ANOMALY-PRESERVING REDUNDANCY-REDUCTION IN HIGH-DIMENSIONAL SIGNALS

RESEARCH THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

OLEG KUYBEDA

SUBMITTED TO THE SENATE OF THE TECHNION -
ISRAEL INSTITUE OF TECHNOLOGY

IYYAR, 5768 HAIFA MAY, 2009

This page is left blank on purpose.

THIS RESEARCH THESIS WAS SUPERVISED BY PROFESSOR DAVID MALAH AND DOCTOR MEIR BARZOHAR UNDER THE AUSPICES OF THE ELECTRICAL ENGINEERING DEPARTMENT

Acknowledgements

I would like to express my deep and sincere gratitude to my supervisor, Prof. David Malah, for all his dedicated personal guidance, understanding and encouragment during all the years of this research.

I am deeply grateful to my supervisor, Dr. Meir Barzohar, for his detailed and constructive comments, and for his important support throughout this work.

During this work I have collaborated with many colleagues for whom I have great regard, and I wish to extend my warmest thanks to Nimrod, Ziva, Avi and Yair, who have helped me with my work in the Signal and Image Processing Laboratory.

I owe my loving thanks to my wife, Alena, for all her support and patience.

The Generous Financial Help Of the Technion Is Gratefully Acknowledged.

This page is left blank on purpose.

Contents

List of Figures

List of Tables

Abstract

In this research we address the problem of redundancy-reduction of high-dimensional noisy signals, which may contain rare/anomaly vectors that we wish to preserve. Since, typically, anomalies contribute weakly to the ℓ_2-norm of the signal as compared to the noise, classical approaches based on the ℓ_2 criterion are unsatisfactory for obtaining a good representation of these vectors. Here we develop new techniques for signal subspace estimation that aim to represent well not only ℓ_2-significant signal contents, but also anomaly vectors, by optimizing another criterion based on the ℓ_∞-norm, which is more sensitive to these vectors.

In the first part of the research, we propose a greedy algorithm for the estimation of an anomaly-preserving signal-subspace and its rank. We call this algorithm: *Maximum Orthogonal-Complements Algorithm (MOCA)*. MOCA combines ℓ_2 and ℓ_∞ norms and considers two aspects: One aspect deals with signal-subspace estimation aiming to minimize the *maximum* of data-residual ℓ_2-norms, denoted as $\ell_{2,\infty}$, for a given rank conjecture. The other determines whether the rank conjecture is valid for the obtained signal-subspace by applying Extreme Value Theory results to model the distribution of the noise $\ell_{2,\infty}$-norm. These two operations are performed alternately using a suboptimal greedy algorithm, which makes the proposed approach practically plausible.

In the next part of the research we propose to adapt MOCA for anomaly detection and discrimination, as well as for population estimation of anomalies, in hyperspectral image cubes. The proposed

approach is denoted as Anomaly Extraction and Discrimination Algorithm (AXDA). The main idea of AXDA is to iteratively reduce the anomaly vector subspace-rank, found by MOCA, making the related anomalies to be poorly represented. This helps to detect them by a statistical analysis of the $\ell_{2,\infty}$-norm of data residuals. As a by-product, AXDA provides also a robust estimate of an anomaly-free background subspace and its rank.

Although the proposed greedy signal-subspace estimation algorithm makes MOCA and AXDA computationly and practically plausible, it is still only approximately minimizes the proposed $\ell_{2,\infty}$-norm of the residuals. In the following part of the research, we develop an optimal algorithm for the minization of the $\ell_{2,\infty}$-norm of data misrepresentation residuals, which we call *Maximum Orthogonal complements Optimal Subspace Estimation* (MOOSE). As any local minimization of a non-convex objective function, MOOSE is prone to getting trapped in a local minimum. Therefore, a proper initialization may be crucial for obtaining a good solution. Since MOCA finds a suboptimal solution using global principles, it provides a good initial point, which is close to the global minimum. The optimization is performed via a natural conjugate gradient learning approach carried out on the set of n dimensional subspaces in $\mathbb{R}^m$, $m > n$, which is a Grassmann manifold.

The wealth of spectral information in hyperspectral images provides plentiful amount of data for anomaly detection algorithms like AXDA. However, hyperspectral imagers are still not versatile enough for anomaly detection applications that seek mobile solutions, since they are too expensive, too heavy and consume too much power. Therefore, there is a demand for multispectral imagers that are much more versatile, although they provide a limited number of spectral channels. The proper design of multispectral filters may be crucial for anomaly detection algorithms. We conclude the research by proposing a novel unsupervised algorithm for channel reduction in hyperspectral images that

allows designing multispectral filters that are tuned for local anomaly detection algorithms. The proposed approach is based on processing a sample hyperspectral image of a typical scene that is likely to be faced by anomaly detection algorithms. Eventually, the problem of designing Multispectral Filters may be formulated as a problem of Channel Reduction in Hyperspectral Images, which is performed by replacing subsets of adjacent spectral bands by their means. An optimal partition of hyperspectral bands is obtained by minimizing the Maximum of Mahalanobis Norms (MXMN) of errors, obtained due to missrepresentation of hyperspectral bands by constants. By minimizing the MXMN of errors, one reduces the anomaly contribution to the errors, which allows to retain more anomaly-related information in the reduced channels.

List of Symbols and Abbreviations

Roman Symbols

$\mathbf{A}$	matrix of pure materials spectra (endmembers) in its columns
$\mathbf{B}$	background subspace basis
$\ell_{2,\infty}$-norm	the maximum of ℓ_2-norms of columns in a matrix
ℓ_2-norm	square root of sum of squares
p	dimensionality of observations
r	dimensionality of sources (equal to the number of endmembers)
$\mathbf{s}_i$	vector of pure materials spectra abundances comprising pixel i
$\mathbf{T}$	anomaly subspace basis
$\mathbf{X}$	matrix with observed vectors as its columns
$\mathbf{x}_i$	column number i of $\mathbf{X}$
$\mathbf{Y}$	hypothetical matrix with noise-free vectors as its columns
$\mathbf{y}_i$	column number i of $\mathbf{Y}$
$\mathbf{Z}$	matrix with noise vectors as its columns
$\mathbf{z}_i$	column number i of $\mathbf{Z}$

Greek Symbols

$\mathbf{\Omega}_h$	matrix composed of h linearly independent columns selected from the data
$\mathbf{\Psi}_{k-h}$	matrix with $k-h$ orthogonal columns, obtained via SVD of the data residuals

Other Symbols

$\mathcal{P}_{\mathbf{\Omega}_h^\perp}$	projection onto complementary space of *range* $\mathbf{\Omega}_h$

