb{R}^d$ is measured and $y \sim \mathcal{N}(\mu\mathbf{H}\theta, \sigma^2\mathbf{R})$. In [51] it was shown that when $\boldsymbol{\theta}$ and $\mathbf{R}$ are unknown, the following GLRT based test can be used to detect the signal presence.

$$l(\mathbf{y}) = \frac{\mathbf{y}^\top \mathbf{S}^{-\frac{1}{2}} \mathbf{P}_\mathbf{D} \mathbf{S}^{-\frac{1}{2}} \mathbf{y}}{\mathbf{y}^\top \mathbf{S}^{-1} \mathbf{y}}, \tag{6.2}$$

where $\mathbf{D}$ is $\mathbf{S}^{-\frac{1}{2}}\mathbf{H}$ and $\mathbf{P}_\mathbf{D} = \mathbf{D}[\mathbf{D}^\top\mathbf{D}]^{-1}\mathbf{D}^\top$ is the projection operator that projects the whitened observation vector $\mathbf{y}$ onto the subspace that is spanned by columns of $\mathbf{D}$. The matrix

$$\mathbf{S} = \frac{1}{K}\sum_{i=1}^{K} \mathbf{n}_i \mathbf{n}_i^\top, \tag{6.3}$$

is the sample covariance matrix, which is estimated from the K training data. This matrix $\mathbf{S}$ is actually an estimation of covariance matrix of noise $\mathbf{R}$, that maximizes likelihood function. This solution for signal detection is referred as ASD.

In general, for signal detection using (6.1) estimation of $p + d(d+1)/2$ unknown parameters is required where p is the number of entries of vector $\boldsymbol{\theta}$ and d is dimension of the symmetric covariance matrix $\mathbf{R}$. Therefore, the number of parameters to be estimated is of order $O(d^2)$. For a limited training data and a high dimensional measurement vector $\mathbf{y}$, (in some applications such as radar and medical imaging) applying ASD directly on data is inefficient and parameters are highly deviated from their actual values. Therefore, introducing knowledge about the covariance matrix is required to improve the performance of ASD.

6.3 Proposed Method

As mentioned earlier, in high dimensional space hypothesis testing, there are many unknown parameters (in the order of $O(d^2)$) that should be estimated. Several studies have addressed this issue by assuming a known structure for the covariance matrix [62]. One of the ways of introducing structure to the covariance matrix in order to improve performance of the detector (6.2), is to assume that the measurement vector $\mathbf{y}$ and noise vector $\boldsymbol{\xi}$ depend on latent variables in a lower dimensional space. In other words, when the number of training data is insufficient, the covariance matrix can be described by a few dominant eigen-vectors/values. Based on this assumption, there are two latent variables $\mathbf{z} \in \mathbb{R}^{d_1}$ and $\mathbf{e} \in \mathbb{R}^{d_1}$ with $d_1 \ll d$ and

$$\mathbf{y} = \mathbf{U}_{d_1}\mathbf{z} \quad \text{and} \quad \boldsymbol{\xi} = \mathbf{U}_{d_1}\mathbf{e}, \tag{6.4}$$

where $\mathbf{U}_{d_1}$ is an unknown $d \times d_1$ matrix whose columns are orthogonal, (i.e. $\mathbf{U}_{d_1}^\top \mathbf{U}_{d_1} = \mathbf{I}_{d_1}$, where $\mathbf{I}_{d_1}$ is the identity matrix of size $d_1 \times d_1$). It is also assumed that $\{\mathbf{e}\}$ is a zero mean Gaussian noise vector with a $d_1 \times d_1$ covariance matrix $\boldsymbol{\Lambda}$. Substituting (6.4) in (6.1) we have

$$\mathbf{U}_{d_1}\mathbf{z} = \mu\mathbf{H}\boldsymbol{\theta} + \mathbf{U}_{d_1}\mathbf{e}. \tag{6.5}$$

Multiplying both sides by $\mathbf{U}_{d1}^\top$ gives

$$\mathbf{z} = \mu\mathbf{U}_{d_1}^\top\mathbf{H}\boldsymbol{\theta} + \mathbf{e}. \tag{6.6}$$

In lower dimensional space, the formulation of (6.6) is the same as (6.1) and we can apply standard detection theory results on the latent variable $\mathbf{z}$ with the difference that the signal subspace in the lower dimensional space is the span of the columns of $\mathbf{U}_{d_1}^\top\mathbf{H}$, instead of $\mathbf{H}$. Therefore, the number of unknown parameters is reduced to $p+d_1(d_1+1)/2$ which is significantly less than $p + d(d+1)/2$, corresponding to the higher dimensional space. In general, using an optimum projection for the lower dimensional space, we may lose some useful information, however, the advantage is that we only need to estimate a small number of parameters. Reducing the number of parameters reduces the parameter estimation error too.

The ASD for the latent variable $\mathbf{z}$ is:

$$l(\mathbf{z}) = \frac{\mathbf{z}^\top \hat{\mathbf{\Lambda}}^{-\frac{1}{2}} \mathbf{P}_\mathbf{B} \hat{\mathbf{\Lambda}}^{-\frac{1}{2}} \mathbf{z}}{\mathbf{z}^\top \hat{\mathbf{\Lambda}}^{-1} \mathbf{z}}, \tag{6.7}$$

where $\mathbf{B}$ is $\hat{\mathbf{\Lambda}}^{-\frac{1}{2}} \mathbf{U}_{d_1}^\top \mathbf{H}$ and

$$\hat{\mathbf{\Lambda}} = \frac{1}{K} \sum_{i=1}^{K} \mathbf{U}_{d_1}^\top \mathbf{n}_i \mathbf{n}_i^\top \mathbf{U}_{d_1} = \mathbf{U}_{d_1}^\top \mathbf{S} \mathbf{U}_{d_1}. \tag{6.8}$$

Finally, using threshold value τ, we can decide between $\mathcal{H}_0$ and $\mathcal{H}_1$. We name our reduced version of ASD, the adaptive reduced subspace detector (ARSD).

6.3.1 Estimation of $\mathbf{U}_{d_1}$

The primary challenge of ARSD is finding the best projection matrix $\mathbf{U}_{d_1}$. When the number of training data is insufficient, the sample covariance matrix is not consistent and the eigen-values and eigen-vectors of the sample covariance matrix can be significantly different from their true values [173].

In order to find a suitable projection basis set, we need to investigate equation (6.7) in detail. After reparameterizing of the reduced version of the test in terms of $\mathbf{y}$, we will have

$$l'(\mathbf{y}) = \frac{\mathbf{y}^\top \mathbf{U}_{d_1} \hat{\mathbf{\Lambda}}^{-1} \mathbf{U}_{d_1}^\top \mathbf{H} (\mathbf{H}^\top \mathbf{U}_{d_1} \hat{\mathbf{\Lambda}}^{-1} \mathbf{U}_{d_1}^\top \mathbf{H})^{-1} \mathbf{H}^\top \mathbf{U}_{d_1} \hat{\mathbf{\Lambda}}^{-1} \mathbf{U}_{d_1}^\top \mathbf{y}}{\mathbf{y}^\top \mathbf{U}_{d_1} \hat{\mathbf{\Lambda}}^{-1} \mathbf{U}_{d_1}^\top \mathbf{y}}. \tag{6.9}$$

In comparison with the test shown in equation (6.2), the only difference is the estimated covariance matrix which is a modified form of the sample covariance matrix, that is:

$$\hat{\mathbf{R}}_{Mod}^{-1} = \mathbf{U}_{d_1} \hat{\Lambda}^{-1} \mathbf{U}_{d_1}^\top = \mathbf{U}_{d_1} (\mathbf{U}_{d_1}^\top \mathbf{S} \mathbf{U}_{d_1})^{-1} \mathbf{U}_{d_1}^\top. \tag{6.10}$$

We would like to find a basis set $\mathbf{U}_{d_1}$ that improves the overall performance of the detector in (6.7). For finding the best projection matrix, we consider two different cases of covariance structures on which our method can be applied.

Algorithm 5 Proposed ARSD

Input: Observation vector $\mathbf{y}$, Training data $\mathbf{N}$, $\mathbf{H}$, d_1,α, τ
Output: Decision on $\mathcal{H}_0$ or $\mathcal{H}_1$

procedure ARSD
  $\mathbf{S} \leftarrow \frac{1}{K}\sum_{i=1}^{K}\mathbf{n}_i\mathbf{n}_i^\top$
  $\mathbf{U}, \Gamma, \mathbf{U}^\top \leftarrow \text{SVD}(\mathbf{S})$
  $\mathbf{U}_{d_1} \leftarrow$ solve equation (6.11) or (6.16)
  $\mathbf{z} \leftarrow \mathbf{U}_{d_1}^\top \mathbf{y}$
  $\hat{\mathbf{\Lambda}} \leftarrow \frac{1}{K}\sum_{i=1}^{K}\mathbf{U}_{d_1}^\top\mathbf{n}_i\mathbf{n}_i^\top\mathbf{U}_{d_1}$
  $\mathbf{B} \leftarrow \hat{\mathbf{\Lambda}}^{-\frac{1}{2}}\mathbf{U}_{d_1}^\top\mathbf{H}$
  $l(\mathbf{z}) \leftarrow \frac{\mathbf{z}^\top\hat{\mathbf{\Lambda}}^{-\frac{1}{2}}\mathbf{P}_B\hat{\mathbf{\Lambda}}^{-\frac{1}{2}}\mathbf{z}}{\mathbf{z}^\top\hat{\mathbf{\Lambda}}^{-1}\mathbf{z}}$
  Output $\leftarrow \mathcal{H}_0$
  if $l(\mathbf{z}) \geq \tau$ **then**
    Output $\leftarrow \mathcal{H}_1$

6.3.1.1 R with Uniform Diagonal Components

When the covariance matrix has the same diagonal components, it means that the variance of the noise is fixed for all components of noise vector $\mathbf{n}$, which can happen in many applications. In this case study, we need to solve the following optimization problem:

$$\hat{\mathbf{U}}_{d_1} = \arg\min_{\mathbf{U}_{d_1}} \|\mathbf{U}_{d_1}\mathbf{U}_{d_1}^\top\mathbf{S}\mathbf{U}_{d_1}\mathbf{U}_{d_1}^\top - \mathbf{S}\|_F^2, \tag{6.11}$$

where $\|.\|_F^2$ shows Frobenius norm of a matrix. This optimization problem means that we need to regularize the sample covariance in a way that the result gets close enough to the sample covariance. For a known d_1 the solution is the first d_1 eigen-vectors of $\mathbf{S}$.

6.3.1.2 Ill-Conditioned R

In the previous section, the solution was the first eigen-vectors of the sample covariance. It means that the last eigen-vectors were in the direction of the noise which is caused by insufficiency in the training data. In this new case, when the covariance matrix is in ill condition, we cannot have the same strategy. In this case, $\mathbf{R}^{-1}$ is significantly dependent on the last eigen-vectors and we should be careful about picking eigen-vectors. A standard approach is to find the maximum likelihood (ML) estimate for $\mathbf{U}_{d_1}$. To

compute the likelihood for the given projected training data we have:

$$f[\mathbf{U}_{d_1}^\top \mathbf{n}_1, ..., \mathbf{U}_{d_1}^\top \mathbf{n}_K] = \big[\frac{1}{\pi^{d1}\det(\mathbf{U}_{d_1}^\top \mathbf{R}\mathbf{U}_{d_1})} \exp\{-\mathrm{tr}((\mathbf{U}_{d_1}^\top \mathbf{R}\mathbf{U}_{d_1})^{-1}\mathbf{U}_{d_1}^\top \mathbf{S}\mathbf{U}_{d_1})\big]^K. \tag{6.12}$$

This likelihood function has two variables $\mathbf{R}$ and $\mathbf{U}_{d_1}$ that need to be estimated via ML. To do this, we follow a two-stage estimation procedure. We first assume $\mathbf{U}_{d_1}$ is fixed and estimate $\mathbf{R}$. In the second stage, we keep $\mathbf{R}$ fixed and estimate $\mathbf{U}_{d_1}$. Assuming $\mathbf{U}_{d_1}$ is fixed, we have:

$$\mathbf{U}_{d_1}^\top \mathbf{R}\mathbf{U}_{d_1} = \mathbf{U}_{d_1}^\top \mathbf{S}\mathbf{U}_{d_1}. \tag{6.13}$$

The optimal $\mathbf{R}$ must satisfy (6.13). Therefore, we now replace $\mathbf{U}_{d_1}^\top \mathbf{R}\mathbf{U}_{d_1}$ in (6.12) with $\mathbf{U}_{d_1}^\top \mathbf{S}\mathbf{U}_{d_1}$ and after simplification

$$f[\mathbf{U}_{d_1}^\top \mathbf{n}_1, ..., \mathbf{U}_{d_1}^\top \mathbf{n}_K] \propto \frac{1}{\det(\mathbf{U}_{d_1}^\top \mathbf{S}\mathbf{U}_{d_1})}. \tag{6.14}$$

Now, maximizing the likelihood with respect to $\mathbf{U}_{d_1}$ can be changed to the following problem

$$\hat{\mathbf{U}}_{d_1} = \arg\min_{\mathbf{U}_{d_1}} \det(\mathbf{U}_{d_1}^\top \mathbf{S}\mathbf{U}_{d_1}), \tag{6.15}$$

In (6.15), if d_1 is assumed to be known, the solution is the last d_1 eigen-vectors which can be obtained by singular value decomposition of the sample covariance matrix $\mathbf{S}$.

So far, we have found that to maximize the likelihood function, we need to pick the last eigen-vectors. However, we know that although $\mathbf{S}$ is highly deviated from the actual covariance when the number of training samples is insufficient, but the sample covariance is a consistent estimation when data is enough. Therefore, our solution must also be as close as possible to the sample covariance matrix $\mathbf{S}$ in the Frobenius norm sense (similar to 6.11). As a result, instead of a conventional maximum likelihood solution for $\mathbf{U}_{d_1}$, we are faced with a constrained optimization problem which can be written as a multi objective optimization problem:

$$\hat{\mathbf{U}}_{d_1} = \arg\min_{\mathbf{U}_{d_1}} \det(\mathbf{U}_{d_1}^\top \mathbf{S}\mathbf{U}_{d_1}) + \alpha\|\mathbf{U}_{d_1}\mathbf{U}_{d_1}^\top \mathbf{S}\mathbf{U}_{d_1}\mathbf{U}_{d_1}^\top - \mathbf{S}\|_F^2, \tag{6.16}$$

which tries to maximize the likelihood such that the estimated covariance remains in the neighborhood of the sample covariance. The eigen-vectors that keep the covariance as close as possible to the sample covariance in terms of Frobenius norm, are the first

eigenvectors of $\mathbf{S}$. Meanwhile, the last eigen-vectors minimize the determinant, the first term of (6.16). Finally a combination of the first a and the last b eigen-vectors of $\mathbf{S}$ is the solution of problem (6.16) where $a + b = d_1$. Algorithm 5 summarizes the basis selection method for ARSD.

6.4 Simulation Results

In this section, we address the performance of ARSD in comparison with conventional ASD when the training data is insufficient. During simulation, we assume $d = 40$, and the number of training data K is 60, which means that K is not significantly larger than d. In lower dimensional space, we can change the dimension to the d_1 that makes the K/d_1 large. We randomly produce the signal subspace matrix $\mathbf{H}$ of size 40×2 (based on event related stimulus function of fMRI data analysis [29, 32]) and it is fixed for all simulations.

First, using equations (6.8), (6.11) and (6.16), parameters $\mathbf{\Lambda}$ and $\mathbf{U}_{d_1}$ are estimated in lower dimensional space. These parameters are then used in ARSD (6.7) and will assess the probability of detection. Figure 6.1 shows the average and standard deviation of Frobenius norm of error between the estimated and actual $\boldsymbol{\theta}$, $\mathbf{R}$ and $\mathbf{R}^{-1}$ where a random uniform covariance is used. It can be seen that when we bring the observation vector to a lower dimensional space for a range of d_1, we can have a better estimation of $\boldsymbol{\theta}$ which makes our test performance better.

Finally figure 6.2 illustrates the performance of the proposed detector in comparison with conventional ASD using a receiver operating characteristic (ROC) curve. In this figure, the performance of the proposed method is shown where we can see that ARSD (black) always outperforms ASD (green). The red ROC curve shows the oracle scenario where the covariance matrix is known. When we increase the value K, both of ASD and ARSD plots converge to the red plot.

6.5 Conclusion

In this study, a reduced version of the adaptive subspace detector (ASD) was proposed, namely adaptive reduced subspace detector (ARSD). It was shown for some scenarios, by bringing data to a lower dimensional space, the likelihood ratio test can provide a

better probability of detection. The most important section of the algorithm is providing proper orthonormal basis to project data. We have used a basis selection method that picks some eigen-vectors among the sorted eigenvectors of the sample covariance to minimize our defined objective function.

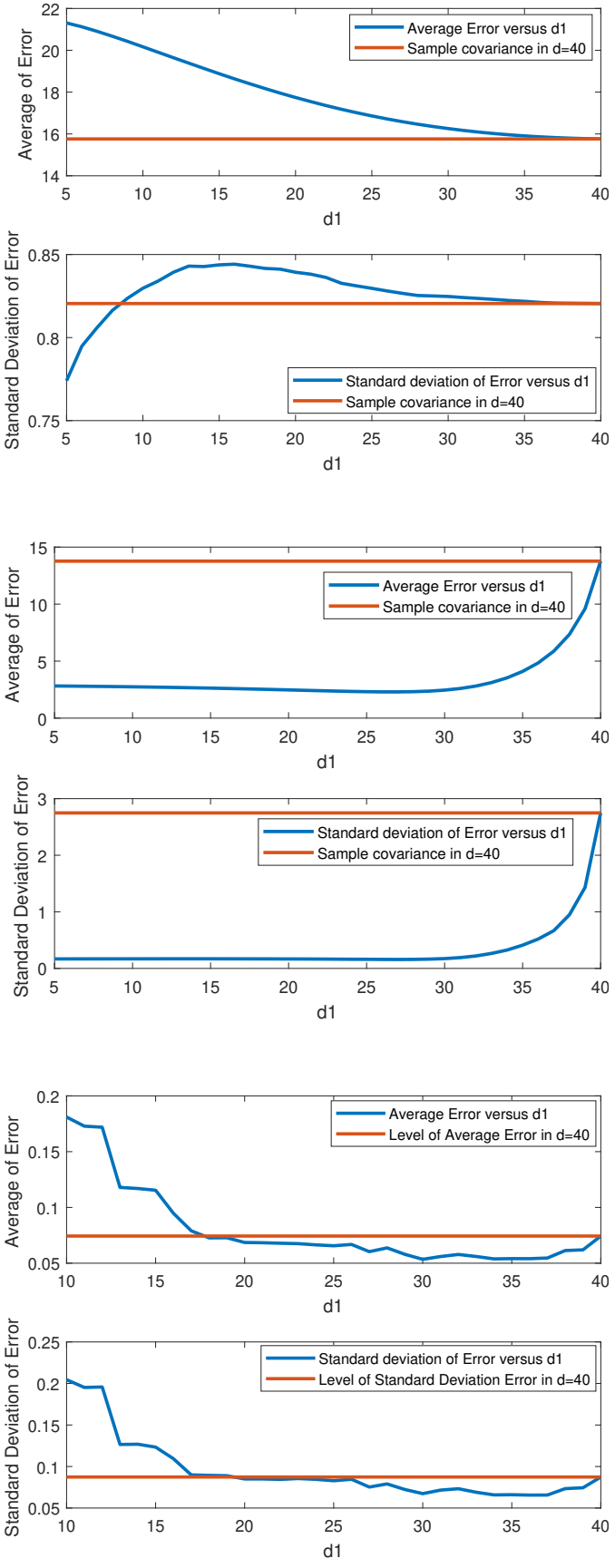

FIGURE 6.1: Average and standard deviation of Frobenius norm of difference between the actual and the estimated covariance (top), covariance inverse (middle) and $\boldsymbol{\theta}$ (bottom).

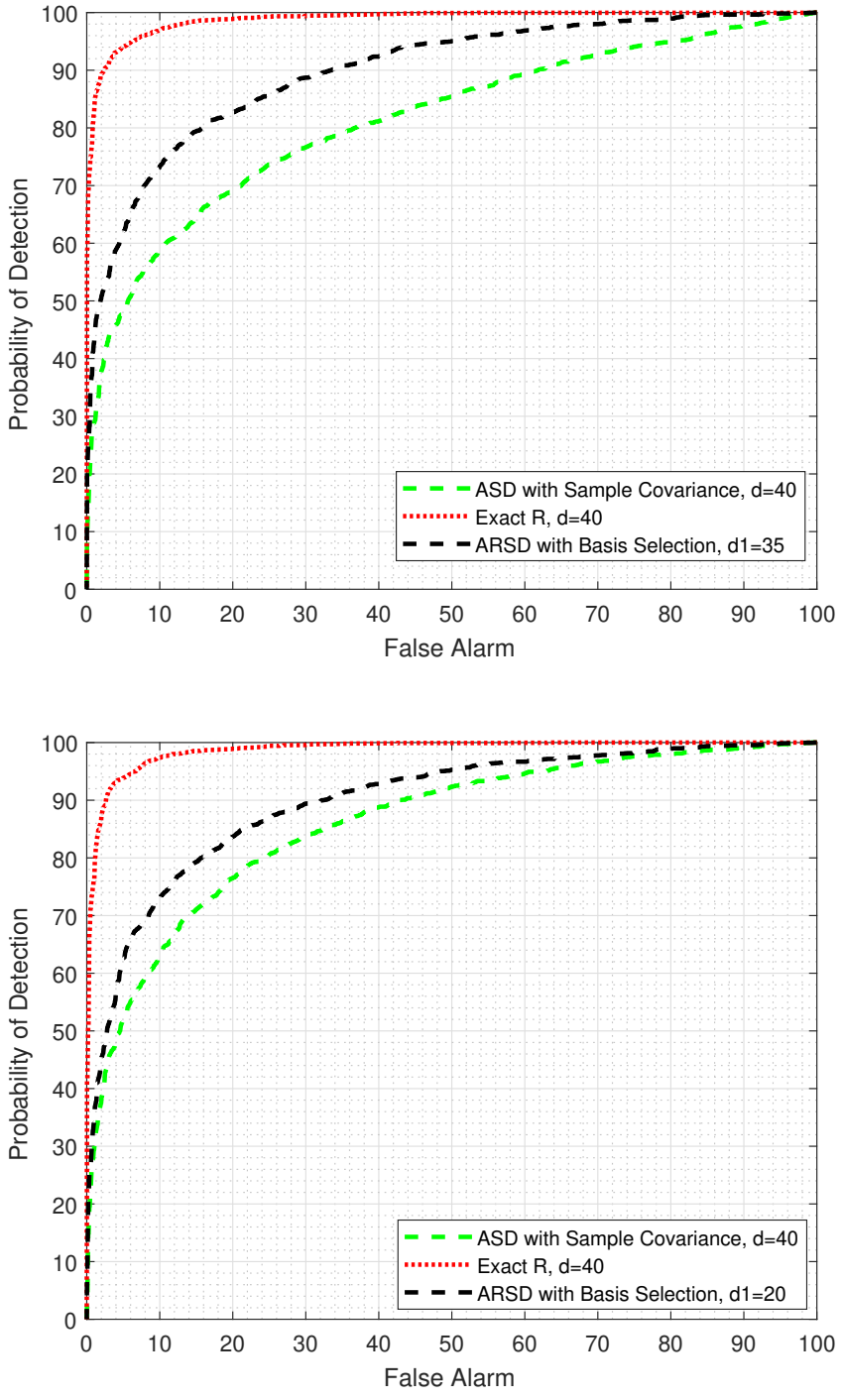

FIGURE 6.2: ROC of detectors in SNR=12 dB, $\mathbf{R}$ with uniform diagonal components (top), Ill-conditioned $\mathbf{R}$ (bottom).

Chapter 7

Adaptive Brain Activity Detection in Structured Interference and Partially Homogeneous Locally Correlated Disturbance

In this chapter, we address the problem of subspace detection in locally correlated complex Gaussian noise and subspace interference with unknown coordinates. For applications where the noise components are only correlated to their neighborhood, using sample covariance estimator as a maximum likelihood estimate (MLE) is not a suitable choice due to significant dependency of the estimation accuracy to the number of observations. In this study, we take advantage of an assumed banded structure in the covariance matrix. In particular, we use the idea of factorizing the joint likelihood function to a few conditional likelihood terms and then maximizing each term independent of others. This process leads toward an explicit estimator of banded covariance matrices which requires less observations to achieve the same accuracy and guaranteed banded structure. Finally, this estimate can be fed into an adaptive matched filter for detection purposes. Simulation results reveal the superiority of the proposed algorithm over well known classical detectors such as adaptive subspace detector (ASD) and adaptive matched filter (AMF)

which are based on the sample covariance estimate. The proposed method is applied to functional magnetic resonance imaging (fMRI) data to localize neural activities in the brain for the task of right finger tapping.

7.1 Introduction

Subspace detection in the presence of disturbance is an important topic which was widely used in various areas such as communications[92], radar [11], classification [93], medical imaging [16, 17] and pattern recognition [94]. Neyman and Pearson [38] showed that the likelihood ratio test (LRT) is the uniformly most powerful (UMP) test for a given false alarm rate when the distributions of the data under both hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ are known. Unfortunately, the UMP test does not exist in many signal processing applications due to the uncertainties and the existence of unknown parameters in assumed models [95]. Naturally, finding suboptimal detectors is always in the interests of people from different applications to improve the overall performance [56].

One of the famous alternatives to LRT is the generalized likelihood ratio test (GLRT) where the unknown parameters in each hypotheses are replaced with their maximum likelihood estimates (MLE). For those readers who are interested, there are also other alternatives such as Rao [58] and Wald [56] tests that have the same asymptotic properties and sometimes can offer a better finite sample performance, but their analysis is not in our interests for the current study.

Kelly [50] proposed a GLRT variant for detecting a target signal in homogeneous colored Gaussian disturbance when the covariance matrix is unknown. It was shown that his proposed detector has a constant false alarm rate (CFAR) with respect to the true covariance, which means for a non-stationary noise environment, the false alarm rate does not change. Due to the complexity of his detector, a modified version of GLRT was proposed where it was first assumed that the covariance matrix is known and then after constructing the GLRT by maximizing with respect to other unknowns, the covariance matrix is replaced with its sample covariance estimated from a secondary target free observations. This detector was called the adaptive matched filter (AMF) [52] and still guarantees the CFARness property and for some applications can perform better than GLRT. Adaptive coherence estimator (ACE) [174] generalizes the results for the partially homogeneous environment where the covariance matrix of the test data is different

from the training data up to a scale factor, and adaptive subspace detector (ASD) [51] extends the ACE for the subspace signals.

Note that in all of the above detectors, no prior knowledge on the structure of the noise covariance matrix is exploited and the covariance matrix can include all possible Hermitian structures. Usually, to have a reasonable accuracy, a training dataset with size at least twice the dimension of the observed data is required. Unfortunately, this requirement may not be satisfied in many applications for example in target detection of radar signals. Therefore, some prior knowledge on the structure of the covariance matrix can be used to achieve the same performance with a smaller number of training data.

In the literature, some well-known structures such as persymmetric [61, 97, 175], low rank [6, 62], block diagonal [176] and Toeplitz covariance matrices [177] have been investigated. In this chapter, we address the detection problem in complex Gaussian noise disturbance with banded covariance. Banded covariance matrices arises frequently in economical, biological or technical time series for example in signal processing applications such as autoregressive, moving average image models, image filtering, drift removal from a given time series [113] and specially in fMRI data analysis [178]. Assuming spatially independent time series of fMRI data, measurements are usually recorded every few seconds and naturally the noise source at time t cannot affect the observation at time $t+\delta$ where $\delta >> 1$. This physical fact brings this idea in mind to consider a banded structure for the covariance matrix of the noise and derive the detector based on this assumption.

In this chapter, as in [52], we use a two steps approach to derive the test. First, it is assumed the covariance matrix is known and use the GLRT approach to construct the test by maximizing the log likelihood function with respect to other unknown parameters. Second, the covariance matrix is replaced by an estimation of a banded covariance matrix. In this chapter, the target signal is presented as the unknown combination of a multi-rank known subspace matrix to improve the robustness of the detector to the mismatching. Moreover, we consider the presence of structured interference with unknown coordinates which may happen in real life due to electronic countermeasure. To make the problem more general, a partially homogeneous environment is considered where the covariance of the training and test data can be different up to a scalar. The main contributions of this chapter may be summarized as follows:

- A general form of subspace model is investigated where the target signal can be

any linear combination of independent vectors. This signal can also be corrupted by a partially homogeneous complex Gaussian noise with unknown covariance in addition to unwanted interference with known subspace.

- An explicit form of banded covariance estimator is used for detector design. This covariance is guaranteed to have a banded structure which is useful for analysis of locally correlated signals especially medical imaging data which are recorded over a long time period.

The remainder of this chapter is organized as follows. Section 7.2 briefly formulates the problem. Section 7.3 contains the background on the ASD and banded matrices. In section 7.4 the proposed method is introduced. The simulation results are presented in section 7.5 and Finally the chapter is concluded in section 7.6.

Notation: Vectors (matrices) are denoted by boldface lower (upper) case letters. For a given matrix $\mathbf{A}$, $a_{i,j}$, $\mathbf{a}^i$ and $\mathbf{a}_j$ show the entry on the $i-$th row and $j-$th column, the $i-$th row of the matrix and the $j-$th column of the matrix, respectively. $\mathbf{A}^{i:j}$ also shows a matrix including row i to j of matrix $\mathbf{A}$. Superscripts $(.)^\top$,$(.)^*$ and $(.)^\dagger$ denote the transpose, complex conjugate and complex conjugate transpose, respectively. The notation $\sim$ means "is distributed as," $\mathcal{CN}$ denotes circularly symmetric complex Gaussian distribution and $\mathcal{N}$ stands for real Gaussian distribution. $\mathbf{I}_N$ and $\mathbf{1}_N$ are the identity matrix and a vector of elements of 1 with the size of N. det(.), tr(.) and log(.) denote the determinant, the trace and the natural logarithm functions. $\mathbf{P_D} = \mathbf{D}[\mathbf{D}^\dagger\mathbf{D}]^{-1}\mathbf{D}^\dagger$ is the orthogonal projection on the subspace $\langle\mathbf{D}\rangle$ and $\mathbf{P}_{\mathbf{D}^\perp} = \mathbf{I} - \mathbf{P_D}$ projects on the subspace that is orthogonal to the subspace $\langle\mathbf{D}\rangle$.

7.2 Problem Formulation

Consider observations which are captured from a linear array with N sensors or N time snapshots from a single sensor given by $\mathbf{y}$. This observation vector can be decomposed using a general linear model (GLM) [49] to

$$\mathbf{y} = \mathbf{H}\boldsymbol{\theta} + \mathbf{B}\boldsymbol{\phi} + \boldsymbol{\xi}, \tag{7.1}$$

where $\mathbf{y} \in \mathbb{C}^N$ is the measurement, $\mathbf{H} \in \mathbb{C}^{N\times p}$ is a known matrix whose columns span the subspace containing target signal and its corresponding deterministic parameter $\boldsymbol{\theta}$ is

unknown. $\mathbf{B} \in \mathbb{C}^{N\times t}$ is also a known matrix whose columns span the subspace containing interference and corresponding coordinate $\boldsymbol{\phi}$ is unknown. $\boldsymbol{\xi}$ is the noise term and $\boldsymbol{\xi} \sim \mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{R})$ where $\mathbf{R}$ is an unknown covariance with banded structure. Note that the subspace $\langle\mathbf{H}\rangle$ and the subspace $\langle\mathbf{B}\rangle$ are linearly independent. It is assumed that a secondary i.i.d., signal-free training dataset is available collected from a partially similar environment as $\boldsymbol{\xi}$, i.e.,$\{\mathbf{n}_k\}_{k=1}^K$ is available for estimating $\mathbf{R}$, where $\mathbf{n}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{R})$. The subspace detection problem now can be formulated as the following hypothesis testing problem:

$$\mathcal{H}_0 = \begin{cases} \mathbf{y} = \mathbf{B}\boldsymbol{\phi} + \boldsymbol{\xi} \sim \mathcal{CN}(\mathbf{B}\boldsymbol{\phi}, \sigma^2\mathbf{R}) \\ \mathbf{y}_k = \mathbf{n}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}) \end{cases}, \tag{7.2}$$

and

$$\mathcal{H}_1 = \begin{cases} \mathbf{y} = \mathbf{H}\boldsymbol{\theta} + \mathbf{B}\boldsymbol{\phi} + \boldsymbol{\xi} \sim \mathcal{CN}(\mathbf{H}\boldsymbol{\theta} + \mathbf{B}\boldsymbol{\phi}, \sigma^2\mathbf{R}) \\ \mathbf{y}_k = \mathbf{n}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}) \end{cases}, \tag{7.3}$$

where $\mathcal{H}_0$ and $\mathcal{H}_1$ are the null and alternative hypotheses, respectively.

7.3 Background

In this section, we briefly describe ASD which is derived for almost the same problem without using any prior information about $\mathbf{R}$.

7.3.1 Adaptive Subspace Detector

In the case of non-homogeneous noise and linear signal approximation, the ASD is a widely used detector based on the GLRT concept. The original form of this detector was proposed without considering the interference. Due to the connection between the ASD and matched subspace detector (MSD) [49] the extension of ASD to include the interference is

$$\ell_{ASD}(\mathbf{y}) = \frac{\mathbf{y}^\dagger\mathbf{S}^{-\frac{1}{2}}\mathbf{P}_{\tilde{\mathbf{B}}^\perp}\mathbf{P}_{\tilde{\mathbf{H}}}\mathbf{P}_{\tilde{\mathbf{B}}^\perp}\mathbf{S}^{-\frac{1}{2}}\mathbf{y}}{\mathbf{y}^\dagger\mathbf{S}^{-\frac{1}{2}}\mathbf{P}_{\tilde{\mathbf{B}}^\perp}\mathbf{S}^{-\frac{1}{2}}\mathbf{y}}, \tag{7.4}$$

where $\tilde{\mathbf{H}} = \mathbf{S}^{-\frac{1}{2}}\mathbf{H}$, $\tilde{\mathbf{B}} = \mathbf{S}^{-\frac{1}{2}}\mathbf{B}$ and $\mathbf{S} = \frac{1}{K}\sum_{k=1}^K \mathbf{n}_k\mathbf{n}_k^\dagger$ is the sample covariance matrix. The exact distributions of above test was derived in [179]. Note that ASD does not consider any structure for the covariance and uses the sample covariance as an estimate of $\mathbf{R}$. We use this test as a benchmark for comparison.

7.3.2 Banded Covariance

Before defining the banded covariance structure, we need to define the bandwidth of a matrix.