Abbreviations

AXDA Anomaly Extraction and Discrimination Algorithm

EVT Extreme Value Theory

FFR Fast Hyperspectral Feature Reduction

GLR Generalized Likelihood Ratio

GLRT Generalized Likelihood Ratio Threshold

GMRX Gaussian Mixture RX algorithm

LDR Linear Dimensionality Reduction

MDL Minimum Description Length

MOCA Maximum Orthogonal-Complements Algorithm

MOOSE Maximum of Orthogonal complements Optimal Subspace Estimation

MSD Matched Subspace Detector

MXMN Minimizing the Maximal Mahalanobis Norm

MX-SVD Min-Max SVD

PAL Potential Anomaly Loss

PCA Principal Components Analysis

RMDL Robust Minimum Description Length algorithm

ROC Receiver Operating Curves

RSNR Ratio between the contribution of Rare-vectors in the direction of the least-significant eigenvector of the rare vector-residuals in the null-space of background vectors

RX Reed-Xiaoli algorithm, a benchmark anomaly detection algorithm for hyperspectral imagery

SNR Signal to Noise Ratio

SVD Singular Value Decomposition

Chapter 1

Introduction

1.1 Anomaly Preserving Redundancy Reduction

Redundancy reduction is one of the central problems faced when dealing with high-dimensional noisy signals. In many sensor-array applications, signal vectors belong to a lower-dimensional subspace than the observed data. This signal-subspace could be estimated and used for redundancy reduction by projecting the observed data vectors onto it. The estimated signal-subspace properties should adequately reflect needs of the application that uses this low-dimensional subspace. In this research, we focus on applications that analyze anomaly vectors, such as anomaly detection in hyperspectral images. Therefore, the estimated signal-subspace should contain (preserve) such vectors.

Recently, several different algorithms have been developed to perform dimensionality reduction of low-dimensional nonlinear manifolds embedded in a high dimensional space [7]. Perhaps the principal method amongst those that provide a mapping from the high dimensional space to the embedded space is Kernel Principal Component Analysis (KPCA) [8]. KPCA first implicitly constructs a higher (sometimes infinite) dimensional space by applying the kernel trick. The nonlinear manifold structure is subsequently captured by applying traditional Principal Cmponent Analysis (PCA) in the obtained higher-dimensional space. Principal curves and manifolds give the natural geometric framework for nonlinear dimensionality reduction and extend the geometric interpretation of PCA by explicitly constructing an embedded manifold, and by encoding using standard

geometric projection onto the manifold [9]. Other nonlinear techniques include techniques for locally linear embedding, such as Locally Linear Embedding (LLE) [10], Hessian LLE [12] and Laplacian Eigenmaps [11]. These techniques construct a low-dimensional data representation using a cost function that retains local properties of the data. They can be also viewed upon as defining a graph-based kernel for KPCA [13]. In this way, the LLE-based techniques above are capable of unfolding datasets such as the Swiss roll.

Unfortunately, none of the presented nonlinear dimensionality reduction techniques explicitely aims to preserve anomaly vectors. Moreover, since anomaly vectors usually do not follow the local structure of the background data, they are likely to be misrepresented by these techniques. In this research, we explicitely deal with the problem of preserving the anomaly vectors withing the reduced dimensional subspace using linear dimensioanlity reduction paradigm. We hope, that in the furture, the developed ideas can be further extended to non-linear problems as well.

The knowledge of signal-subspace implies also a knowledge of the corresponding signal-subspace rank. In a number of applications in the literature the signal rank (order) is assumed to be known - such as the number of independent source signals in Blind Source Separation via Independent Components Analysis [6]; the order of the channel FIR model in blind single-input/multiple-output channel identification [14], [15], [16]; the signal-subspace rank in linear system identification algorithms [17], [18], [19]; the number of individual pure spectra (endmembers) in hyperspectral image processing [51], etc.

In practice, the signal-subspace and rank have to be estimated from observed vectors $\{\mathbf{x}_i\}_{i=1}^N$, assumed to satisfy the following linear model:

$$\mathbf{x}_i = \mathbf{A}\mathbf{s}_i + \mathbf{z}_i, \qquad i = 1, \ldots, N, \tag{1.1}$$

where $\mathbf{x}_i \in \mathbb{R}^p$ is the observed random vector, $\mathbf{z}_i \in \mathbb{R}^p$ is the data-acquisition or/and model noise; $\mathbf{s}_i \in \mathbb{R}^r$, and $\mathbf{A} \in \mathbb{R}^{p \times r}$, $(r \leq p)$. The observed dimension p is obviously known, whereas the signal-intrinsic dimension (rank) r is not always known.

In some of the applications above, $\mathbf{s}_i$ is a vector of hidden source signals, $\mathbf{A}$ is some "mixing" matrix through which the sources are observed; while in a hyperspectral application [1] $\mathbf{x}_i$ is an observed hyperspectral pixel, the columns of $\mathbf{A}$ are the pure materials spectra (endmembers) and $\mathbf{s}_i$ their corresponding abundances in $\mathbf{x}_i$ [51]. A schematic outline of the hyperspectral image model is presented in Fig. 1.1. For the sake of convenience, unlike in eqn. (1.1), where the pixel index i is subscripted, we superscript the pixel index i in the figure.

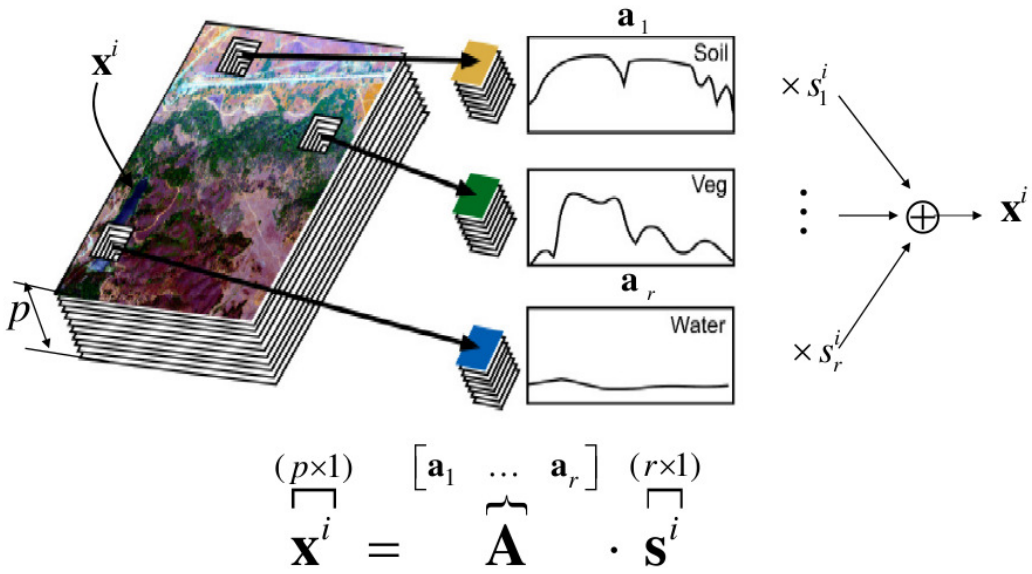


Figure 1.1: $\mathbf{x}^i$ is an observed hyperspectral pixel, the columns of $\mathbf{A}$ are the pure materials spectra (endmembers) and $\mathbf{s}^i$ their corresponding abundances in $\mathbf{x}^i$.