Definition1: The Bandwidth of a given squared matrix $\mathbf{A}$ is equal to b when $\forall i, j$ when $|i-j| > b$ then $a_{i,j} = 0$ and $\mathbf{A}$ has the following visual structure:

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & \dots & a_{1,1+b} & & & & \\ & \ddots & \ddots & \ddots & & 0 & \\ & & a_{i,i-b} & \dots & a_{i,i+b} & & \\ & 0 & & \ddots & \ddots & \ddots & \\ & & & & a_{N,N-b} & \dots & a_{N,N} \end{bmatrix}.$$

Definition2: The matrix $\mathbf{A}$ of size N is called a banded matrix if its bandwidth (b) is sufficiently small and it is represented by $\mathbf{A}_N^b$.

7.4 Proposed Method

Now, we develop our proposed detector for the model (7.1) using an a two steps procedure. Note that the joint likelihood of secondary data and received test data in receiver side is

$$f\left(\mathbf{y}, \mathbf{n}_1, \dots, \mathbf{n}_K\right) = f\left(\mathbf{y}\right) \prod_{k=1}^{K} f\left(\mathbf{n}_k\right), \tag{7.5}$$

where for a zero mean $\mathbf{y}$

$$f(\mathbf{y}) = \frac{1}{\pi^N \det(\sigma^2 \mathbf{R})} \exp\{-\frac{1}{\sigma^2} \mathbf{y}^\dagger \mathbf{R}^{-1} \mathbf{y}\}. \tag{7.6}$$

Using (7.1), (7.5) and (7.6), and by defining $\mathbf{C} = [\mathbf{H}, \mathbf{B}]$ and $\boldsymbol{\beta} = [\boldsymbol{\theta}^\top, \boldsymbol{\phi}^\top]^\top$, under $\mathcal{H}_1$ we have

$$\begin{aligned} f_1\left(\mathbf{y}, \mathbf{n}_1, \dots, \mathbf{n}_K\right) &= \frac{1}{\pi^N \det(\sigma^2 \mathbf{R})} \exp\{-\frac{1}{\sigma^2} (\mathbf{y} - \mathbf{C}\boldsymbol{\beta})^\dagger \mathbf{R}^{-1} (\mathbf{y} - \mathbf{C}\boldsymbol{\beta})\} \\ &\times \left\{ \frac{1}{\pi^N \det(\mathbf{R})} \right\}^K \prod_{k=1}^{K} \exp\{-\mathbf{n}_k^\dagger \mathbf{R}^{-1} \mathbf{n}_k\}. \end{aligned} \tag{7.7}$$

Under $\mathcal{H}_0$, the density $f_0 = f_1|_{\boldsymbol{\theta}=\mathbf{0}}$. Finally, noting that $\mathbf{y}^\dagger\mathbf{R}^{-1}\mathbf{y} = \mathrm{tr}(\mathbf{y}^\dagger\mathbf{R}^{-1}\mathbf{y}) = \mathrm{tr}(\mathbf{R}^{-1}\mathbf{y}\mathbf{y}^\dagger)$, we can rewrite f_1 as

$$f_1\left(\mathbf{y}, \mathbf{n}_1, \ldots, \mathbf{n}_K\right) = \left\{ \frac{1}{\pi^N \det(\mathbf{R}) \sigma^{2N/K+1}} \exp\{-\mathrm{tr}(\mathbf{R}^{-1}\mathbf{T}_1)\} \right\}^{K+1}, \tag{7.8}$$

where $\mathbf{T}_1$ is the total sample covariance constructed using both the training and test data given by

$$\mathbf{T}_1 = \frac{K}{K+1}\mathbf{S} + \frac{1}{K+1}\frac{1}{\sigma^2}(\mathbf{y} - \mathbf{C}\boldsymbol{\beta})(\mathbf{y} - \mathbf{C}\boldsymbol{\beta})^\dagger. \tag{7.9}$$

Using a similar approach under $\mathcal{H}_0$

$$f_0\left(\mathbf{y}, \mathbf{n}_1, \ldots, \mathbf{n}_K\right) = \left\{ \frac{1}{\pi^N \det(\mathbf{R}) \sigma^{2N/K+1}} \exp\{-\mathrm{tr}(\mathbf{R}^{-1}\mathbf{T}_0)\} \right\}^{K+1}, \tag{7.10}$$

where

$$\mathbf{T}_0 = \frac{K}{K+1}\mathbf{S} + \frac{1}{K+1}\frac{1}{\sigma^2}(\mathbf{y} - \mathbf{B}\boldsymbol{\phi})(\mathbf{y} - \mathbf{B}\boldsymbol{\phi})^\dagger. \tag{7.11}$$

Now for known $\mathbf{R}$, the GLRT is given by

$$\ell_{GLRT}(\mathbf{y}) = \frac{\sup_{\boldsymbol{\beta},\sigma^2} f_1\left(\mathbf{y}, \mathbf{n}_1, \ldots, \mathbf{n}_K\right)}{\sup_{\boldsymbol{\phi},\sigma^2} f_0\left(\mathbf{y}, \mathbf{n}_1, \ldots, \mathbf{n}_K\right)}. \tag{7.12}$$

We begin with the denominator estimate $\boldsymbol{\phi}$ under $\mathcal{H}_0$. After some algebraic simplification (see appendix E) it can be shown that

$$\hat{\boldsymbol{\phi}} = \left(\bar{\mathbf{B}}^\dagger\bar{\mathbf{B}}\right)^{-1}\bar{\mathbf{B}}^\dagger\bar{\mathbf{y}}, \tag{7.13}$$

where $\bar{\mathbf{B}} = \mathbf{R}^{-\frac{1}{2}}\mathbf{B}$ and $\bar{\mathbf{y}} = \mathbf{R}^{-\frac{1}{2}}\mathbf{y}$. therefore, the denominator of (7.12) after one step maximization can be recast as

$$\left\{ \frac{1}{\pi^N \det(\mathbf{R}) \sigma^{2N/K+1}} \right\}^{K+1} \exp\{-\mathrm{tr}(K\mathbf{R}^{-1}\mathbf{S} + \frac{1}{\sigma^2}\bar{\mathbf{y}}^\dagger\mathbf{P}_{\bar{\mathbf{B}}^\perp}\bar{\mathbf{y}})\}.$$

After replacing $\boldsymbol{\phi}$ with $\hat{\boldsymbol{\phi}}$ in f_0, we estimate σ^2. It is easy to show that σ^2 can be estimated from

$$\hat{\sigma}_0^2 \;=\; \arg\min_{\sigma^2} \frac{1}{\sigma^2}(\bar{\mathbf{y}}^\dagger\mathbf{P}_{\bar{\mathbf{B}}^\perp}\bar{\mathbf{y}}) + N\log(\sigma^2), \tag{7.14}$$

which yields

$$\hat{\sigma}_0^2 = (\bar{\mathbf{y}}^\dagger \mathbf{P}_{\bar{\mathbf{B}}^\perp} \bar{\mathbf{y}})/N, \tag{7.15}$$

where the index 0 shows the estimation under the null hypothesis. Similarly, the MLE of unknown parameters under $\mathcal{H}_1$ are given by

$$\hat{\boldsymbol{\beta}} = \left(\bar{\mathbf{C}}^\dagger \bar{\mathbf{C}}\right)^{-1} \bar{\mathbf{C}}^\dagger \bar{\mathbf{y}}, \tag{7.16}$$

and

$$\hat{\sigma}_1^2 = (\bar{\mathbf{y}}^\dagger \mathbf{P}_{\bar{\mathbf{C}}^\perp} \bar{\mathbf{y}})/N, \tag{7.17}$$

where $\bar{\mathbf{C}} = \mathbf{R}^{-\frac{1}{2}}\mathbf{C}$. Replacing the above estimated parameters in (7.12) and using a monotone increasing function, the GLRT based test is given by

$$\ell_{GLRT-I}(\mathbf{y}) = \frac{\bar{\mathbf{y}}^\dagger \mathbf{P}_{\bar{\mathbf{B}}^\perp} \bar{\mathbf{y}}}{\bar{\mathbf{y}}^\dagger \mathbf{P}_{\bar{\mathbf{C}}^\perp} \bar{\mathbf{y}}}. \tag{7.18}$$

Note that the above test depends on the measurement $\mathbf{y}$, the matrices $\mathbf{H}$ and $\mathbf{B}$, and finally the covariance matrix $\mathbf{R}$ that initially was assumed known. Now, to be able to construct the above test, we replace $\mathbf{R}$ in all the above derived expressions by a consistent estimator which is described in the following proposition. We partition the banded covariance $\mathbf{R}_N^b$ as

$$\mathbf{R}_N^b = \begin{bmatrix} \mathbf{R}_{N-1}^b & \tilde{\mathbf{r}}_{1N} \\ \tilde{\mathbf{r}}_{1N}^\dagger & r_{N,N} \end{bmatrix},$$

where $\tilde{\mathbf{r}}_{1N}^\dagger = (0, \ldots, 0, r_{N,N-b}, \ldots, r_{N,N-1})$. The top left corner matrix is also banded with the same bandwidth but reduced matrix size. This property can be used to estimate the unknown covariance matrix in a gradually increasing dimension scheme introduced in the following proposition.

Proposition 1: Assume K i.i.d. samples $\mathbf{Z} = \{\mathbf{n}_1, \cdots, \mathbf{n}_K\} \in \mathbb{C}^{N \times K}$ are available where $\mathbf{n}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{R})$ and $\mathbf{R}$ is a banded covariance with known bandwidth. The estimator $\hat{\mathbf{R}}$ is given by following two steps:

- Use the MLE for $\mathbf{R}_{b+1}^{b}$ (from now on the upper index will be removed when it is clear from the text).

- For $i = b+2, \ldots, N$, sequentially increase the dimension and use the following estimators where for each i, let $j = i-b, \cdots, i-1$.

$$\begin{aligned}
\hat{r}_{i,j} &= \hat{\boldsymbol{\lambda}}_{i,j} \frac{\det(\hat{\mathbf{R}}_{i-1})}{\det(\hat{\mathbf{R}}_{i-2})}, \\
\hat{r}_{i,i} &= \frac{1}{K} \mathbf{z}^{i*} \left(\mathbf{I}_K - \hat{\mathbf{F}}_{i-1} \left(\hat{\mathbf{F}}_{i-1}^{\dagger} \hat{\mathbf{F}}_{i-1} \right)^{-1} \hat{\mathbf{F}}_{i-1}^{\dagger} \right) \mathbf{z}^{i\top} \\
&\quad + \tilde{\mathbf{r}}_{1i}^{\dagger} \hat{\mathbf{R}}_{i-1}^{-1} \tilde{\mathbf{r}}_{1i},
\end{aligned}$$

where

$$\begin{aligned}
\tilde{\mathbf{r}}_{1i}^{\dagger} &= (0, \ldots, 0, \hat{r}_{i,i-b}, \ldots, \hat{r}_{i,i-1}), \\
\hat{\boldsymbol{\lambda}}_i &= (\hat{\lambda}_{i,i-b}, \cdots, \hat{\lambda}_{i,i-1})^{\top} \\
&= \left(\hat{\mathbf{F}}_{i-1}^{\dagger} \hat{\mathbf{F}}_{i-1} \right)^{-1} \hat{\mathbf{F}}_{i-1}^{\dagger} \mathbf{z}^{i\top}, \\
\hat{\mathbf{F}}_{i-1} &= (\hat{\mathbf{z}}_{i-1,i-b}, \cdots, \hat{\mathbf{z}}_{i-1,i-1}),
\end{aligned}$$

and

$$\hat{\mathbf{z}}_{i-1,m} = \sum_{n=1}^{i-1} (-1)^{m+n} \frac{\det(\hat{\mathbf{M}}_{i-1}^{nm})}{\det(\hat{\mathbf{R}}_{i-2})} \mathbf{z}^{n\top},$$

where $\hat{\mathbf{M}}_{i-1}^{nm}$ is the matrix which is obtained when the $n-$th row and $m-$column have been removed from $\hat{\mathbf{R}}_{i-1}$.

Proof: See Appendix F.

7.5 Simulation Results

In order to show the superiority of our proposed algorithm over the conventional ASD and AMF, we provide extended simulation based studies.

7.5.1 Synthetic Data

In this simulation, the signal to noise ratio (SNR) is defined as

$$\text{SNR} = 10\log_{10}(\boldsymbol{\theta}^{\dagger}\mathbf{H}^{\dagger}(\sigma^2\mathbf{R})^{-1}\mathbf{H}\boldsymbol{\theta}). \tag{7.19}$$

The subspace matrix $\mathbf{H}$ is chosen to be

$$\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_p], \tag{7.20}$$

where

$$\mathbf{h}_i = 1/\sqrt{N}[1, e^{-j2\pi f_i}, \cdots, e^{-j2\pi f_i(N-1)}]^{\top}, \tag{7.21}$$

and $f_i = i \times 0.05 + 0.05, i = 1, \cdots, p$. The subspace matrix $\mathbf{B}$ is chosen as

$$\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_t], \tag{7.22}$$

where

$$\mathbf{b}_i = 1/\sqrt{N}[1, e^{-j2\pi g_i}, \cdots, e^{-j2\pi g_i(N-1)}]^{\top}, \tag{7.23}$$

and $g_i = i \times 0.02 - 0.12, i = 1, \cdots, t$. The number of sensor is $N = 15$ and $p = t = 2$. To illustrate the performance of our proposed detector, 5000 test vectors are generated using Monte Carlo, and the receiver operating characteristic (ROC) curves calculated and shown in figure 7.1 for different K values and randomly generated covariance with $b = 2$ and $\sigma^2 = 9$. The target signal is added with the probability of $1/2$ to the test data with SNR= 10dB. It can be seen that starting with a small K value, the proposed banded AMF has significantly better performance than the classical ASD and AMF. Figure 7.2 shows the histogram of three tests for generated samples and as it is shown, the banded AMF provides the highest separation between $\mathcal{H}_0$ and $\mathcal{H}_1$. This makes it a better detector. In general, as K increases towards asymptotic values (see figure 7.3 and its caption for more discussion), all three detectors use almost the same estimate of the covariance matrix which yields the same performance. However, for small K, the prior information about the covariance structure plays an important role and provides a better estimate in term of Frobenius norm of error matrix.

7.5.2 Real fMRI Data

In this section, we apply the proposed detector to the real fMRI data for the task of neural activation detection [117]. The main aim of fMRI data analysis is to localize the neural activity in the brain for a specific task (in a task based fMRI) or in resting state fMRI. Two task based fMRI datasets are selected to show the power of our proposed algorithm, the block-paradigm right finger tapping task (BPRFT) and event related right finger tapping (ERRFT) data [28]. As a summery of data generating process, in BPRFT data, a 15 sec task period alternated with 72 sec resting period was repeated for 4 times followed by 30 sec rest. The total recording time was 480 sec. In ERRFT total recording time was 650 sec. After dummy scan, right finger tapping is done 40 times randomly and then it had a 30 sec resting time. These datasets have 592895 voxels. The data was preprocessed by following steps (similar to [180]) including motion correction, temporal pre-whitening, drift removal and slice time correction. The matrix $\mathbf{H}$ is approximated by a rank-1 signal which is computed by applying the convolution operator to the task stimuli and hemodynamic response function (HRF). In order to omit the effect of scanner in the recorded data, an interference matrix $\mathbf{B}$ with DC values in the first column and two low frequency components in other columns are formed. We have repeated the experiment for 3 cases of ($b = 0, b = 5$ and $b = 20$). The results show the activity maps over fixed brain slices from different angels together with the activity maps computed by AMF. Note that, in this experiment we did not have access to a secondary data to estimate the covariance matrix in exact steps of the assumed model. To make this assumption valid, we use a classical test statistic (MSD with small threshold [49]) to detect the voxels that have minimum impact from the signal of interest and interference. This step is repeated until we could provide at least $3 \times K$ samples as the secondary or training data. The results are shown in figures 7.4 and 7.5. In both figures, the first row shows the localized active areas using AMF. It can be observed that some false positive voxels exist in the map (e.g., a small active circle in top right part of BPRFT data) which are not consistent with our knowledge about neural activation (e.g., left hand side of the brain is responsible of right hand side of human bodies and smooth activation change over the voxels) and the existing results in the literature using a data driven approaches (e.g., [181]). On the other hand using our proposed method with $b = 0$, we obtain a slightly different activation map with a smaller number of false positive areas. However, this small bandwidth may not model the true underlying

covariance, so we gradually increase the bandwidth and show the activation map in the third rows. The model is still within the assumptions that we made for our proposed method (i.e., $b << N$) and the result is consistent with previous bandwidth and has small false positives. In the last experiment, we increase b to 20, to show the case that the assumption on the covariance structure is no longer valid and it can be observed that the activation map is again becoming full of false positives and gradually tends to the result of AMF.

7.6 Conclusion

In this chapter, the subspace detection problem in the presence of structured noise and interference is addressed. Instead of using the sample covariance estimate, in the adaptive matched filter approach, a new explicit estimator is used. This estimator is based on successive maximizing of conditional factors of joint likelihood function. The advantage of this estimator over the classical sample covariance estimate is the use of prior information about the structure of the true covariance. As a consequence, it was shown in simulations that the proposed estimator has a better estimation accuracy in finite sample regime which happens a lot in signal processing applications. Naturally, this improved estimation leads to better detection rate for a fixed false alarm rate. Using, simulated and real fMRI data, it was shown that the proposed method can be useful in subspace detection and neural activity detection due to the lack of correlation between distinct observations and the results can be more reliable.

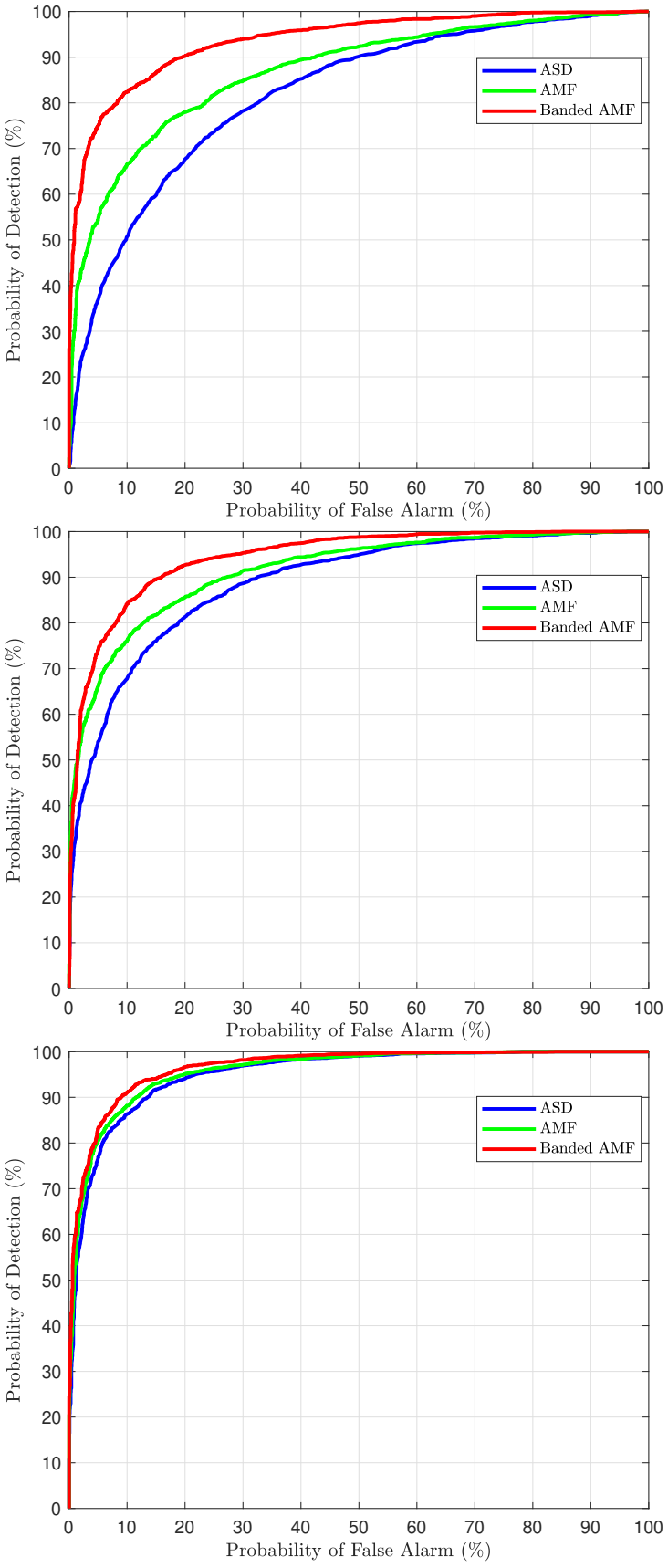

FIGURE 7.1: ROC curve of ASD, AMF and banded AMF for designed experiment with $K = 20$ (top), $K = 30$ (middle) and $K = 100$ (buttom).

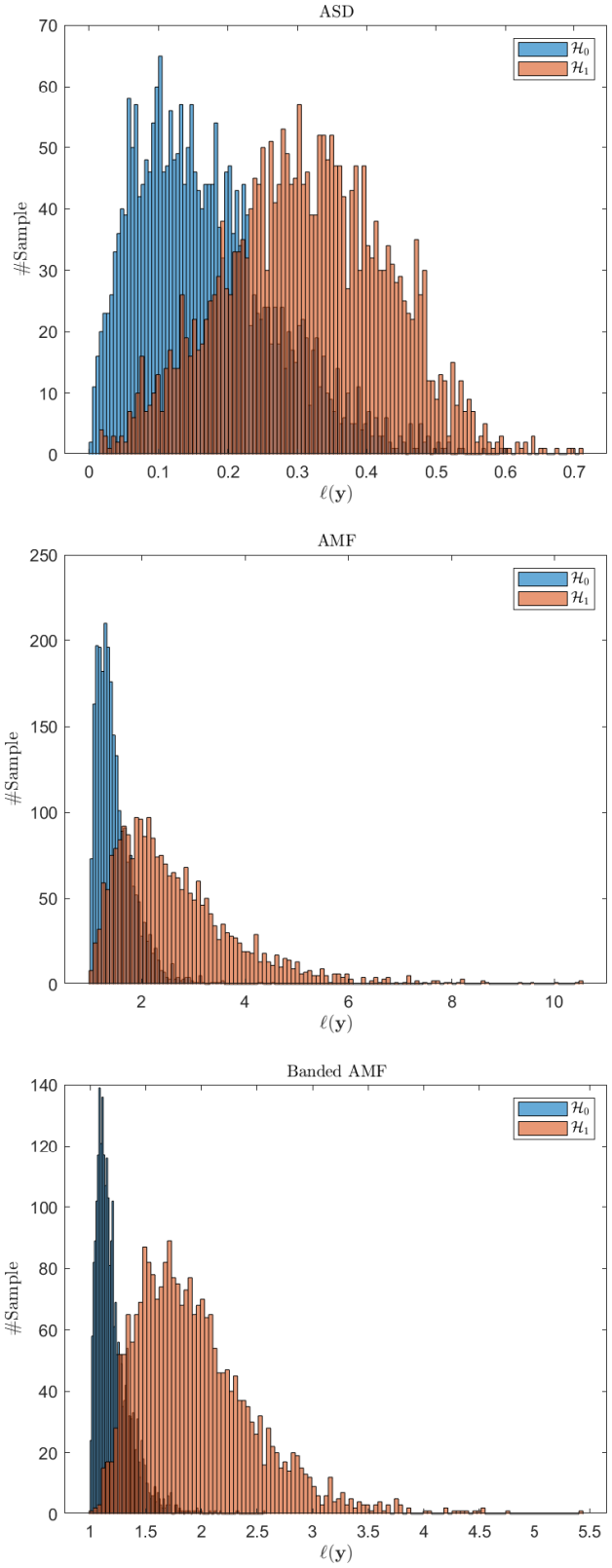

FIGURE 7.2: Histogram comparison of the tests statistics when $N = 15$, $K = 20$ and $\sigma^2 = 9$. The proposed banded AMF has the most separability and consequently the maximum detection rate was achieved.

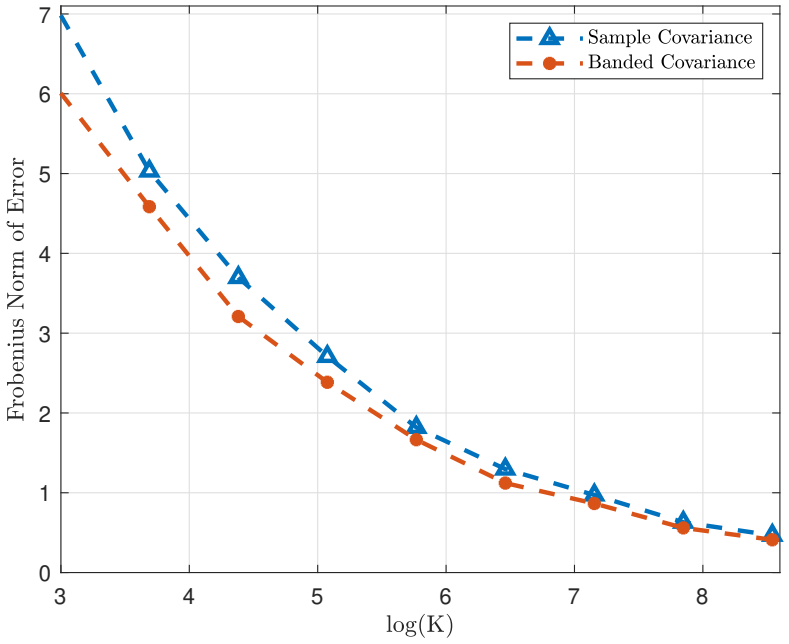

FIGURE 7.3: Average Frobenius norm of error between the estimated and true covariance over 30 random trials ($N = 5, b = 2$). It can be seen that gradually, with increasing the number of samples, both estimators converge to their true value which shows that they are both consistent. In low sample regime, the proposed estimator has uniformly a better error rate and gradually with increasing K, it has the same error rate as sample covariance. This behavior makes this estimator more suitable for the applications that large number of training samples is not available.

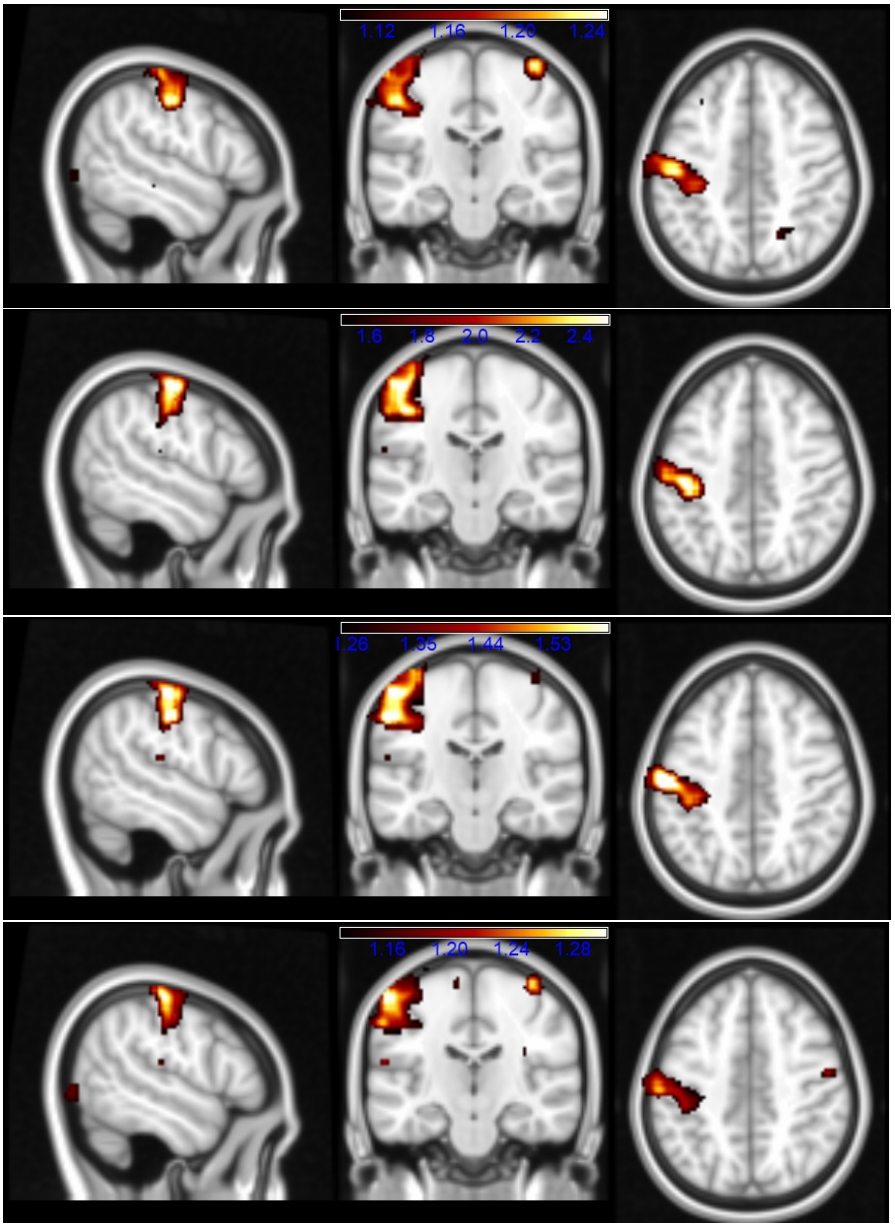

FIGURE 7.4: Active area of brain in BPRFT data, from the top, AMF, banded AMF with $b = 0, b = 5$ and $b = 20$.

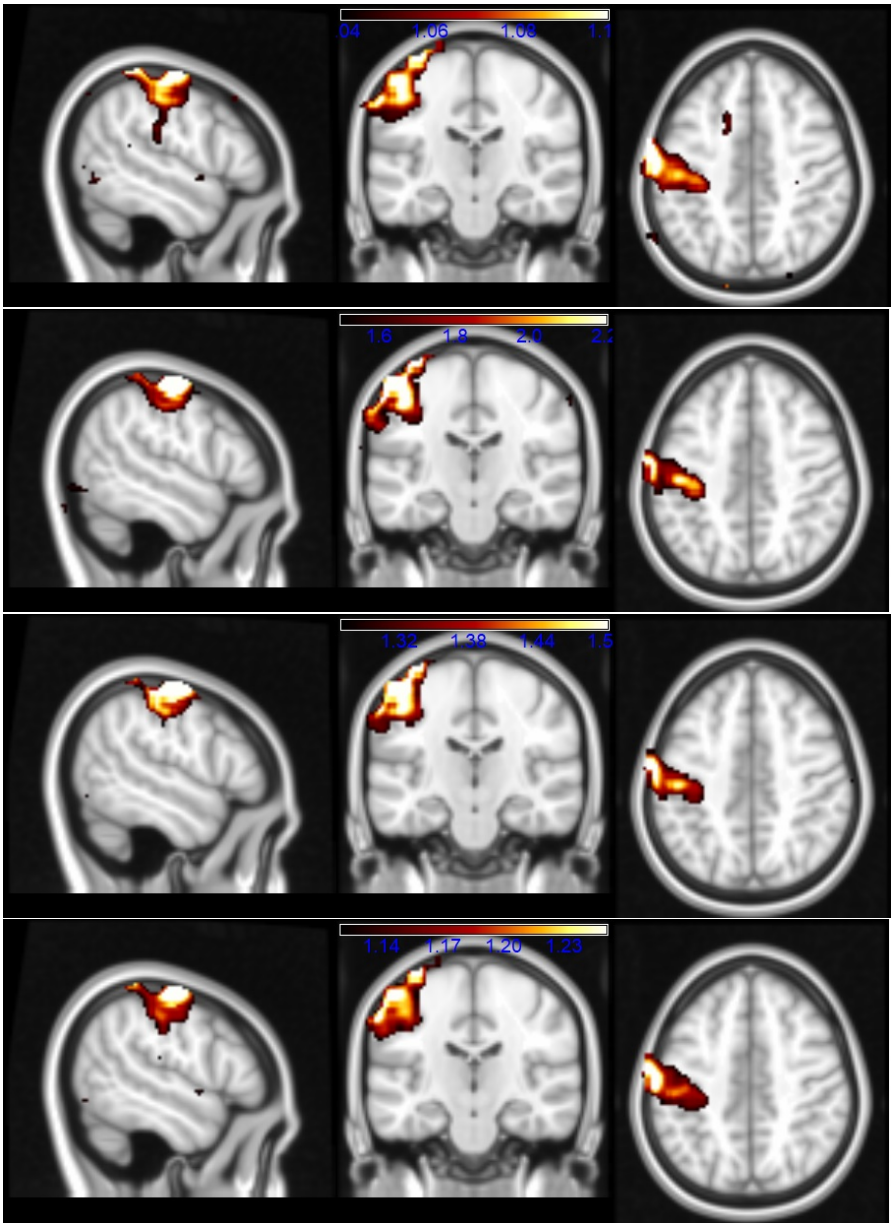

FIGURE 7.5: Active area of brain in ERRFT data, from the top, AMF, banded AMF with $b = 0, b = 5$ and $b = 20$.

Chapter 8

Conclusion and Future Work

8.1 Conclusions

In this thesis, various detectors and estimators were provided to be used in fMRI data analysis, in particular activation detection task, and also in other imaging applications. The main goal of theses algorithms was to furnish and extend classical methods, using a natural way to achieve a better performance and robustness.

In the third chapter of this thesis, we proposed three robust detectors using the GLRT, Wald and Rao framework. The $\alpha-$divergence was used as the generalization of the KL divergence to derive a new form of estimator which is adjustable with a single parameter. Furthermore, a detailed algorithmic approach for tuning this parameter and asymptotic analysis of the proposed detectors were provided. Simulations and experimental results were also shown for the task of subspace detection, fMRI activation detection, hyperspectral image analysis and SAR detection. All comparisons showed the superiority of the proposed detectors over the existing well known methods in the literature.

In the fourth chapter of this thesis, four variants of PCA algorithm were proposed using the same robust measure to achieve robustness against outliers in the task of estimation of principal components. These methods can be used to derive the existing patterns in the data in the case that the signal subspace is unknown. These methods were also extensively evaluated in the case of synthetic data, simulated fMRI data, foreground-background separation in the videos and finally saliency map identification of the images. The results indicated the superiority of the proposed methods.