A number of approaches have been proposed in the literature for signal-subspace and rank estimation under the assumption that $\mathbf{s}_i$ and $\mathbf{z}_i$ are independent, stationary, zero-mean and ergodic random Gaussian processes. The use of information theoretic criteria such as minimum description length,(MDL) and Akaike's information criterion(AIC) [40], have become a solid basis for many rank

[1] Hyperspectral remote sensors collect image data simultaneously in dozens or hundreds of narrow, adjacent spectral bands. These measurements make it possible to derive a continuous spectrum for each image pixel. After adjustments for sensor, atmospheric, and terrain effects are applied, these image spectra can be used for recognition, mapping, detection and classification of surface materials [5].

estimation techniques [20], [23], [24], [39]. According to these criteria the signal rank is determined as the value k, which minimizes either one of:

$$AIC = -2\log f(\mathbf{x}_1,\ldots,\mathbf{x}_N|\hat{\mathbf{\Theta}}(k)) + 2k \quad (1.2)$$

$$MDL = -\log f(\mathbf{x}_1,\ldots,\mathbf{x}_N|\hat{\mathbf{\Theta}}(k)) + \frac{1}{2}k\log N, \quad (1.3)$$

where $f(\mathbf{x}_1,\ldots,\mathbf{x}_N|\hat{\mathbf{\Theta}}(k))$ is a parameterized family of probability densities, $\hat{\mathbf{\Theta}}(k)$ is the maximum likelihood estimate of a parameter vector $\mathbf{\Theta}(k)$, and k is a number of free adjusted parameters in $\mathbf{\Theta}(k)$. The resulting maximum likelihood estimate $\hat{\mathbf{\Theta}}(k)$ relies on the structure of the eigenvalues of the noisy signal covariance matrix $\mathbf{R}_{\mathbf{x}_i} = E\{\mathbf{x}_i\mathbf{x}_i^*\}$, where the index i could be omitted due to the stationarity assumption. Therefore, the subspace is optimal in the least-squares sense. The covariance matrix is assumed to satisfy $\mathbf{R}_{\mathbf{x}} = \mathbf{\Psi} + \sigma^2\mathbf{I}$, where $\mathbf{\Psi} = \mathbf{A}E\{\mathbf{s}\mathbf{s}^T\}\mathbf{A}^T$ and σ^2 denotes the noise variance. It is shown in [40] that the MDL yields a consistent estimate of the rank, while AIC yields an inconsistent estimate, which tends, asymptotically, to overestimate the signal rank. The resulting signal subspace estimate is obtained via Singular Value Decomposition (SVD) [48] of the observed data matrix $\mathbf{X}$. The signal subspace basis contains maximum likelihood estimates of the eigen-vectors of the matrix $\mathbf{\Psi}$, and it minimizes the ℓ_2-norm of misrepresentation residuals belonging to the complementary subspace.

It is commonly known that least-squares techniques are not robust in the sense that outliers can arbitrarily skew the desired solution [50]. By outliers we mean measurements corresponding to deviations from the nominal signal and noise characteristics for which the scheme is designed.

We start this work by proposing a redundancy reduction approach for high-dimensional noisy signals containing anomaly (rare) vectors that, typically, contribute weakly to the ℓ_2-norm of the signal as compared to the noise. This makes ℓ_2-based criteria unsatisfactory for obtaining a good representation of rare vectors, which may be of high importance in denoising and dimensionality reduction applications that aim to preserve all the signal-related information, including rare vectors, within the estimated low-dimensional signal-subspace. For example, in a problem of redundancy reduction in hyperspectral images, rare (anomalous) endmembers that are present in just a few data pixels contribute weakly to the

ℓ_2-norm of the signal, compared to the noise. Therefore, their contribution to the signal-subspace cannot be reliably estimated using an ℓ_2-based criterion, as will be shown in more detail in the following sections. Yet, the representation of the rare vectors can be crucial for anomaly detection that might follow the redundancy reduction stage.

The problem of representing well and compactly all signal vectors, including rare ones, in a low-dimensional subspace didn't attain much attention in the literature. The opposite is true: there are applications where the rare-vectors are treated as outliers that may skew the nominal signal-subspace estimation. The problem of dealing with outliers has been extensively studied in the literature. Related works ([50],[28],[29] and many others) propose robust parameter estimation techniques, which are designed to exclude the outlying measurements.

In contrast to robust parameter estimation techniques, the proposed method is designed to represent well both abundant and rare measurements, irrespective of their frequentness in the data. In other words, a good representation of all measured vectors is equally important. For this purpose, we define a deterministic matrix $\mathbf{Y} \in \mathbb{R}^{p\times N}$ that consists of signal components only. Our goal is to find the column space and the rank of $\mathbf{Y}$, given an observed matrix $\mathbf{X} \in \mathbb{R}^{p\times N}$ with columns $\mathbf{x}_1, \ldots, \mathbf{x}_N$,

$$\mathbf{X} = \mathbf{Y} + \mathbf{Z}, \tag{1.4}$$

where $rank\ \mathbf{Y} = r$ is unknown, $r < p, N$, and $\mathbf{Z} \in \mathbb{R}^{p\times N}$ is a noise matrix with i.i.d. zero-mean Gaussian entries.

Our approach combines two norms, ℓ_2 and ℓ_∞ for both signal-subspace and rank determination and considers two aspects: One aspect deals with the determination of the signal-subspace for a given rank conjecture. The other determines whether the rank conjecture is valid, given the obtained signal-subspace. The corresponding operations are performed alternately for an increasing sequence of tested subspace rank values, until the rank conjecture is affirmed. The signal-subspace is estimated by a greedy algorithm, which aims to minimize the *maximum* of misrepresentation-residual ℓ_2-norms denoted as $\ell_{2,\infty}$-norm. Mathemati-

cally, the $\ell_{2,\infty}$-norm of a matrix $\mathbf{X}$ is defined as follows:

$$\|\mathbf{X}\|_{2,\infty} \triangleq \max_{i=1,\ldots,N} \|\mathbf{x}_i\|_2, \tag{1.5}$$

where $\mathbf{x}_i$ denote columns of $\mathbf{X}$. It is easy to see that $\ell_{2,\infty}$ is a norm on a vector space $\mathcal{V}$ of $p \times N$ matrices, since for any $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2 \in \mathcal{V}$ the following holds:

1. $\|\alpha\mathbf{X}\|_{2,\infty} = |\alpha| \|\mathbf{X}\|_{2,\infty}$,
2. $\|\mathbf{X}_1 + \mathbf{X}_2\|_{2,\infty} \leq \max_i \left(\|\mathbf{x}_{1,i}\|_2 + \|\mathbf{x}_{2,i}\|_2\right) \leq \max_i \|\mathbf{x}_{1,i}\|_2 + \max_i \|\mathbf{x}_{2,i}\|_2 = \|\mathbf{X}_1\|_{2,\infty} + \|\mathbf{X}_2\|_{2,\infty}$,
3. $\|\mathbf{X}\|_{2,\infty} \geq 0$,
4. $\|\mathbf{X}\|_{2,\infty} = 0 \iff \mathbf{X} = 0$.

The proposed signal-subspace estimation process offers an appropriate compromise between the following two approaches: The first approach is based on selecting the signal-subspace basis vectors directly from the data as presented in [35]. This approach is good for representing anomalies, since it is capable of selecting anomalies from the data. However, due to noise in the obtained basis vectors, it may perform poorly in representing background pixels. The second approach is based on SVD, which represents well the background pixels. Yet, it may perform poorly in representing anomalies. Thus, the estimated admits the following form:

$$\hat{\mathcal{S}}_k = range\ \left[\mathbf{\Psi}_{k-h} | \mathbf{\Omega}_h\right], \tag{1.6}$$

i.e., $\hat{\mathcal{S}}_k$ is the space linearly spanned by columns of matrices $\mathbf{\Omega}_h$ and $\mathbf{\Psi}_{k-h}$, where $\mathbf{\Omega}_h$ is a matrix composed of h linearly independent columns selected from the data, k is the estimated signal rank and $\mathbf{\Psi}_{k-h}$ is a matrix with $k-h$ orthogonal columns, obtained via SVD of the data residuals $\mathcal{P}_{\mathbf{\Omega}_h^\perp}\mathbf{x}_i$, i=1,...,N, where $\mathcal{P}_{\mathbf{\Omega}_h^\perp}$ is a projection onto $(range\ \mathbf{\Omega}_h)^\perp$. This notation is equivalent to $\mathcal{P}^\perp_{\mathbf{\Omega}_h}$ used in [35]. The columns in $\mathbf{\Omega}_h$ are data vectors that are, typically, rare and, therefore, have a low contribution to the ℓ_2-norm of data that is not ample enough to be captured by SVD. They are selected by applying the proposed $\ell_{2,\infty}$-norm on the data misrepresentation errors by SVD. This property proposes columns of $\mathbf{\Omega}_h$ as

a good candidate for anomaly subspace basis. We use this observation later on in order to develop an anomaly detection algorithm.

The signal-subspace rank is determined by applying Extreme Value Theory results [52] to model the distribution of the misrepresentation $\ell_{2,\infty}$-norm. Since $\ell_{2,\infty}$ penalizes individual data-vector misrepresentations, it helps to represent well not only abundant-vectors, but also rare-vectors. Since we use the maximum-orthogonal-complements (residuals) for the determination of both signal-subspace and rank, we call the proposed algorithm: *Maximum Orthogonal-Complements Algorithm (MOCA).*

We present simulation results of comparing the performance of classical MDL with the proposed approach for signal-subspace rank determination. The comparison is performed on both synthetically simulated data and on a real hyperspectral image.

1.2 Adapting MOCA for Anomaly Detection

The merits of combining ℓ_2-norm and $\ell_{2,\infty}$-norm are not limited to the anomaly preserving signal representation problem only. This combination can be further extended for developing an anomaly detection algorithm. In the next part of our research we propose to adapt MOCA for anomaly detection, discrimination and population estimation of anomalies of the same type, in hyperspectral image cubes (denoted in this research as hyperspectral images). The considered sceneries are composed of reflected spectra of abundant natural ground materials such as vegetation, soil, minerals, etc., along with anomalies such as localized man-made objects. E.g., small buildings, vehicles, etc. The wealth of spectral information in hyperspectral images provides plentiful amount of data for classification tasks. One such task relates to anomaly detection, in which hyperspectral pixels have to be classified into either background material spectra class or anomaly material spectra class.

Since most often, neither prior anomaly signatures nor their statistical model, are known, anomaly detection methods first model the background and then detect anomalies by finding pixels that are not well-described by the background model. It turns out that the problem of background pixels modelling is a critical

and a subtle task. As a matter of fact, it poses a two-fold problem: On one hand, the model has to be general enough in order to accurately represent the wealth of background material spectra, so as to avoid false alarms due background pixel deviations from the model. On the other hand, the model has to be concise enough (e.g., in terms of its order/rank), limiting its ability to adapt to anomalies, and leaving anomalies to disagree with the model, which is essential for a high probability of detection.

1.2.1 Background-modelling literature review

A variety of background modelling methods appears in the literature. One type of these methods is based on estimating the underlying probability density function (pdf) of the background signature, and applying a threshold to the likelihood of tested pixels. The Reed-Xiaoli (RX) algorithm [32], is a benchmark anomaly detector for hyperspectral imagery. According to this algorithm, the background pixels in a local neighborhood of a tested pixel are assumed to be independent, identically distributed, Gaussian random vectors. After estimating the background mean vector and covariance matrix, the Mahalanobis distance between the tested pixel and the background mean vector is compared to a threshold to detect an anomaly [32]. Unfortunately, in many environments, it has been shown empirically that local background modelling by a single Gaussian provides an inadequate representation of the underlying distribution [38], leading to poor false alarm performance. This is especially true when the local background contains multiple classes of terrain.

To properly characterize nonhomogeneous backgrounds, researchers have employed a Gaussian Mixture Model (GMM) [37], [38]. This approach models the background signature distribution as a linear combination of Gaussian distributions. The Gaussian Mixture distribution is applied as a global model since the parameters are estimated over large regions. Anomaly detection may be achieved by applying the generalized likelihood ratio test (GLRT) to the model. The authors of [38] denote the related approach by GMRX.

While GMRX provides good performance, it is limited by the simplicity of Gaussian components. GMRX is further limited by the need to know or estimate

a priori the number of terrain classes in the image.