In the fifth chapter, a new robust adaptive matched filter was proposed that is beneficial in the case that the secondary dataset (training data) is not signal free which may happen in fMRI and Radar signals. In this setup, using the classical estimation of the covariance matrix may generate a significantly biased estimator. A robust covariance estimator was fed into adaptive matched filter to fix such flaws.
In the sixth chapter, we focused on using the prior knowledge about the structure of the data to significantly decrease the number of required samples to be able to effectively perform the detection task. Therefore, a low rank assumption was considered for the observed data and the detection rate showed the improvement in the final accuracy due to use of prior information.
Finally, in the seventh chapter, we continued to use the prior information in fMRI data analysis. In this work, a banded structure for the covariance matrix of the complex Gaussian noise was assumed which represents the case that the noise is locally correlated. Using this assumption, a modified adaptive matched filter was developed which was applied to the real fMRI data, and successfully offered a better activation map in the term of spatial consistency.

8.2 Future Works

In presented works, we mostly used a single Gaussian density as the nominal model. However, in a more complex scenarios, it is always useful to use mixture models (e.g., Gaussian mixture model) to obtain a high order model for a better approximation of the data. Therefore, a natural extension of the proposed method can be developing a new expectation maximization algorithm with the use of the robust measure while under single component the presented results can be recovered. The benefit of this method can be keeping robustness property even with increasing the model order.
Most detection approaches in correlated noise multivariate signals are based on strong assumptions. There is not enough literature for the case that a robust measure is used as an alternative of maximum likelihood approach. As the future work, it would be also interesting to use $\alpha-$divergence in adaptive detection framework.
In this thesis, we focused on using $\alpha-$divergence as an appropriate alternative of the KL divergence. For extending the results to a more general family of the estimators and detectors, the $f-$divergence and Bregman divergences are also provided in information

geometry that can naturally offer a broad ranges of properties.

In the case of principal component and subspace estimation, we just focused on the case that the subspace is stationary and does not change over time or space. There is another category of work where the signal subspace is allowed to smoothly change. This type of the work is called "Subspace Tracking". Since existing robust subspace trackers use the simple form of Huber and classical robust loss functions, the robustification of the subspace tracker methods can be also improved by robust measures form information geometry.

Appendix A

The Proof of Proposition 1, Ch. 3

With considering $L_\alpha(\mathbf{y},\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N}\log_\alpha\{f(\mathbf{y}_i,\boldsymbol{\beta})\}$ and expanding the $\frac{\partial L_\alpha(\mathbf{y},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$ using Taylor theorem around $\boldsymbol{\beta}^*$ gives

$$\mathbf{0} = \frac{\partial L_\alpha(\mathbf{y},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = \frac{\partial L_\alpha(\mathbf{y},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} + \frac{\partial^2 L_\alpha(\mathbf{y},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^2}\big|_{\boldsymbol{\beta}=\bar{\boldsymbol{\beta}}}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*),$$

where $\bar{\boldsymbol{\beta}}$ lies on the line that connects $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^*$. Based on above Taylor expansion, we have

$$\frac{\partial L_\alpha(\mathbf{y},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} = -\frac{\partial^2 L_\alpha(\mathbf{y},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^2}\big|_{\boldsymbol{\beta}=\bar{\boldsymbol{\beta}}}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*). \tag{A.1}$$

The first conclusion is based on the consistency of $\hat{\boldsymbol{\beta}}$ as $N\to\infty$, $\mathbf{E}\left[\frac{\partial L_\alpha(\mathbf{y},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right]\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} = \mathbf{0}$ and we can conclude that $\frac{\partial^2 L_\alpha(\mathbf{y},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\big|_{\boldsymbol{\beta}=\bar{\boldsymbol{\beta}}} \to \frac{\partial^2 L_\alpha(\mathbf{y},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}$ w.p.1. Using the results in [90, 113], $(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*) \to_d \mathcal{N}(\mathbf{0},\mathbf{K}(\boldsymbol{\beta}^*))$ as $N\to\infty$ where $\mathbf{K}(\boldsymbol{\beta}) = \mathbf{E}\left[\frac{\partial^2\ell_\alpha(\mathbf{y},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^2}\right]^{-1}\mathbf{E}\left[\frac{1}{N}\frac{\partial\ell_\alpha(\mathbf{y},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\frac{\partial\ell_\alpha(\mathbf{y},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}^\top\right]\mathbf{E}\left[\frac{\partial^2\ell_\alpha(\mathbf{y},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^2}\right]^{-1}$. Therefore by expression (A.1) the asymptotic variance of the score function evaluated at the true value can be computed by

$$\begin{aligned}
\mathrm{Var}\left(\frac{\partial L_\alpha(\mathbf{y},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} &= \mathbf{E}\left[\frac{\partial^2\ell_\alpha(\mathbf{y},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^2}\right]\Big|_{\boldsymbol{\beta}=\bar{\boldsymbol{\beta}}}\mathbf{K}(\boldsymbol{\beta}^*)\mathbf{E}\left[\frac{\partial^2\ell_\alpha(\mathbf{y},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^2}\right]\Big|_{\boldsymbol{\beta}=\bar{\boldsymbol{\beta}}} \\
&= \mathbf{E}\left[\frac{1}{N}\frac{\partial\ell_\alpha(\mathbf{y},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\frac{\partial\ell_\alpha(\mathbf{y},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}^\top\right]\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} \\
&= \mathbf{F}(\boldsymbol{\beta}^*),
\end{aligned}$$

where $\mathbf{F}$ is the defined $\alpha-$information matrix. Finally based on central limit theorem asymptotically $\frac{\partial L_\alpha(\mathbf{y},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} \rightarrow_d \mathcal{N}(\mathbf{0}, \mathbf{F}(\hat{\boldsymbol{\beta}}))$ and the proof is complete.

Appendix B

The Proof of Proposition 2, Ch. 3

Proposition 2 reflects a known result in literature [182], but is proven for our defined $\alpha-$logarithm function to make the paper self-contained. Let's define $g_N(\boldsymbol{\omega}) = \sum_{i=1}^{N} \log_\alpha \{f(y_i, \boldsymbol{\omega})\}$ and assume that $\hat{\boldsymbol{\omega}}_1 = [\hat{\boldsymbol{\theta}}_1^\top, \hat{\boldsymbol{\lambda}}_1^\top]^\top$ and $\hat{\boldsymbol{\omega}}_0 = [\mathbf{0}^\top, \hat{\boldsymbol{\lambda}}_0^\top]^\top$ represent the estimated parameters under the alternative and null hypotheses, respectively. Expanding the $g_N(\boldsymbol{\omega}^*)$ into Taylor series around $\hat{\boldsymbol{\omega}}_1$ gives

$$g_N(\boldsymbol{\omega}^*) \simeq \quad g_N(\hat{\boldsymbol{\omega}}_1) + (\boldsymbol{\omega}^* - \hat{\boldsymbol{\omega}}_1)^\top \nabla_{\boldsymbol{\omega}} g_N(\hat{\boldsymbol{\omega}}_1) + \frac{1}{2}(\boldsymbol{\omega}^* - \hat{\boldsymbol{\omega}}_1)^\top \nabla^2_{\boldsymbol{\omega}} g_N(\bar{\boldsymbol{\omega}})(\boldsymbol{\omega}^* - \hat{\boldsymbol{\omega}}_1),$$

where $\bar{\boldsymbol{\omega}}$ lies on the line that connects $\hat{\boldsymbol{\omega}}_1$ and $\boldsymbol{\omega}^*$. Note that when $N \to \infty$, $\hat{\boldsymbol{\omega}} \to_p \boldsymbol{\omega}^*$ and $\nabla_{\boldsymbol{\omega}} g_N(\hat{\boldsymbol{\omega}}) = \mathbf{0}$ and empirically $\mathbf{M} = \frac{1}{N}\sum_{i=1}^{N} -\frac{\partial \mathbf{S}}{\partial \boldsymbol{\omega}}$ where $\mathbf{s} = \nabla_{\boldsymbol{\omega}} g_N$. Therefore, we can rewrite above approximation as

$$2[g_N(\hat{\boldsymbol{\omega}}_1) - g_N(\boldsymbol{\omega}^*)] \simeq N(\hat{\boldsymbol{\omega}}_1 - \boldsymbol{\omega}^*)^\top \mathbf{M}(\hat{\boldsymbol{\omega}}_1 - \boldsymbol{\omega}^*).$$

Partitioning $\mathbf{M}$ to 4 sub blocks of $\mathbf{M}_{\theta\theta}, \mathbf{M}_{\theta\lambda}, \mathbf{M}_{\lambda\theta}$ and $\mathbf{M}_{\lambda\lambda}$ and also using similar approach we can write the approximation under the null hypothesis as

$$2[g_N(\hat{\boldsymbol{\lambda}}_0) - g_N(\boldsymbol{\lambda}^*)] \simeq N(\hat{\boldsymbol{\lambda}}_0 - \boldsymbol{\lambda}^*)^\top \mathbf{M}_{\lambda\lambda}(\hat{\boldsymbol{\lambda}}_0 - \boldsymbol{\lambda}^*).$$

Now we can conclude that

$$\begin{aligned} NT_{G_r}(\mathbf{y}) &= 2[g_N(\hat{\boldsymbol{\omega}}_1) - g_N(\hat{\boldsymbol{\omega}}_0)] \\ &\simeq N(\hat{\boldsymbol{\omega}}_1 - \boldsymbol{\omega}^*)^\top \mathbf{M}(\hat{\boldsymbol{\omega}}_1 - \boldsymbol{\omega}^*) - N(\hat{\boldsymbol{\lambda}}_0 - \boldsymbol{\lambda}^*)^\top \mathbf{M}_{\lambda\lambda}(\hat{\boldsymbol{\lambda}}_0 - \boldsymbol{\lambda}^*). \end{aligned}$$

It is shown in [113] that asymptotically $\sqrt{N}(\hat{\boldsymbol{\omega}}_1 - \boldsymbol{\omega}^*) \simeq \mathbf{M}^{-1}\mathbf{s} \rightarrow_d \mathcal{N}(\mathbf{0},\mathbf{V})$ where $\mathbf{V} = \mathbf{M}^{-1}\mathbf{Q}\mathbf{M}^{-1}$ and $\mathbf{Q} = \mathbf{E}[\mathbf{s}\mathbf{s}^\top]$. Defining $\mathbf{R} = [\mathbf{0}_{(t+1)\times p}, \mathbf{I}_{(t+1)}]^\top$, and partitioning $\mathbf{s}$ to $\mathbf{s}_\theta$ and $\mathbf{s}_\lambda$, we have $\mathbf{R}^\top\mathbf{s} = \mathbf{s}_\lambda$. Therefore

$$\begin{aligned} NT_{G_r}(\mathbf{y}) &\simeq \mathbf{s}^\top\mathbf{M}^{-1}\mathbf{s} - \mathbf{s}^\top\mathbf{R}\mathbf{M}_{\lambda\lambda}^{-1}\mathbf{R}^\top\mathbf{s} \\ &= \mathbf{s}^\top\big[\mathbf{M}^{-1} - \mathbf{R}\mathbf{M}_{\lambda\lambda}^{-1}\mathbf{R}^\top\big]\mathbf{s} \\ &= \mathbf{s}^\top\mathbf{M}^{-1}\mathbf{M}\big[\mathbf{M}^{-1} - \mathbf{R}\mathbf{M}_{\lambda\lambda}^{-1}\mathbf{R}^\top\big]\mathbf{M}\mathbf{M}^{-1}\mathbf{s} \\ &= N(\hat{\boldsymbol{\omega}}_1 - \boldsymbol{\omega}^*)^\top\mathbf{M}\big[\mathbf{M}^{-1} - \mathbf{R}\mathbf{M}_{\lambda\lambda}^{-1}\mathbf{R}^\top\big]\mathbf{M}(\hat{\boldsymbol{\omega}}_1 - \boldsymbol{\omega}^*) \\ &= N(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^*)^\top\big[\mathbf{M}_{\theta\theta} - \mathbf{M}_{\theta\lambda}\mathbf{M}_{\lambda\lambda}^{-1}\mathbf{M}_{\lambda\theta}\big](\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^*), \end{aligned}$$

where in this paper $\boldsymbol{\theta}^* = \mathbf{0}$ and the first part of the proposition is proved. Let's define $\mathbf{e} = \sqrt{N}\mathbf{V}^{-\frac{1}{2}}(\hat{\boldsymbol{\omega}}_1 - \boldsymbol{\omega}^*)$, then $\mathbf{e} \rightarrow_d \mathcal{N}(\mathbf{0}, \mathbf{I}_{p+t+1})$ when $N \rightarrow \infty$. Therefore

$$\begin{aligned} NT_{G_r}(\mathbf{y}) &\simeq N(\hat{\boldsymbol{\omega}}_1 - \boldsymbol{\omega}^*)^\top\mathbf{M}\big[\mathbf{M}^{-1} - \mathbf{R}\mathbf{M}_{\lambda\lambda}^{-1}\mathbf{R}^\top\big]\mathbf{M}(\hat{\boldsymbol{\omega}}_1 - \boldsymbol{\omega}^*) \\ &= \mathbf{e}^\top\mathbf{V}^{\frac{1}{2}}\mathbf{M}\big[\mathbf{M}^{-1} - \mathbf{R}\mathbf{M}_{\lambda\lambda}^{-1}\mathbf{R}^\top\big]\mathbf{M}\mathbf{V}^{\frac{1}{2}}\mathbf{e} \\ &= \mathbf{e}^\top\mathbf{V}^{\frac{1}{2}}\mathbf{M}\left[\mathbf{M}^{-1} - \begin{bmatrix}\mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{\lambda\lambda}^{-1}\end{bmatrix}\right]\mathbf{M}\mathbf{V}^{\frac{1}{2}}\mathbf{e}, \end{aligned}$$

For the general quadratic form of $\mathbf{e}^\top\mathbf{A}\mathbf{e}$, the eigendecomposition of $\mathbf{A} = \mathbf{V}^{\frac{1}{2}}\mathbf{M}\left[\mathbf{M}^{-1} - \begin{bmatrix}\mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{\lambda\lambda}^{-1}\end{bmatrix}\right]\mathbf{M}\mathbf{V}^{\frac{1}{2}}$ has the form of $\mathbf{U}\boldsymbol{\Gamma}\mathbf{U}^\top$. Defining $\mathbf{z} = \mathbf{U}^\top\mathbf{e}$, it's easy to show $\mathbf{z} \rightarrow_d \mathcal{N}(\mathbf{0}, \mathbf{I}_{p+t+1})$ and we have

$$NT_{G_r}(\mathbf{y}) \simeq \mathbf{z}^\top\boldsymbol{\Gamma}\mathbf{z},$$

where

$$\begin{aligned} \boldsymbol{\Gamma} &= \operatorname{diag}\left(\operatorname{eig}\left(\mathbf{V}^{\frac{1}{2}}\mathbf{M}\left[\mathbf{M}^{-1} - \begin{bmatrix}\mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{\lambda\lambda}^{-1}\end{bmatrix}\right]\mathbf{M}\mathbf{V}^{\frac{1}{2}}\right)\right) \\ &= \operatorname{diag}\left(\operatorname{eig}\left(\mathbf{M}\mathbf{V}\mathbf{M}\left[\mathbf{M}^{-1} - \begin{bmatrix}\mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{\lambda\lambda}^{-1}\end{bmatrix}\right]\right)\right) \\ &= \operatorname{diag}\left(\operatorname{eig}\left(\mathbf{Q}\left[\mathbf{M}^{-1} - \begin{bmatrix}\mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{\lambda\lambda}^{-1}\end{bmatrix}\right]\right)\right), \end{aligned}$$

and finally

$$NT_{G_r} \rightarrow_d \sum_{i=1}^{p} \gamma_i z_i^2,$$

where γ_i are the nonzero diagonal elements of $\mathbf{\Gamma}$ and the proof is complete.