Some works, like [33], propose a nonlinear version - the Kernel-RX algorithm, in which RX is applied in an extended high-dimensional feature space associated with the original input via a certain nonlinear mapping function. In [34], the authors use anisotropic kernel reconstruction of the source image using the reference image as a way to robustly model pattern variations (that can be also viewed as a background process) in order to detect defects in patterned wafers. Another approach to local background modelling corresponds to the so-called large-margin techniques, such as support vector machines (SVMs), which detect anomalies by directly estimating a decision boundary with maximal separability. The authors of [36] propose to determine the minimal enclosing hypersurface that contains a training set of background data pixels. A training set is sampled from a window enclosing the tested pixel, excluding pixels belonging to its adjacent neighborhood (which makes this method local) that is supposed to be large enough to contain a maximum-size anomaly. The anomaly is detected by thresholding the distance from the tested pixel to the obtained hypersurface.

1.2.2 Anomaly detection via MSD

At first glance, once MOCA estimates the anomaly and background subspaces, one may apply to the result the classical Matched Subspace Detection (MSD) [30] for detection of anomalies. This method is widely used in the literature for anomaly detection in hyperspectral images when anomaly and background subspaces are known in advance (see for example references [43], [44], [45], [46], and there exist many more).

According to the MSD method, two hypotheses are defined:

$$H_0 \;:\; \mathbf{x}_i \sim \mathcal{N}[\mathbf{B}\mathbf{b}_i, \sigma^2\mathbf{I}], \tag{1.7}$$

$$H_1 \;:\; \mathbf{x}_i \sim \mathcal{N}[\mathbf{B}\mathbf{b}_i + \mathbf{T}\boldsymbol{\theta}_i, \sigma^2\mathbf{I}], \tag{1.8}$$

where $\mathcal{N}$ denotes the normal distribution and σ corresponds to the noise std; $\mathbf{B}$ and $\mathbf{T}$ are background and anomaly subspace bases with $\mathbf{b}_i$ and $\boldsymbol{\theta}_i$ background and anomaly subspace expansion coefficients of data vector $\mathbf{x}_i$, respectively.

The matrices $\mathbf{B}$ and $\mathbf{T}$, comprising the signal-subspace basis, are not necessarily orthogonal each other (i.e. $\mathbf{B}^T\mathbf{T} \neq 0$), but they are linearly independent, meaning that there is no element in $\mathbf{B}$ that can be represented as a linear combination of vectors in $\mathbf{T}$. The hypothesis H_0 corresponds to the case in which the observed vector is drawn from the interference/background subspace, contaminated by white Gaussian noise. Whereas, the hypothesis H_1 corresponds to the case in which the observed vector is a superposition of a vector from the interference/background subspace and a vector from the anomaly subspace, contaminated by white Gaussian noise.

The Generalized Log-Likelihood Ratio (GLR) is given by

$$L(\mathbf{x}) = \frac{1}{\sigma^2}\mathbf{x}^T\mathcal{P}_{\mathbf{B}^\perp\mathbf{T}}\mathbf{x}, \tag{1.9}$$

where $\mathcal{P}_{\mathbf{B}^\perp\mathbf{T}}$ is a projection onto $(range\ \mathbf{B})^\perp \bigcap range\ \mathbf{T}$ - a low-rank anomaly-matched subspace, such that interference contribution contained in $range\ \mathbf{B}$, and noise contribution contained in both $range\ \mathbf{B}$ and $(range\ \mathbf{T})^\perp$, are removed. This filter is usually called a *matched subspace filter* or a *matched field filter*. The energy of the filter output (corresponding to $L(\mathbf{x})$) is computed and compared to a threshold. The problem with this approach is that it is not well-adapted to the $\ell_{2,\infty}$ optimality criterion of the signal-subspace basis $[\mathbf{\Psi}_{k-h}|\mathbf{\Omega}_h]$ found by MOCA. Although, according to MOCA, the maximum of the data residual norms $\eta = \|\mathcal{P}_{[\mathbf{\Psi}_{k-h}|\mathbf{\Omega}_h]^\perp}\mathbf{X}\|_{\ell_{2,\infty}}$ is minimized, it may still leave large residuals (with norms below η) belonging to $(range\ \ [\mathbf{\Psi}_{k-h}|\mathbf{\Omega}_h])^\perp$. Obviously, these residuals don't contribute to $L(\mathbf{x}_i)$, which measures the norms of data vector projections onto $(range\ \mathbf{\Psi})^\perp \bigcap range\ \mathbf{\Omega}$ (see (1.9) above). This reduces the probability of anomaly detection by MSD in cases of anomalies having residual norms below η. Note, that the value of η is determined by statistics of the maximum-norm noise realization that has a narrow distribution centered around a value that is not insignificant. This value is a function of σ and the noise subspace rank.

Another disadvantage of using MSD in conjunction with a subspace determination by MOCA, is that the anomaly subspace basis $\mathbf{\Omega}$ found by MOCA is composed of vectors that were directly selected from the data, whereas the

background subspace basis satisfies *range* $\mathbf{\Psi} \subset$ *null* $\mathbf{\Omega}^T$. Therefore, the estimated anomaly subspace, as well as the background subspaces are deflected by noise. Since in hyperspectral images the background subspace and anomaly subspace are far from being orthogonal, even small deviations of the anomaly and background subspace estimations may cause background vectors to have a strong contribution to $L(\mathbf{x}_i)$ of (1.9), which rapidly increases false alarm rate by MSD. This observation is experimentally substantiated in section 5.2.

1.2.3 Combining ℓ_2-norm and $\ell_{2,\infty}$-norm for anomaly detection

The proposed algorithm, denoted here as Anomaly Extraction and Discrimination Algorithm (AXDA), is based on using the background and anomaly subspace estimates by MOCA and is designed to cope with the above MSD drawbacks. The key-point of the proposed algorithm is that it iteratively modifies both $\mathbf{\Omega}$ and $\mathbf{\Psi}$. The modification is performed by removing columns from the matrix $\mathbf{\Omega}$, one at a time, and updating the matrix $\mathbf{\Psi}$ to match the modified $\mathbf{\Omega}$. This significantly reduces the effect of noise on the anomaly detection process. AXDA uses the $\ell_{2,\infty}$-optimality criterion of MOCA to extract all anomaly pixels belonging to the same anomaly endmember, where anomaly endmembers correspond to columns of $\mathbf{\Omega}_h$. Thus, AXDA combined with MOCA, allows determination of the number of anomalies and the extraction of all pixels belonging to the same type in an unsupervised way. It still applies Extreme Value Theory (EVT) [52] to model the $\ell_{2,\infty}$-norm to construct a sharp, robust and adaptive anomaly detector, which doesn't rely on any prior knowledge about the dimensionality or statistical model of the background, without the need for tuning a one-sided hypothesis threshold, and without any prior knowledge about the number of anomaly classes and/or anomaly endmembers.

We demonstrate results of applying AXDA to real hyperspectral images. We also show there a comparison of AXDA vs. GMRX [38] and MSD algorithms in terms of Receiver Operating Curves (ROC) obtained by applying the algorithms on 5 hyperspectral images.

1.3 $\ell_{2,\infty}$-Optimal Subspace Estimation

On one hand, the greedy algorithm used by MOCA for signal-subspace estimation for a given rank only approximately minimizes the $\ell_{2,\infty}$-norm of misrepresentation residuals. On the other hand, the special form of the signal-subspace basis, used in the greedy algorithm, i.e., $[\mathbf{\Psi}_{k-h}|\mathbf{\Omega}_h]$ (see (1.6)), facilitated developing AXDA- an anomaly detection algorithm.

In the next part of our research, we propose an optimal algorithm for the signal-subspace estimation that utilizes a natural conjugate gradient learning approach proposed in [62] to minimize $\ell_{2,\infty}$-norm of the misrepresentation residuals. During the minimization process, the signal-subspace basis matrix is constrained to the Grassmann manifold defined as the set of all n dimensional subspaces in $\mathbb{R}^m$, $n \leq m$ [62]. Since $\ell_{2,\infty}$-norm of the misrepresentation residuals can be also referenced as the maximum orthogonal complement norm, we denote the proposed algorithm as Maximum of *Orthogonal complements Optimal Subspace Estimation* (MOOSE).

The optimal signal-subspace obtained by MOOSE, can be used to improve the performance of MOCA in terms of signal-subspace estimation error. Unfortunately, MOOSE does not produce a signal-subspace basis with a special structure like MOCA, which makes an adaptation of MOOSE for anomaly detection purposes much more difficult.

1.4 Multispectral Filters Design for Anomaly Detection

We conclude this research by proposing a novel unsupervised technique for Designing Multispectral Filters that facilitates a better performance of local anomaly detection algorithms. The proposed approach is based on processing a sample hyperspectral image of a typical scene that is likely to be faced by anomaly detection algorithms, where the sample image is not neccessarily required to include anomalies. Eventually, the problem of Multispectral Filters design may be formulated as a problem of Redundancy Reduction in Hyperspectral Channels,

which is performed by replacing adjacent spectral bands by their means. This is a real-world Redundancy Reduction problem that requires preserving anomalies.

A common problem of local anomaly detection algorithms is so-called *Hughes phenomenon* [73], according to which the perfomance of anomaly detection algorithms significantly deteriorates when the number of pixels is severely limited for an accurate learning of the local background models. In order to alleviate the effect of *Hughes phenomenon*, one has to reduce the number of hyperspectral bands, since the complexity of background models is proportional to the hyperspectral data dimensionality.

Linear Dimensionality Reduction (LDR) is a widely used preprocessing technique for the alleviation of *Hughes phenomenon* in classification problems [74], [77], [75]. LDR also allows to eliminate redundancies occuring due to high correlations among adjacet bands. Of particular interest are techniques that reduce the dimensionality of hyperspectral data by replacing subsets of adjacent bands by their means, since the resulting features can be physically interpreted as responses of *multispectral* filters, which may be tuned to aplication-dependent needs. Thus, the authors of [76] propose top-down and bottom-up algorithms designed to find subsets of bands yielding high Fisher discrimination among classes. In [74] one can find an approach that groups the channels into a partition that increases interclass distance computed on a training set. Another approach, based on dynamic programming, is proposed in [77]. It minimizes the mean squared error of representing all hyperspectral pixels in the image by piece-wise constant spectral segments.

Unfortunately, little attention has been drawn in the literature to channel reduction techniques designed to improve the performance of local anomaly detection algorithms. This problem is of high importance in applications that seek a technology to construct high performance multispectral filters for anomaly detection. An appealing approach for this purpose is proposed in [77], denoted as Fast Hyperspectral Feature Reduction (FFR). It looks for a best piece-wise constant represenation of the hypersectral data and does not assume any prior knowledge about the data. However, FFR is not well-taylored to data that contains anomalies, since it uses the mean squared error based (ℓ_2-norm based) criterion. As

discussed earlier, this criterion is known to be insensitive to anomaly contributions and, as a result, may lead to a poor represention of anomalies.

The novel approach proposed here is based on a new criterion that is designed to retain spectral channels containing valuable anomaly-related information for anomaly detection algorithms. The optimal partition of the spectrum is obtained by *Minimizing the Maximal Mahalanobis Norm* of errors, obtained due to the misrepresentaion of spectral intervals by constants. Therfore, we denote the proposed technique as Min-Max MN or, in short, MXMN. By minimizing the MXMN of errors, one reduces the anomaly contribution to the errors, which allows to retain more anomaly-related information in the reduced channels, if there are anomalies in the sample image. In the case that the sample scene does not contain anomalies, minimizing the MXMN of errors allows smoothing out spectral bands containing background clutter, which are unfavorable for the anomaly detection since they are likely to mask possible subtle anomaly contributions to other bands.

We compare MXMN with other dimensionality reduction techniques, such as classical principal components analysis (PCA) and FFR, by examining the results of the Reed-Xiaoli (RX) algorithm [32], a benchmark anomaly detector for hyperspectral imagery, applied after the dimensionality reduction. We demonstrate that the proposed approach corresponds to a better ROC curve, as compared to PCA and FFR, for a wide range of false alarm rates, and even better than obtained by applying RX on the original data (without the dimensionality reduction) for the important range of low false-alarm rates.

The thesis is organized as follows: In chapter 2 we present the proposed redundancy reduction approach for high-dimensional noisy signals containing anomaly (rare) vectors, named MOCA [1]. In chapter 3 MOCA is used for developing an anomaly detections algorithm, named AXDA [3]. Chapter 4 deals with an $\ell_{2,\infty}$-optimal signal subspace estimation via a natural learning on a Grassmann manifold [2]. Finally, in chapter 5 we develop a technique for Designing Multispectral Filters that facilitates a better performance of local anomaly detection algorithms [4].

Chapter 2

Anomaly Preserving Redundancy Reduction

2.1 Optimality criterion for subspace estimation

Before getting into the development of an estimator of a subspace that may include rare vectors, we first characterize the presence of rare-vectors. For demonstrational purposes, we show in Fig. 2.1 a schematic plot of a subspace of abundant vectors and rare-vectors. The abundant vectors (marked by dots) lie close to a subspace spanned by the vector $\mathbf{v}_0$. As it is seen in the figure, the rare vectors (marked by circles and dashed arrows) $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4$ don't belong to the abundant vector subspace spanned by $\mathbf{v}_0$. Obviously, rare-vectors are characterized by their low number compared to the number of abundant vectors. Rare vectors are supposed to lie far from the abundant vector subspace. They, however, are allowed to belong to a subspace of a dimension lower than their number. It is important to stress that unlike in the example (for $p = 2$), the observed dimensionality p (in the general case) is expected to exceed the dimension of the subspace spanned by abundant and rare vectors combined.