Appendix C

Derivation of U, Ch. 4

Fixing $\mathbf{V}$ reduces the original problem of robust sparse PCA to

$$\begin{aligned}\hat{\mathbf{U}} &= \arg\min_{\mathbf{U}} l_\alpha(\mathbf{Y}-\mathbf{U}\mathbf{V}) + \sum_{j=1}^{m}\pi_j\sqrt{r}\|\mathbf{u}^j\|_2 \\ &= \arg\min_{\mathbf{U}} \sum_{j=1}^{m}\sum_{i=1}^{N} l_\alpha(y_{ji}-\mathbf{u}^j\mathbf{v}_i) + \sum_{j=1}^{m}\pi_j\sqrt{r}\|\mathbf{u}^j\|_2.\end{aligned}$$

Taking the derivative with respect to $\mathbf{u}^j, j=1,\cdots,m$ gives

$$\pi_j\sqrt{r}\mathbf{s}^j - \sum_{i=1}^{N}\left(y_{ji}-\mathbf{u}^j\mathbf{v}_i\right)\mathbf{v}_i^\top \times \exp\{-\frac{1-\alpha}{2}\left(y_{ji}-\mathbf{u}^j\mathbf{v}_i\right)^2\} = \mathbf{0},$$

where $\mathbf{s}^j = \frac{\mathbf{u}^j}{\|\mathbf{u}^j\|_2}$, if $\|\mathbf{u}^j\|_2 \neq 0$. Otherwise if $\|\mathbf{u}^j\|_2 = 0$, $\mathbf{s}^j$ is a vector such that $\|\mathbf{s}^j\|_2 \leq 1$. In the case of $\|\mathbf{u}^j\|_2 \neq 0$ and defining $w_i = \exp\{-\frac{1-\alpha}{2}\left(y_{ji}-\mathbf{u}^j\mathbf{v}_i\right)^2\}$ the above equation can be written to

$$\sum_{i=1}^{N} w_i y_{ji}\mathbf{v}_i^\top = \pi_j\sqrt{r}\frac{\mathbf{u}^j}{\|\mathbf{u}^j\|_2} + \sum_{i=1}^{N} w_i\mathbf{u}^j\mathbf{v}_i\mathbf{v}_i^\top.$$

Rewriting the above equation in a matrix form by putting w_i in the main diagonal elements of matrix $\mathbf{W}$ gives

$$\begin{aligned}\mathbf{y}^j\mathbf{W}\mathbf{V}^\top &= \mathbf{u}^j\frac{\pi_j\sqrt{r}}{\|\mathbf{u}^j\|_2} + \mathbf{u}^j\mathbf{V}\mathbf{W}\mathbf{V}^\top \\ &= \mathbf{u}^j\left(\frac{\pi_j\sqrt{r}}{\|\mathbf{u}^j\|_2}\mathbf{I}_r + \mathbf{V}\mathbf{W}\mathbf{V}^\top\right). \qquad \text{(C.1)}\end{aligned}$$

and finally

$$\hat{\mathbf{u}}^j = \mathbf{y}^j\mathbf{W}\mathbf{V}^\top\left(\frac{\pi_j\sqrt{r}}{\|\mathbf{u}^j\|_2}\mathbf{I}_r + \mathbf{V}\mathbf{W}\mathbf{V}^\top\right)^{-1}, j = 1, \cdots, m.$$

In the case of $\|\mathbf{u}^j\|_2 = 0$ we have

$$\pi_j\sqrt{r}\mathbf{s}^j - \sum_{i=1}^{N} y_{ji}\mathbf{v}_i^\top \exp\{-\frac{1-\alpha}{2}y_{ji}^2\} = \mathbf{0},$$

or by defining $\mathbf{W}_0$ is a diagonal matrix with main diagonal elements of $w_0(i,i) = \exp\{-\frac{1-\alpha}{2}y_{ji}^2\}$, we have

$$\begin{aligned} & \mathbf{y}^j\mathbf{W}_0\mathbf{V}^\top = \mathbf{s}^j\pi_j\sqrt{r} \\ \Rightarrow \quad & \mathbf{s}^j = \frac{1}{\pi_j\sqrt{r}}\mathbf{y}^j\mathbf{W}_0\mathbf{V}^\top \end{aligned}$$

and if $\frac{1}{\pi_j\sqrt{r}}\|\mathbf{y}^j\mathbf{W}_0\mathbf{V}^\top\|_2 \leq 1$ holds, $\mathbf{u}^j$ should be set to zero. Combining this result with (C.1) gives $\hat{\mathbf{u}}^j$ as

$$\begin{cases} \mathbf{y}^j\mathbf{W}\mathbf{V}^\top\left(\frac{\pi_j\sqrt{r}}{\|\mathbf{u}^j\|_2}\mathbf{I}_r + \mathbf{V}\mathbf{W}\mathbf{V}^\top\right)^{-1}, & \text{if } \|\mathbf{y}^j\mathbf{W}_0\mathbf{V}^\top\|_2 \geq \pi_j\sqrt{r} \\ \mathbf{0}, & \text{otherwise.} \end{cases}$$

Appendix D

Analysis of Estimated V, Ch. 4

Starting from the constrained optimization problem

$$\hat{\mathbf{V}} = \arg\min_{\mathbf{V}} l_\alpha(\mathbf{Y} - \mathbf{UV}) \text{ s.t. } \mathbf{VV}^\top = \mathbf{I}_r, \tag{D.1}$$

we use the Lagrangian function and make this problem an unconstrained problem of

$$\begin{aligned} \hat{\mathbf{V}} &= \arg\min_{\mathbf{V}} L(\mathbf{V}, \mathbf{\Lambda}) \\ &= \arg\min_{\mathbf{V}} l_\alpha(\mathbf{Y} - \mathbf{UV}) - \text{tr}\left(\mathbf{\Lambda}\left(\mathbf{VV}^\top - \mathbf{I}_r\right)\right), \end{aligned}$$

where $\mathbf{\Lambda}$ is the KKT multiplier. Taking the derivative with respect to $\mathbf{V}$ gives

$$\nabla L = \nabla l_\alpha - 2\mathbf{\Lambda V} = \mathbf{0}, \tag{D.2}$$

knowing $\mathbf{\Lambda}^\top = \mathbf{\Lambda}$. To find $\mathbf{\Lambda}$ using above expression we have

$$\begin{aligned} & 2\mathbf{\Lambda V} = \nabla l_\alpha \\ \Rightarrow \quad & 2\mathbf{\Lambda} = \nabla l_\alpha \mathbf{V}^\top \\ \Rightarrow \quad & \mathbf{\Lambda} = \frac{1}{2}\nabla l_\alpha \mathbf{V}^\top = \frac{1}{2}\mathbf{V}\nabla l_\alpha^\top. \end{aligned} \tag{D.3}$$

Plugging (D.3) into (D.2) gives

$$\begin{aligned} & \nabla l_\alpha - \mathbf{V}\nabla l_\alpha^\top \mathbf{V} = \mathbf{0} \\ \Rightarrow \quad & \mathbf{V}\mathbf{V}^\top \nabla l_\alpha - \mathbf{V}\nabla l_\alpha^\top \mathbf{V} = \mathbf{0} \\ \Rightarrow \quad & \mathbf{V}\left(\mathbf{V}^\top \nabla l_\alpha - \nabla l_\alpha^\top \mathbf{V}\right) = \mathbf{0} \\ \Rightarrow \quad & \mathbf{A} = \mathbf{0}, \end{aligned} \tag{D.4}$$

where $\mathbf{A} = \mathbf{V}^\top \nabla l_\alpha - \nabla l_\alpha^\top \mathbf{V}$. Therefore in summary $\nabla L = \mathbf{V}\mathbf{A}$ and the solution only needs to satisfy (D.4). To be able to find the solution of (D.4) we use the following updating rule which was suggested in [160] and is somehow similar to Gradient updating rule with the learning rate η.

$$\mathbf{V}_{t+1} \quad = \quad \mathbf{V}_t - \eta\left(\mathbf{V}_t + \mathbf{V}_{t+1}\right)\mathbf{A}. \tag{D.5}$$

To show that above updating rule guarantees the convergence to a local minimum we need to first prove that at each step the orthogonality constraint is met and also at each step $l_\alpha(\mathbf{V}_{t+1}) \leq l_\alpha(\mathbf{V}_t)$. For the first part we have

$$\begin{aligned} & \mathbf{V}_{t+1} = \mathbf{V}_t - \eta\left(\mathbf{V}_t + \mathbf{V}_{t+1}\right)\mathbf{A} \\ \Rightarrow \quad & \mathbf{V}_{t+1} + \eta\mathbf{V}_t\mathbf{A} = \mathbf{V}_t - \eta\mathbf{V}_t\mathbf{A} \\ \Rightarrow \quad & \mathbf{V}_{t+1}\left(\mathbf{I}_N + \eta\mathbf{A}\right) = \mathbf{V}_t\left(\mathbf{I}_N - \eta\mathbf{A}\right) \\ \Rightarrow \quad & \mathbf{V}_{t+1} = \mathbf{V}_t\left(\mathbf{I}_N - \eta\mathbf{A}\right)\left(\mathbf{I}_N + \eta\mathbf{A}\right)^{-1} \\ \Rightarrow \quad & \mathbf{V}_{t+1} = \mathbf{V}_t\mathbf{Q}, \end{aligned}$$

where $\mathbf{Q} = (\mathbf{I}_N - \eta\mathbf{A})(\mathbf{I}_N + \eta\mathbf{A})^{-1}$. Now the only thing which we should show is that $\mathbf{V}_{t+1}\mathbf{V}_{t+1}^\top = \mathbf{I}_r$. In particular, starting from

$$\mathbf{V}_{t+1}\mathbf{V}_{t+1}^\top = \mathbf{V}_t\mathbf{Q}\mathbf{Q}^\top\mathbf{V}_t^\top,$$

if we can show $\mathbf{Q}\mathbf{Q}^\top = \mathbf{I}_N$, then $\mathbf{V}_{t+1}\mathbf{V}_{t+1}^\top = \mathbf{I}_r$ and the proof is complete. To do so, we have

$$\begin{aligned}
\mathbf{Q}\mathbf{Q}^\top &= (\mathbf{I}_N - \eta\mathbf{A})(\mathbf{I}_N + \eta\mathbf{A})^{-1}{(\mathbf{I}_N + \eta\mathbf{A})^{-1}}^\top(\mathbf{I}_N - \eta\mathbf{A})^\top \\
&= (\mathbf{I}_N - \eta\mathbf{A})\big((\mathbf{I}_N - \eta\mathbf{A})(\mathbf{I}_N + \eta\mathbf{A})\big)^{-1}(\mathbf{I}_N + \eta\mathbf{A}) \\
&= (\mathbf{I}_N - \eta\mathbf{A})\big((\mathbf{I}_N + \eta\mathbf{A})(\mathbf{I}_N - \eta\mathbf{A})\big)^{-1}(\mathbf{I}_N + \eta\mathbf{A}) \\
&= (\mathbf{I}_N - \eta\mathbf{A})(\mathbf{I}_N - \eta\mathbf{A})^{-1}(\mathbf{I}_N + \eta\mathbf{A})^{-1}(\mathbf{I}_N + \eta\mathbf{A}) \\
&= \mathbf{I}_N\mathbf{I}_N = \mathbf{I}_N.
\end{aligned}$$

It means if we start with an orthogonal initial point, at each update the orthogonality constraint is satisfied. For the second part of the proof as a summary we can see that l_α is a function of $\mathbf{V}_{t+1}$ and $\mathbf{V}_{t+1}$ is a function of η. Therefore, l_α is a function of η. Thus, it can be shown that the Gradient of l_α at $\eta = 0$ and at the direction of step length is negative, then there exist a small enough η such that $l_\alpha(\mathbf{V}_{t+1}) \leq l_\alpha(\mathbf{V}_t)$ and the proof is complete. For more details see [160].

Appendix E

Deriving Expression (7.13), Ch. 7

We start with the maximization of denominator of (7.12).

$$\begin{aligned}
\hat{\boldsymbol{\phi}} &= \arg\max_{\boldsymbol{\phi}} f_0\left(\mathbf{y}, \mathbf{n}_1, \ldots, \mathbf{n}_K\right) \\
&= \arg\max_{\boldsymbol{\phi}} \left\{ \frac{1}{\pi^N \det(\mathbf{R}) \sigma^{2N/K+1}} \exp\{-\mathrm{tr}(\mathbf{R}^{-1}\mathbf{T}_0)\} \right\}^{K+1} \\
&= \arg\max_{\boldsymbol{\phi}} \left\{ \exp\{-\mathrm{tr}(\mathbf{R}^{-1}\mathbf{T}_0)\} \right\} \\
&= \arg\min_{\boldsymbol{\phi}} \{\mathrm{tr}(\mathbf{R}^{-1}\mathbf{T}_0)\} \\
&= \arg\min_{\boldsymbol{\phi}} \{\mathrm{tr}(\mathbf{R}^{-1}\{\frac{K}{K+1}\mathbf{S} \\
&\quad + \frac{1}{K+1}\frac{1}{\sigma^2}(\mathbf{y}-\mathbf{B}\boldsymbol{\phi})(\mathbf{y}-\mathbf{B}\boldsymbol{\phi})^\dagger\})\} \\
&= \arg\min_{\boldsymbol{\phi}} \{\mathrm{tr}(\mathbf{R}^{-1}(\mathbf{y}-\mathbf{B}\boldsymbol{\phi})(\mathbf{y}-\mathbf{B}\boldsymbol{\phi})^\dagger)\} \\
&= \arg\min_{\boldsymbol{\phi}} \{\mathrm{tr}((\bar{\mathbf{y}}-\bar{\mathbf{B}}\boldsymbol{\phi})(\bar{\mathbf{y}}-\bar{\mathbf{B}}\boldsymbol{\phi})^\dagger)\} \\
&= \arg\min_{\boldsymbol{\phi}} \{\mathrm{tr}(\bar{\mathbf{y}}\bar{\mathbf{y}}^\dagger - \bar{\mathbf{y}}\boldsymbol{\phi}^\dagger\bar{\mathbf{B}}^\dagger - \bar{\mathbf{B}}\boldsymbol{\phi}\bar{\mathbf{y}}^\dagger + \bar{\mathbf{B}}\boldsymbol{\phi}\boldsymbol{\phi}^\dagger\bar{\mathbf{B}}^\dagger)\},
\end{aligned}$$

where $\bar{\mathbf{B}} = \mathbf{R}^{-\frac{1}{2}}\mathbf{B}$ and $\bar{\mathbf{y}} = \mathbf{R}^{-\frac{1}{2}}\mathbf{y}$. Taking the Gradient with respect to $\boldsymbol{\phi}$ and equalizing to zero gives

$$\begin{aligned}
&\nabla_{\boldsymbol{\phi}}\{\mathrm{tr}(\bar{\mathbf{y}}\bar{\mathbf{y}}^\dagger - \bar{\mathbf{y}}\boldsymbol{\phi}^\dagger\bar{\mathbf{B}}^\dagger - \bar{\mathbf{B}}\boldsymbol{\phi}\bar{\mathbf{y}}^\dagger + \bar{\mathbf{B}}\boldsymbol{\phi}\boldsymbol{\phi}^\dagger\bar{\mathbf{B}}^\dagger)\} \\
&= -2\bar{\mathbf{B}}^\dagger\bar{\mathbf{y}} + 2\bar{\mathbf{B}}^\dagger\bar{\mathbf{B}}\boldsymbol{\phi} = \mathbf{0}, \qquad \text{(E.1)}
\end{aligned}$$

and finally

$$\hat{\boldsymbol{\phi}} = \left(\bar{\mathbf{B}}^{\dagger}\bar{\mathbf{B}}\right)^{-1}\bar{\mathbf{B}}^{\dagger}\bar{\mathbf{y}}. \tag{E.2}$$

Appendix F

The Proof of Proposition 1, Ch. 7

The proof is a modified version of proposition 1 in [178] when the mean is zero and the noise is circularly complex Gaussian. Instead of maximizing the complete likelihood, we factorize the joint likelihood using the conditional likelihoods. Thus, the joint likelihood is given by

$$f\left(\mathbf{Z}\right) = f\left(\mathbf{z}^{N}|\mathbf{Z}^{1:N-1}\right)\cdots f\left(\mathbf{z}^{b+2}|\mathbf{Z}^{1:b+1}\right) f\left(\mathbf{Z}^{1:b+1}\right).$$

For the last term of $f\left(\mathbf{Z}^{1:b+1}\right)$, there is no specific structure in the covariance, then the solution is the well known sample covariance matrix. For other terms, we will maximize each factor individually by considering other observed variables using following relationship

$$\mathbf{z}^{i}|\mathbf{Z}^{1:i-1} \sim \mathcal{CN}(\boldsymbol{\mu}_{i|1:i-1}^{\top}, r_{i|1:i-1}, \mathbf{I}_{K}),$$

where

$$\boldsymbol{\mu}_{i|1:i-1}^{\top} = \tilde{\mathbf{r}}_{1i}^{\dagger}\mathbf{R}_{i-1}^{-1}\mathbf{Z}^{1:i-1}, \tag{F.1}$$

and

$$r_{i|1:i-1} = r_{i,i} - \tilde{\mathbf{r}}_{1i}^{\dagger}\mathbf{R}_{i-1}^{-1}\tilde{\mathbf{r}}_{1i}. \tag{F.2}$$

Using (F.1) we have

$$\begin{aligned}
\boldsymbol{\mu}_{i|1:i-1} &= (\tilde{\mathbf{r}}_{1i}^{\dagger}\mathbf{R}_{i-1}^{-1}\mathbf{Z}^{1:i-1})^{\top} \\
&= \sum_{j=i-b}^{i-1} r_{i,j}\frac{\det(\mathbf{R}_{i-2})}{\det(\mathbf{R}_{i-1})}\sum_{n=1}^{i-1}(-1)^{j+n}\frac{\det(\mathbf{M}_{i-1}^{nj})}{\det(\mathbf{R}_{i-2})}\mathbf{z}^{n\top} \\
&= \sum_{j=i-b}^{i-1}\hat{\lambda}_{i,j}\hat{\mathbf{z}}_{i-1,j} \\
&= \hat{\mathbf{F}}_{i-1}\hat{\boldsymbol{\lambda}}_i \qquad \text{(F.3)}
\end{aligned}$$

where $\hat{\boldsymbol{\lambda}}_i = (\hat{\lambda}_{i,i-b}, \cdots, \hat{\lambda}_{i,i-1})^{\top}$ and $\hat{\mathbf{F}}_{i-1} = (\hat{\mathbf{z}}_{i-1,i-b}, \cdots, \hat{\mathbf{z}}_{i-1,i-1})$, while $\hat{\lambda}_{i,j} = r_{i,j}\frac{\det(\mathbf{R}_{i-2})}{\det(\mathbf{R}_{i-1})}$ and $\hat{\mathbf{z}}_{i-1,m} = \sum_{n=1}^{i-1}(-1)^{m+n}\frac{\det(\mathbf{M}_{i-1}^{nm})}{\det(\mathbf{R}_{i-2})}\mathbf{z}^{n\top}$. $\mathbf{M}_{i-1}^{nm}$ is the matrix which is obtained when the $n-$th row and $m-$column have been removed from $\mathbf{R}_{i-1}$. The prposed estimator for $\boldsymbol{\lambda}_i$ is

$$\hat{\boldsymbol{\lambda}}_i = \left(\hat{\mathbf{F}}_{i-1}^{\dagger}\hat{\mathbf{F}}_{i-1}\right)^{-1}\hat{\mathbf{F}}_{i-1}^{\dagger}\mathbf{z}^{i\top}.$$

For the conditional variance we have

$$\begin{aligned}
\hat{r}_{i|1:i-1} &= \frac{1}{K}\left(\mathbf{z}^{i\top} - \hat{\boldsymbol{\mu}}_{i|1:i-1}\right)^{\dagger}\left(\mathbf{z}^{i\top} - \hat{\boldsymbol{\mu}}_{i|1:i-1}\right) \\
&= \frac{1}{K}\mathbf{z}^{i*}\left(\mathbf{I}_K - \mathbf{F}_{i-1}\left(\hat{\mathbf{F}}_{i-1}^{\dagger}\hat{\mathbf{F}}_{i-1}\right)^{-1}\hat{\mathbf{F}}_{i-1}^{\dagger}\right)\mathbf{z}^{i\top}.
\end{aligned}$$

Therefore, using (F.2) we have

$$\begin{aligned}
r_{i,i} &= \frac{1}{K}\mathbf{z}^{i*}\left(\mathbf{I}_K - \mathbf{F}_{i-1}\left(\hat{\mathbf{F}}_{i-1}^{\dagger}\hat{\mathbf{F}}_{i-1}\right)^{-1}\hat{\mathbf{F}}_{i-1}^{\dagger}\right)\mathbf{z}^{i\top} \\
&+ \tilde{\mathbf{r}}_{1i}^{\dagger}\mathbf{R}_{i-1}^{-1}\tilde{\mathbf{r}}_{1i}.
\end{aligned}$$

Adopting the same steps as [178], and using induction it can be shown that above covariance estimator is consistent which verifies figure 7.3. As a summary we know that in the first step we have the maximum likelihood estimation of $\mathbf{R}_{b+1}^{b}$ which is a consistent estimator. Then for coming observation, it can be shown that the new covariance estimator based on previous consistent estimator will be again consistent. Finally at the end, the full covariance estimator will remain consistent since we are using a sequence of consistent estimators.