The example above can be generalized by the following property:

Rare vector presence property: The $p \times N$ matrix $\mathbf{Y}$ is said to contain rare-vectors if there exists a decomposition $\mathbf{Y} = [\mathbf{Y}_1|\mathbf{Y}_2]\mathbf{\Pi}$, where $\mathbf{\Pi}$ is some permutation matrix, $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are $p \times N_1$ and $p \times N_2$ submatrices of $\mathbf{Y}$, such that $N_1 + N_2 = N$, $N_1 \gg N_2$, and *range* $\mathbf{Y}_1 \subset$ *range* $\mathbf{Y}$.

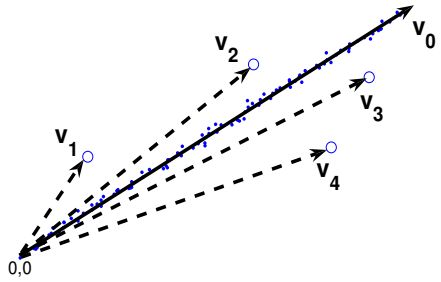


Figure 2.1: **Schematic plot demonstrating rare vectors presence in data.** $\mathbf{v}_0$ spans abundant vectors (dots) subspace; $\mathbf{v}1, \mathbf{v}_2, \mathbf{v}_3$ and $\mathbf{v}_4$ denote rare vectors (circles).

This property states that the matrix $\mathbf{Y}$ is considered to contain rare vectors if the number of $\mathbf{Y}$ columns that are linearly independent of all the other $\mathbf{Y}$ columns is relatively small.

In order to develop a rare-vector preserving signal-subspace estimator, we should define an optimality criterion that is sensitive to the appearance of rare-vectors in the data. First, we consider an ℓ_2-based optimality criterion, since it appears in Singular Value Decomposition (SVD) - a well-known technique for the signal-subspace estimation [48]. Then, we show that ℓ_2-based criteria are not appropriate for estimating a signal-subspace that contains rare vectors, and propose combining ℓ_2 and ℓ_∞-based criteria as a remedy.

As noted above, at the signal-subspace determination stage, the rank is assumed to be known, say $rank\ \mathbf{Y} = k$.

2.1.1 Signal-subspace estimation via SVD

According to the SVD approach, the signal-subspace $\mathcal{S}_k = range\ \mathbf{Y}$ is estimated by minimizing the ℓ_2 norm of the residuals:

$$\begin{aligned} \hat{\mathcal{S}}_k \quad &= \quad \underset{\mathcal{L}}{\operatorname{argmin}} \ \|\mathbf{X} - \mathcal{P}_{\mathcal{L}}\mathbf{X}\|_{Fb}^2 = \underset{\mathcal{L}}{\operatorname{argmin}} \ \|\mathcal{P}_{\mathcal{L}^\perp}\mathbf{X}\|_{Fb}^2 \\ &\quad \text{s.t. } rank\ \mathcal{L} = k, \end{aligned} \tag{2.1}$$

where $\|\cdot\|_F b$ denotes Frobenius norm, $\mathcal{L} \subset \mathbb{R}^p$ and $\mathcal{P}_{\mathcal{L}}$ denotes an orthogonal projection onto subspace $\mathcal{L}$. It can also be shown that under a Gaussian as-

sumption on the columns of $\mathbf{Y}$, $\hat{\mathcal{S}}_k$ coincides with the maximum-likelihood (ML) estimation of $\mathcal{S}_k$ [41]. The estimated signal-subspace $\hat{\mathcal{S}}_k$ is obtained via SVD of the observation matrix $\mathbf{X}$ as $\mathbf{X} = \hat{\mathbf{U}}\hat{\mathbf{S}}\hat{\mathbf{V}}'$, where $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$ are $p \times p$ and $N \times p$ matrices with orthonormal columns, respectively, and $\hat{\mathbf{S}} = \text{ diag } \{\hat{s}_1, \ldots, \hat{s}_p\}$, $\hat{s}_1 \geq \hat{s}_2 \geq \ldots \geq \hat{s}_p$. The signal subspace $\hat{\mathcal{S}}_k$ is equal to span of $\{\hat{\mathbf{u}}_\mathbf{1}, \ldots, \hat{\mathbf{u}}_\mathbf{k}\}$ - the first k columns of $\hat{\mathbf{U}}$ (see [48] for details).

2.1.2 Drawbacks of minimizing the ℓ_2 norm in the presence of rare-vectors

Intuitively, it seems that minimizing the observation residuals $\mathcal{P}_{\mathcal{L}^\perp}\mathbf{x}_i$, $i = 1, \ldots N$, as a function of $\mathcal{L}$, could be appropriate for estimating $\mathcal{S}$. Indeed, for $\mathcal{L} = \mathcal{S}$,

$$\mathcal{P}_{\mathcal{L}^\perp}\mathbf{x}_i = \mathcal{P}_{\mathcal{L}^\perp}\mathbf{z}_i, \qquad i = 1, \ldots, N, \tag{2.2}$$

which means that given a precise signal-subspace estimation, the data residuals are equal to the corresponding noise residuals. Whereas, for $\mathcal{L} \neq \mathcal{S}$, one expects to obtain signal contributions in the residual subspace $\mathcal{L}^\perp$, which are likely to increase the residual squared norm $\|\mathcal{P}_{\mathcal{L}^\perp}\mathbf{x}_i\|^2$. This can be seen from the fact that since $\mathbf{z}$ is statistically independent of $\mathbf{y}$, so are their projections onto the null-space of $\mathcal{L}$. Moreover, since z_i are zero mean i.i.d., it is expected that

$$\|\mathcal{P}_{\mathcal{L}^\perp}\mathbf{x}_i\|^2 \approx \|\mathcal{P}_{\mathcal{L}^\perp}\mathbf{y}_i\|^2 + \|\mathcal{P}_{\mathcal{L}^\perp}\mathbf{z}_i\|^2. \tag{2.3}$$

Therefore, looking for $\hat{\mathcal{S}}$ that minimizes residual norms is reasonable. However, using an ℓ_2 norm (like in (2.1)) can be inappropriate in the presence of rare-vectors, since the contribution of rare-vector residuals to the ℓ_2-norm may be much weaker than the contribution of noise-residuals. As a result, the estimated subspace $\hat{\mathcal{S}}$ may be skewed by noise in a way that completely misrepresents the rare-vectors. In some practical cases this miss-representation may occur with high probability, as demonstrated in simulations below.