Bibliography

[1] Aref Miri Rekavandi, Abd-Krim Seghouane, and Robin J Evans. Robust subspace detectors based on α-divergence with application to detection in imaging. *IEEE Transactions on Image Processing*, 30:5017–5031, 2021.

[2] Aref Miri Rekavandi, Abd-Krim Seghouane, and Robin J Evans. Robust likelihood ratio test using α- divergence. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1150–1154. IEEE, 2020.

[3] Aref Miri Rekavandi, Abd-Krim Seghouane, and Robin J Evans. Learning robust and sparse principal components with the α-divergence. *To be Submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[4] Aref Miri Rekavandi and Abd-Krim Seghouane. Robust principal component analysis using Alpha divergence. In *ICIP 2020-2020 IEEE International Conference on Image Processing (ICIP)*, pages 1–5. IEEE, 2020.

[5] Aref Miri Rekavandi, Abd-Krim Seghouane, and Robin J Evans. Adaptive matched filter using non-target free training data. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1090–1094. IEEE, 2020.

[6] Aref Miri Rekavandi, Abd-Krim Seghouane, and Robin J Evans. Adaptive subspace detector in high dimensional space with insufficient training data. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1164–1168. IEEE, 2019.

[7] Aref Miri Rekavandi, Abd-Krim Seghouane, and Robin J Evans. Adaptive brain activity detection in structured interference and partially homogeneous locally

correlated disturbance. *To be Submitted to IEEE Transactions on Medical Imaging*, 2021.

[8] Martin A Lindquist et al. The statistical analysis of fMRI data. *Statistical Science*, 23(4):439–464, 2008.

[9] Abdelhak M Zoubir, Visa Koivunen, Esa Ollila, and Michael Muma. *Robust statistics for signal processing*. Cambridge University Press, 2018.

[10] Stephane Heritier, Eva Cantoni, Samuel Copt, and Maria-Pia Victoria-Feser. *Robust methods in biostatistics*, volume 825. John Wiley & Sons, 2009.

[11] Wai-Sheou Chen and Irving S Reed. A new CFAR detection test for radar. *Digital Signal Processing*, 1(4):198–214, 1991.

[12] Heesung Kwon and Nasser M Nasrabadi. Kernel matched subspace detectors for hyperspectral target detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(2):178–194, 2005.

[13] Jiaqiu Ai, Xiangyang Qi, Weidong Yu, Yunkai Deng, Fan Liu, and Li Shi. A new CFAR ship detection algorithm based on 2-D joint log-normal distribution in SAR images. *IEEE Geoscience and Remote Sensing Letters*, 7(4):806–810, 2010.

[14] Mohammad Taghi Dabiri, Seyed Mohammad Sajad Sadough, and Hossein Safi. GLRT-based sequence detection of OOK modulation over FSO turbulence channels. *IEEE Photonics Technology Letters*, 29(17):1494–1497, 2017.

[15] Daniel J Ryan, Iain B Collings, and I Vaughan L Clarkson. GLRT-optimal noncoherent lattice decoding. *IEEE Transactions on Signal Processing*, 55(7):3773–3786, 2007.

[16] Abd-Krim Seghouane and Ju Lynn Ong. A Bayesian model selection approach to fMRI activation detection. In *2010 IEEE International Conference on Image Processing*, pages 4401–4404. IEEE, 2010.

[17] Abd-Krim Seghouane. fMRI activation detection using a variant of Akaike information criterion. In *2012 IEEE Statistical Signal Processing Workshop (SSP)*, pages 233–236. IEEE, 2012.

[18] Gudrun Lange, Jason Steffener, Dane B Cook, Benjamin Martin Bly, Christopher Christodoulou, W-C Liu, J Deluca, and BH Natelson. Objective evidence of cognitive complaints in chronic fatigue syndrome: a BOLD fMRI study of verbal working memory. *Neuroimage*, 26(2):513–524, 2005.

[19] Seiji Ogawa, David W Tank, Ravi Menon, Jutta M Ellermann, Seong G Kim, Helmut Merkle, and Kamil Ugurbil. Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. *Proceedings of the National Academy of Sciences*, 89(13):5951–5955, 1992.

[20] Eugene Lin and Adam Alessio. What are the basic concepts of temporal, contrast, and spatial resolution in cardiac CT? *Journal of Cardiovascular Computed Tomography*, 3(6):403–408, 2009.

[21] Jeff H Duyn, Yihong Yang, Joseph A Frank, Venkata S Mattay, and Lei Hou. Functional magnetic resonance neuroimaging data acquisition techniques. *Neuroimage*, 4(3):S76–S83, 1996.

[22] Karl J Friston, P Fletcher, Oliver Josephs, ANDREW Holmes, MD Rugg, and Robert Turner. Event-related fMRI: characterizing differential responses. *Neuroimage*, 7(1):30–40, 1998.

[23] Martijn P Van Den Heuvel and Hilleke E Hulshoff Pol. Exploring the brain network: a review on resting-state fMRI functional connectivity. *European Neuropsychopharmacology*, 20(8):519–534, 2010.

[24] Geoffrey Karl Aguirre, Eric Zarahn, and Mark D'Esposito. The variability of human, BOLD hemodynamic responses. *Neuroimage*, 8(4):360–369, 1998.

[25] Geoffrey M Boynton, Stephen A Engel, Gary H Glover, and David J Heeger. Linear systems analysis of functional magnetic resonance imaging in human V1. *Journal of Neuroscience*, 16(13):4207–4221, 1996.

[26] Shang-Hong Lai and Ming Fang. A novel local PCA-based method for detecting activation signals in fMRI. *Magnetic Resonance Imaging*, 17(6):827–836, 1999.

[27] Martin J McKeown and Terrence J Sejnowski. Independent component analysis of fMRI data: examining the assumptions. *Human Brain Mapping*, 6(5-6):368–372, 1998.

[28] Kangjoo Lee, Sungho Tak, and Jong Chul Ye. A data-driven sparse GLM for fMRI analysis using sparse dictionary learning with MDL criterion. *IEEE Transactions on Medical Imaging*, 30(5):1076–1089, 2010.

[29] Abd-Krim Seghouane and Asif Iqbal. Sequential dictionary learning from correlated data: Application to fMRI data analysis. *IEEE Transactions on Image Processing*, 26(6):3002–3015, 2017.

[30] Ronald Sladky, Karl J Friston, Jasmin Tröstl, Ross Cunnington, Ewald Moser, and Christian Windischberger. Slice-timing effects and their correction in functional MRI. *Neuroimage*, 58(2):588–594, 2011.

[31] Maxim Zaitsev, Burak Akin, Pierre LeVan, and Benjamin R Knowles. Prospective motion correction in functional MRI. *Neuroimage*, 154:33–42, 2017.

[32] Abd-Krim Seghouane and Asif Iqbal. Basis expansion approaches for regularized sequential dictionary learning algorithms with enforced sparsity for fMRI data analysis. *IEEE Transactions on Medical Imaging*, 36(9):1796–1807, 2017.

[33] Jean-Baptiste Poline and Matthew Brett. The general linear model and fMRI: does love last forever? *Neuroimage*, 62(2):871–880, 2012.

[34] Karl J Friston, Andrew P Holmes, JB Poline, PJ Grasby, SCR Williams, Richard SJ Frackowiak, and Robert Turner. Analysis of fMRI time-series revisited. *Neuroimage*, 2(1):45–53, 1995.

[35] Joonki Noh and Victor Solo. Rician distributed fMRI: asymptotic power analysis and Cramer–Rao lower bounds. *IEEE Transactions on Signal Processing*, 59(3): 1322–1328, 2010.

[36] Steven M Kay. *Detection theory*. Prentice Hall PTR, 2009.

[37] Gökhan Gül. *Robust and distributed hypothesis testing*. Springer, 2017.

[38] Jerzy Neyman and Egon Sharpe Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933.

[39] Bernard C Levy. *Principles of signal detection and parameter estimation*. Springer Science & Business Media, 2008.

[40] Abdelhak M Zoubir, Visa Koivunen, Yacine Chakhchoukh, and Michael Muma. Robust estimation in signal processing: A tutorial-style treatment of fundamental concepts. *IEEE Signal Processing Magazine*, 29(4):61–80, 2012.

[41] Sidney Siegel. Nonparametric statistics. *The American Statistician*, 11(3):13–19, 1957.

[42] RA Maronna, RD Martin, and VJ Yohai. Robust statistics theory and methods john wiley & sons. *Inc., USA*, 2006.

[43] Peter J Huber. Robust estimation of a location parameter. pages 73–101. 1964.

[44] Peter J Huber. *Robust statistical procedures*. SIAM, 1996.

[45] Peter J Huber. A robust version of the probability ratio test. *The Annals of Mathematical Statistics*, pages 1753–1758, 1965.

[46] H Vincent Poor. *An introduction to signal detection and estimation*. Springer Science & Business Media, 2013.

[47] Louis L Scharf. *Statistical signal processing*, volume 98. Addison-Wesley Reading, MA, 1991.

[48] John CW Rayner, Olivier Thas, and Donald John Best. *Smooth tests of goodness of fit: using R*. John Wiley & Sons, 2009.

[49] Louis L Scharf and Benjamin Friedlander. Matched subspace detectors. *IEEE Transactions on Signal Processing*, 42(8):2146–2157, 1994.

[50] Edward J Kelly. An adaptive detection algorithm. *IEEE Transactions on Aerospace and Electronic Systems*, (2):115–127, 1986.

[51] Shawn Kraut, Louis L Scharf, and L Todd McWhorter. Adaptive subspace detectors. *IEEE Transactions on Signal Processing*, 49(1):1–16, 2001.

[52] Daniel R Fuhrmann, Edward J Kelly, and Ramon Nitzberg. A CFAR adaptive matched filter detector. *IEEE Transactions on Aerospace and Electronic Systems*, 28(1):208–216, 1992.

[53] Zheran Shang, Kai Huo, Weijian Liu, Yongliang Wang, and Xiang Li. Interference environment model recognition for robust adaptive detection. *IEEE Transactions on Aerospace and Electronic Systems*, 2019.

[54] DR Cox and DV Hinkley. Theoretical statistics chapman and hall, london. *See Also*, 1974.

[55] Antonio De Maio. A new derivation of the adaptive matched filter. *IEEE Signal Processing Letters*, 11(10):792–793, 2004.

[56] Masoud Naderpour, Ali Ghobadzadeh, Aliakbar Tadaion, and Saeed Gazor. Generalized wald test for binary composite hypothesis test. *IEEE Signal Processing Letters*, 22(12):2239–2243, 2015.

[57] Jun Liu, Shenghua Zhou, Weijian Liu, Jibin Zheng, Hongwei Liu, and Jian Li. Tunable adaptive detection in colocated MIMO radar. *IEEE Transactions on Signal Processing*, 66(4):1080–1092, 2017.

[58] Antonio De Maio. Rao test for adaptive detection in Gaussian interference with unknown covariance matrix. *IEEE Transactions on Signal Processing*, 55(7):3577–3584, 2007.

[59] Jun Liu, Weijian Liu, Bo Chen, Hongwei Liu, Hongbin Li, and Chengpeng Hao. Modified Rao test for multichannel adaptive signal detection. *IEEE Transactions on Signal Processing*, 64(3):714–725, 2015.

[60] Steven M Kay. *Fundamentals of statistical signal processing*. Prentice Hall PTR, 1993.

[61] Linlin Mao, Yongchan Gao, Shefeng Yan, and Lijun Xu. Persymmetric subspace detection in structured interference and non-homogeneous disturbance. *IEEE Signal Processing Letters*, 26(6):928–932, 2019.

[62] Satyabrata Sen. Low-rank matrix decomposition and spatio-temporal sparse recovery for STAP radar. *IEEE Journal of Selected Topics in Signal Processing*, 9 (8):1510–1523, 2015.

[63] Kiseon Kim and Georgy Shevlyakov. Why Gaussianity? *IEEE Signal Processing Magazine*, 25(2):102–113, 2008.

[64] David Middleton. Non-Gaussian noise models in signal processing for telecommunications: new methods an results for class a and class b noise models. *IEEE Transactions on Information Theory*, 45(4):1129–1149, 1999.

[65] Yuri I Abramovich and Pavel Turcaj. Impulsive noise mitigation in spatial and temporal domains for surface-wave over-the-horizon radar. Technical report, Cooperative Research Centre for Sensor Signal and Information Processing, 1999.

[66] Paul C Etter. *Underwater acoustic modeling and simulation*. CRC Press, 2018.

[67] T Keith Blankenship, DM Kriztman, and Theodore S Rappaport. Measurements and simulation of radio frequency impulsive noise in hospitals and clinics. In *1997 IEEE 47th Vehicular Technology Conference. Technology in Motion*, volume 3, pages 1942–1946. IEEE, 1997.

[68] DC Alexander, GJ Barker, and SR Arridge. Detection and modeling of non-Gaussian apparent diffusion coefficient profiles in human brain data. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 48(2):331–340, 2002.

[69] Mukund N Desai and Rami S Mangoubi. Robust Gaussian and non-Gaussian matched subspace detection. *IEEE Transactions on Signal Processing*, 51(12): 3115–3127, 2003.

[70] J v Neumann. Zur theorie der gesellschaftsspiele. *Mathematische annalen*, 100(1): 295–320, 1928.

[71] Max Shiffman. On the equality min max= max min, and the theory of games. 1949.

[72] Peter J Huber and Volker Strassen. Minimax tests and the Neyman-Pearson lemma for capacities. *The Annals of Statistics*, pages 251–263, 1973.

[73] R Martin and S Schwartz. Robust detection of a known signal in nearly Gaussian noise. *IEEE Transactions on Information Theory*, 17(1):50–56, 1971.

[74] R Martin and C McGath. Robust detection to stochastic signals (Corresp.). *IEEE Transactions on Information Theory*, 20(4):537–541, 1974.

[75] S Kassam and J Thomas. Asymptotically robust detection of a known signal in contaminated non-Gaussian noise. *IEEE Transactions on Information Theory*, 22 (1):22–26, 1976.

[76] A El-Sawy and V VandeLinde. Robust detection of known signals. *IEEE Transactions on Information Theory*, 23(6):722–727, 1977.

[77] S Kassam, George Moustakides, and Jung Shin. Robust detection of known signals in asymmetric noise. *IEEE Transactions on Information Theory*, 28(1):84–91, 1982.

[78] Bernard C Levy. Robust hypothesis testing with a relative entropy tolerance. *IEEE Transactions on Information Theory*, 55(1):413–421, 2008.

[79] Gökhan Gül and Abdelhak M Zoubir. Robust hypothesis testing with α -divergence. *IEEE Transactions on Signal Processing*, 64(18):4737–4750, 2016.

[80] Gökhan Gül. Asymptotically minimax robust hypothesis testing. *arXiv preprint arXiv:1711.07680*, 2017.

[81] S Kassam. Robust hypothesis testing for bounded classes of probability densities (corresp.). *IEEE Transactions on Information Theory*, 27(2):242–247, 1981.

[82] Michael Fauß and Abdelhak M Zoubir. Old bands, new tracks—revisiting the band model for robust hypothesis testing. *IEEE Transactions on Signal Processing*, 64 (22):5875–5886, 2016.

[83] Michael Fauß, Abdelhak M Zoubir, and H Vincent Poor. On the equivalence of f-divergence balls and density bands in robust detection. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4399–4403. IEEE, 2018.

[84] Yingxi Liu and Ahmed Tewfik. Empirical likelihood ratio test with distribution function constraints. *IEEE Transactions on Signal Processing*, 61(18):4463–4472, 2013.

[85] John W Tukey. A survey of sampling from contaminated distributions. *Contributions to Probability and Statistics*, pages 448–485, 1960.

[86] Frank R Hampel. Contribution to the theory of robust estimation. *Ph. D. Thesis, University of California, Berkeley*, 1968.

[87] Frédéric Pascal, Lionel Bombrun, Jean-Yves Tourneret, and Yannick Berthoumieu. Parameter estimation for multivariate generalized Gaussian distributions. *IEEE Transactions on Signal Processing*, 61(23):5960–5971, 2013.

[88] Ebrahim Baktash, Mahmood Karimi, Zhenliang Zhang, and Xiaodong Wang. Detection of unknown signals under complex elliptically symmetric distributions. *IEEE Transactions on Communications*, 65(4):1662–1674, 2017.

[89] Peter J Huber and Elvezio M Ronchetti. Robust statistics. 2009. *Hoboken: John Wiley & Sons*, 2.

[90] Leonard A Stefanski and Dennis D Boos. The calculus of M-estimation. *The American Statistician*, 56(1):29–38, 2002.

[91] Albert E Beaton and John W Tukey. The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, 16(2):147–185, 1974.

[92] Michael L McCloud and Louis L Scharf. Interference estimation with applications to blind multiple-access communication over fading channels. *IEEE Transactions on Information Theory*, 46(3):947–961, 2000.

[93] Ian L Dryden. Shape analysis. *Wiley StatsRef: Statistics Reference Online*, 2014.

[94] E OJE. Subspace methods of pattern recognition. In *Pattern Recognition and Image Processing Series*, volume 6. Research Studies Press, 1983.

[95] Sandip Bose and Allan O Steinhardt. A maximal invariant framework for adaptive detection with structured and unstructured covariance matrices. *IEEE Transactions on Signal Processing*, 43(9):2164–2175, 1995.

[96] Pu Wang, Jun Fang, Hongbin Li, and Braham Himed. Detection with target-induced subspace interference. *IEEE Signal Processing Letters*, 19(7):403–406, 2012.

[97] Jun Liu, Weijian Liu, Yongchan Gao, Shenghua Zhou, and Xiang-Gen Xia. Persymmetric adaptive detection of subspace signals: Algorithms and performance analysis. *IEEE Transactions on Signal Processing*, 66(23):6124–6136, 2018.

[98] Domenico Ciuonzo, Antonio De Maio, and Danilo Orlando. A unifying framework for adaptive radar detection in homogeneous plus structured interference—part ii: Detectors design. *IEEE Transactions on Signal Processing*, 64(11):2907–2919, 2016.

[99] Weijian Liu, Jun Liu, Hai Li, Qinglei Du, and Yong-Liang Wang. Multichannel signal detection based on Wald test in subspace interference and Gaussian noise. *IEEE Transactions on Aerospace and Electronic Systems*, 55(3):1370–1381, 2018.

[100] Domenico Ciuonzo, Antonio De Maio, and Danilo Orlando. On the statistical invariance for adaptive radar detection in partially homogeneous disturbance plus structured interference. *IEEE Transactions on Signal Processing*, 65(5):1222–1234, 2016.

[101] Zeyu Wang, Ming Li, Hongmeng Chen, Lei Zuo, Peng Zhang, and Yan Wu. Adaptive detection of a subspace signal in signal-dependent interference. *IEEE Transactions on Signal Processing*, 65(18):4812–4820, 2017.

[102] Andrzej Cichocki and Shun-ichi Amari. Families of Alpha-Beta-and Gamma-divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568, 2010.

[103] Narayanaswamy Balakrishnan, Elena Castilla, Nirian Martín, and Leandro Pardo. Robust estimators and test statistics for one-shot device testing under the exponential distribution. *IEEE Transactions on Information Theory*, 65(5):3080–3096, 2019.

[104] Ayanendranath Basu, Abhijit Mandal, Nirian Martín, and Leandro Pardo. A robust Wald-type test for testing the equality of two means from log-normal samples. *Methodology and Computing in Applied Probability*, 21(1):85–107, 2019.

[105] Nirian Martin, Raquel Mata, and Leandro Pardo. Wald type and Phi-divergence based test-statistics for isotonic binomial proportions. *Mathematics and Computers in Simulation*, 120:31–49, 2016.

[106] Abhik Ghosh, Nirian Martin, Ayanendranath Basu, and Leandro Pardo. A new class of robust two-sample Wald-type tests. *The International Journal of Biostatistics*, 14(2), 2018.

[107] Ayanendranath Basu, Abhik Ghosh, Abhijit Mandal, Nirian Martin, Leandro Pardo, et al. A Wald-type test statistic for testing linear hypothesis in logistic regression models based on minimum density power divergence estimator. *Electronic Journal of Statistics*, 11(2):2741–2772, 2017.

[108] Arnt-Brre Salberg, Alfred Hanssen, and Louis L Scharf. Robust multidimensional matched subspace classifiers based on weighted least-squares. *IEEE Transactions on Signal Processing*, 55(3):873–880, 2007.

[109] Shun-ichi Amari. *Differential-geometrical methods in statistics*, volume 28. Springer Science & Business Media, 2012.

[110] Shun-Ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2007.

[111] Shun-Ichi Amari. α-divergence is unique, belonging to both f-divergence and Bregman divergence classes. *IEEE Transactions on Information Theory*, 55(11): 4925–4931, 2009.

[112] Asif Iqbal and Abd-Krim Seghouane. An α-divergence-based approach for robust dictionary learning. *IEEE Transactions on Image Processing*, 28(11):5729–5739, 2019.

[113] Abd-Krim Seghouane and Davide Ferrari. Robust hemodynamic response function estimation from fNIRS signals. *IEEE Transactions on Signal Processing*, 67(7): 1838–1848, 2019.

[114] Frank R Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974.

[115] Eva Cantoni and Elvezio Ronchetti. Robust inference for generalized linear models. *Journal of the American Statistical Association*, 96(455):1022–1030, 2001.

[116] Peter J Huber et al. Robust regression: asymptotics, conjectures and monte Carlo. *Annals of Statistics*, 1(5):799–821, 1973.

[117] Babak A Ardekani, Jeff Kershaw, Kenichi Kashikura, and Iwao Kanno. Activation detection in functional MRI using subspace modeling and maximum likelihood estimation. *IEEE Transactions on Medical Imaging*, 18(2):101–114, 1999.

[118] Vince Daniel Calhoun, T Adali, GD Pearlson, and James J Pekar. Spatial and temporal independent component analysis of functional MRI data containing a pair of task-related waveforms. *Human Brain Mapping*, 13(1):43–53, 2001.

[119] Deanna M Barch, Gregory C Burgess, Michael P Harms, Steven E Petersen, Bradley L Schlaggar, Maurizio Corbetta, Matthew F Glasser, Sandra Curtiss, Sachin Dixit, Cindy Feldt, et al. Function in the human connectome: task-fMRI and individual differences in behavior. *Neuroimage*, 80:169–189, 2013.

[120] Pasquale Iervolino and Raffaella Guida. A novel ship detector based on the generalized-likelihood ratio test for SAR imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(8):3616–3630, 2017.

[121] Gan Rongbing and Wang Jianguo. Distribution-based CFAR detectors in SAR images. *Journal of Systems Engineering and Electronics*, 17(4):717–721, 2006.

[122] Gui Gao, Kewei Ouyang, Yongbo Luo, Sheng Liang, and Shilin Zhou. Scheme of parameter estimation for generalized Gamma distribution and its application to ship detection in SAR images. *IEEE Transactions on Geoscience and Remote Sensing*, 55(3):1812–1832, 2016.

[123] Yi Cui, Jian Yang, Yoshio Yamaguchi, Gulab Singh, Sang-Eun Park, and Hirokazu Kobayashi. On semiparametric clutter estimation for ship detection in synthetic aperture radar images. *IEEE Transactions on Geoscience and Remote Sensing*, 51(5):3170–3180, 2012.

[124] Yehuda Koren and Liran Carmel. Visualization of labeled data using linear transformations. In *IEEE Symposium on Information Visualization 2003 (IEEE Cat. No. 03TH8714)*, pages 121–128. IEEE, 2003.

[125] Ian Jolliffe. *Principal component analysis*. Springer, 2011.

[126] Thierry Bouwmans, Sajid Javed, Hongyang Zhang, Zhouchen Lin, and Ricardo Otazo. On the applications of robust PCA in image and video processing. *Proceedings of the IEEE*, 106(8):1427–1457, 2018.

[127] Thomas P Minka. Automatic choice of dimensionality for PCA. In *Advances in Neural Information Processing Systems*, pages 598–604, 2001.

[128] Abd-Krim Seghouane and Andrzej Cichocki. Bayesian estimation of the number of principal components. *Signal Processing*, 87(3):562–568, 2007.

[129] Abd-Krim Seghouane, Navid Shokouhi, and Inge Koch. Sparse principal component analysis with preserved sparsity pattern. *IEEE Transactions on Image Processing*, 28(7):3274–3285, 2019.

[130] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.

[131] Lei Xu and Alan L Yuille. Robust principal component analysis by self-organizing rules based on statistical physics approach. *IEEE Transactions on Neural Networks*, 6(1):131–143, 1995.

[132] Geoffrey E Hinton, Peter Dayan, and Michael Revow. Modeling the manifolds of images of handwritten digits. *IEEE Transactions on Neural Networks*, 8(1):65–74, 1997.

[133] Michael E Tipping and Christopher M Bishop. Mixtures of probabilistic principal component analyzers. *Neural Computation*, 11(2):443–482, 1999.

[134] Ralf Möller and Heiko Hoffmann. An extension of neural gas to local PCA. *Neurocomputing*, 62:305–326, 2004.

[135] Venkat Chandrasekaran, Sujay Sanghavi, Pablo A Parrilo, and Alan S Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

[136] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[137] Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Robust PCA via outlier pursuit. In *Advances in Neural Information Processing Systems*, pages 2496–2504, 2010.

[138] Praneeth Netrapalli, UN Niranjan, Sujay Sanghavi, Animashree Anandkumar, and Prateek Jain. Non-convex robust PCA. In *Advances in Neural Information Processing Systems*, pages 1107–1115, 2014.

[139] Xinyang Yi, Dohyung Park, Yudong Chen, and Constantine Caramanis. Fast algorithms for robust PCA via Gradient descent. In *Advances in neural information processing systems*, pages 4152–4160, 2016.

[140] Alain Baccini, Ph Besse, and AD Falguerolles. A L1-norm PCA and a heuristic approach. *Ordinal and symbolic data analysis*, 1(1):359–368, 1996.

[141] Qifa Ke and Takeo Kanade. Robust $l_1$norm factorization in the presence of outliers and missing data by alternative convex programming. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 739–746. IEEE, 2005.

[142] Hojjat Akhondi-Asl and James DB Nelson. M-estimate robust PCA for seismic noise attenuation. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 1853–1857. IEEE, 2016.

[143] Norm A Campbell. Robust procedures in multivariate analysis i: Robust covariance estimation. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 29(3):231–237, 1980.

[144] Md Nurul Haque Mollah, Nayeema Sultana, Mihoko Minami, and Shinto Eguchi. Robust extraction of local structures by the minimum β-divergence method. *Neural Networks*, 23(2):226–238, 2010.

[145] Peter J Huber. *Robust statistics*. New York, USA, Wiley, 1981.

[146] Mostafa Rahmani and George K Atia. Coherence pursuit: Fast, simple, and robust principal component analysis. *IEEE Transactions on Signal Processing*, 65(23): 6260–6275, 2017.

[147] Vishnu Menon and Sheetal Kalyani. Structured and unstructured outlier identification for robust PCA: A fast parameter free algorithm. *IEEE Transactions on Signal Processing*, 67(9):2439–2452, 2019.

[148] Zhihui Lai, Yong Xu, Qingcai Chen, Jian Yang, and David Zhang. Multilinear sparse principal component analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 25(10):1942–1950, 2014.

[149] Daniela M Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 2009.

[150] Zihan Zhou, Xiaodong Li, John Wright, Emmanuel Candes, and Yi Ma. Stable principal component pursuit. In *2010 IEEE international symposium on information theory*, pages 1518–1522. IEEE, 2010.

[151] Pratik Prabhanjan Brahma, Yiyuan She, Shijie Li, Jiade Li, and Dapeng Wu. Reinforced robust principal component pursuit. *IEEE Transactions on Neural Networks and Learning Systems*, 29(5):1525–1538, 2017.

[152] Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[153] DN Lawley. A modified method of estimation in factor analysis and some large sample results. In *Uppsala Symposium on Psychological Factor Analysis*, volume 17, pages 35–42. Taylor & Francis, 1953.

[154] Christian D Sigg and Joachim M Buhmann. Expectation-maximization for sparse and non-negative PCA. In *Proceedings of the 25th International Conference on Machine Learning*, pages 960–967. ACM, 2008.

[155] Alexander Ilin and Tapani Raiko. Practical approaches to principal component analysis in the presence of missing values. *Journal of Machine Learning Research*, 11(Jul):1957–2000, 2010.

[156] Michael P Windham. Robustifying model fitting. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 599–609, 1995.

[157] Edwin Choi, Peter Hall, and Brett Presnell. Rendering parametric procedures more robust by empirically tilting the model. *Biometrika*, 87(2):453–465, 2000.

[158] Davide Ferrari and Davide La Vecchia. On robust estimation via pseudo-additive information. *Biometrika*, 99(1):238–244, 2011.

[159] Jonathan T Barron. A general and adaptive robust loss function. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4331–4339, 2019.

[160] Zaiwen Wen and Wotao Yin. A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142(1-2):397–434, 2013.

[161] Asif Iqbal and Abd-Krim Seghouane. An algorithm for multi subject fMRI analysis based on the SVD and penalized rank-1 matrix approximation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2721–2725. IEEE, 2018.

[162] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

[163] Tianyi Zhou and Dacheng Tao. Godec: Randomized low-rank & sparse matrix decomposition in noisy case. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*, 2011.

[164] Liyuan Li, Weimin Huang, Irene Yu-Hua Gu, and Qi Tian. Statistical modeling of complex backgrounds for foreground object detection. *IEEE Transactions on Image Processing*, 13(11):1459–1472, 2004.

[165] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):569–582, 2014.

[166] Adnan Shah and Abd-Krim Seghouane. An integrated framework for joint HRF and drift estimation and Hbo/Hbr signal improvement in fNIRS data. *IEEE Transactions on Medical Imaging*, 33(11):2086–2097, 2014.

[167] Nicholas B Pulsone and Charles M Rader. Adaptive beamformer orthogonal rejection test. *IEEE Transactions on Signal Processing*, 49(3):521–529, 2001.

[168] Jun Liu, Zi-Jing Zhang, and Yun Yang. Optimal waveform design for generalized likelihood ratio and adaptive matched filter detectors using a diversely polarized antenna. *Signal Processing*, 92(4):1126–1131, 2012.

[169] Peter J Huber. *Robust statistics*, volume 523. John Wiley & Sons, 2004.

[170] Ali Ghobadzadeh, Saeed Gazor, Masoud Naderpour, and Ali Akbar Tadaion. Asymptotically optimal CFAR detectors. *IEEE Transactions on Signal Processing*, 64(4):897–909, 2015.

[171] Abd-Krim Seghouane. Asymptotic bootstrap corrections of AIC for linear regression models. *signal Processing*, 90(1):217–224, 2010.

[172] Abd-Krim Seghouane. New AIC corrected variants for multivariate linear regression model selection. *IEEE Transactions on Aerospace and Electronic Systems*, 47 (2):1154–1165, 2011.

[173] Iain M Johnstone and Arthur Yu Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693, 2009.

[174] Shawn Kraut and Louis L Scharf. The cfar adaptive subspace detector is a scale-invariant glrt. *IEEE Transactions on Signal Processing*, 47(9):2538–2541, 1999.

[175] Guilhem Pailloux, Philippe Forster, Jean-Philippe Ovarlez, and Frederic Pascal. Persymmetric adaptive radar detectors. *IEEE Transactions on Aerospace and Electronic Systems*, 47(4):2376–2390, 2011.

[176] Sergiy A Vorobyov, Alex B Gershman, and Kon Max Wong. Maximum likelihood direction-of-arrival estimation in unknown noise fields using sparse sensor arrays. *IEEE Transactions on Signal Processing*, 53(1):34–43, 2004.

[177] Daniel R Fuhrmann. Application of Toeplitz covariance estimation to adaptive beamforming and detection. *IEEE Transactions on Signal Processing*, 39(10): 2194–2198, 1991.

[178] Martin Ohlson, Zhanna Andrushchenko, and Dietrich von Rosen. Explicit estimators under m-dependence for a multivariate normal distribution. *Annals of the Institute of Statistical Mathematics*, 63(1):29–42, 2011.

[179] Jun Liu, Zi-Jing Zhang, Peng-Lang Shui, and Hongwei Liu. Exact performance analysis of an adaptive subspace detector. *IEEE Transactions on Signal Processing*, 60(9):4945–4950, 2012.

[180] Muhammad Ali Qadar and Abd-Krim Seghouane. A projection CCA method for effective fMRI data analysis. *IEEE Transactions on Biomedical Engineering*, 66 (11):3247–3256, 2019.

[181] Asif Iqbal, Abd-Krim Seghouane, and Tülay Adalı. Shared and subject-specific dictionary learning (shSSDL) algorithm for multisubject fMRI data analysis. *IEEE Transactions on Biomedical Engineering*, 65(11):2519–2528, 2018.

[182] Stéphane Heritier and Elvezio Ronchetti. Robust bounded-influence tests in general parametric models. *Journal of the American Statistical Association*, 89(427): 897–904, 1994.