First, we define the Rare-vector Signal-to-Noise Ratio as follows:

$$RSNR \triangleq \frac{s^2_{min}(\mathcal{P}_{\mathbf{Y}^{\perp}_{abund}}\mathbf{Y}_{rare})}{E\{\|\mathcal{P}_{\mathbf{Y}^{\perp}_{abund}}\mathbf{z}\|_2^2\}} = \frac{s^2_{min}(\mathcal{P}_{\mathbf{Y}^{\perp}_{abund}}\mathbf{Y}_{rare})}{(p-k)\sigma^2}, \tag{2.4}$$

where $\mathbf{Y}_{rare}$ is a submatrix of $\mathbf{Y}$ composed of all rare-vectors, $\mathbf{Y}_{abund}$ is a submatrix of $\mathbf{Y}$ composed of the remaining (abundant) vectors; $\mathcal{P}_{\mathbf{Y}^{\perp}_{abund}}$ is a projection onto the null-space of $\mathbf{Y}_{abund}$; $s^2_{min}(\mathbf{D})$ is the squared minimal non-zero singular value of the argument matrix $\mathbf{D}$, and σ^2 is the noise variance. The choice of the minimal non-zero singular value is essential, since it reflects the rare-vectors subspace perturbation by additive noise [71], i.e., the error in rare-vector subspace estimation. That is, RSNR measures the ratio between the contribution of rare-vectors in the direction of the least-significant eigenvector of the rare vector-residuals in the null-space of background vectors, and the contribution of noise in that direction. We also define SNR as follows:

$$SNR \triangleq \frac{\|\mathbf{Y}_{abund}\|^2_{Fb}}{pN\sigma^2}. \tag{2.5}$$

Now, we describe the setup of simulations that show a typical case for which rare-vectors are misrepresented by SVD. A $p \times N = 10^2 \times 10^5$ signal matrix $\mathbf{Y}$ (which corresponds to a typical hyperspectral image cube consisting of 10^5 pixel-vectors of dimension 10^2, each) was generated, such that $\mathbf{Y} = [\mathbf{Y}_{abund}|\mathbf{y}_{rare}]$, using a Gaussian distribution for the columns of $\{\mathbf{Y}_{abund}$ with a covariance matrix $\mathbf{C}_{\mathbf{Y}_{abund}} = 100\sigma^2\mathbf{I}_{p,q}$, $q = 5$, and $\mathbf{y}_{rare} \in null\ \mathbf{Y}^T_{abund}$, where $\mathbf{I}_{p,q}$ denotes a diagonal $p \times p$ matrix with $q \leq p$ nonzero diagonal entries, all equal to 1. Since $s^2_{min}(\mathbf{y}_{rare}) = \|\mathbf{y}_{rare}\|_2^2$, it follows that

$$RSNR = \frac{\|\mathbf{y}_{rare}\|_2^2}{E\{\|\mathcal{P}_{\mathbf{C}_{\mathbf{Y}_{abund}}{}^{\perp}}\mathbf{z}\|_2^2\}} = \frac{\|\mathbf{y}_{rare}\|_2^2}{(p-k)\sigma^2}, \tag{2.6}$$

Then, the "measured" data-matrix $\mathbf{X}$ was obtained as $\mathbf{X} = \mathbf{Y} + \mathbf{Z}$, where the $\mathbf{Z}$ columns are Gaussian with a covariance matrix $\mathbf{C}_z = \sigma^2\mathbf{I}_{p,p}$.

In our simulations, for each $RSNR$ value $\mathbf{X}$ was generated 100 times. We consider 50 RSNR values, sampled uniformly in the range $[0, \ldots, 170]$ for $\sigma^2 = 1$, as shown in Fig. 2.2 (a). In a dashed (dot-dashed) line we plot the minimum

(maximum) of 100 generated values (per RSNR value) of

$$\nu_k \triangleq \|\mathcal{P}_{\hat{\mathbb{S}}_k^{\perp}}\mathbf{X}\|_{2,\infty}^2 \triangleq \max_{j=1,\ldots,N} \|\mathcal{P}_{\hat{\mathbb{S}}_k^{\perp}}\mathbf{x}_j\|_2^2, \tag{2.7}$$

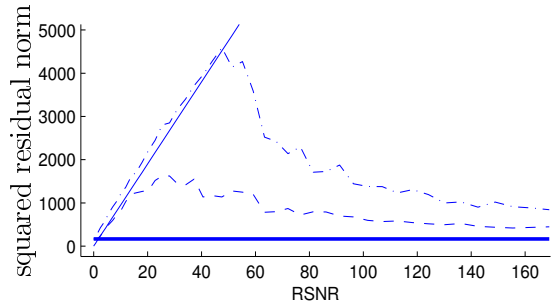
where $\ell_{2,\infty}$ is a norm defined by selecting the *maximum* ℓ_2*-norm* of the data vector residuals (corresponding to the null-space of $\hat{\mathbb{S}}_k$, $k = q+1 = 6$ in (2.7)). In a thin solid line we plot $\|\mathbf{y}_{rare}\|_2^2$ as a function of RSNR.

We repeated the simulation above for matrices $\mathbf{X}$ with $\mathbf{Y} = \mathbf{Y}_{rare}$ (i.e., there are no rare-vectors in the data). The horizontal heavy solid line shows the mean value of ν_k, $k = q = 5$, corresponding to data without rare-vectors. In both cases - with and without a rare-vector, $\hat{\mathbb{S}}_k$ was obtained via SVD.

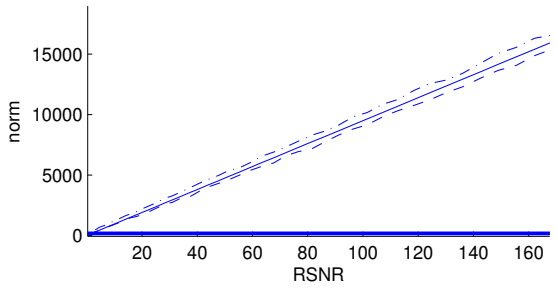
The maximum residual norm $\nu_k = \|\mathcal{P}_{\hat{\mathbb{S}}_k^{\perp}}\mathbf{X}\|_{2,\infty}^2$ in data without rare-vectors has a narrow distribution, since it approximately equals to the maximum norm of the noise residuals $\|\mathcal{P}_{\hat{\mathbb{S}}_k^{\perp}}\mathbf{Z}\|_{2,\infty}^2$, which has a narrow distribution, explained by Extreme Value Theory results, as shown in Appendix A. Therefore, ν_k has nearly a "deterministic" behavior in data without rare-vectors.

However, in the presence of rare-vectors (for $k = q+1$), it is likely to obtain ν_k values that are higher than $\|\mathcal{P}_{\hat{\mathbb{S}}_k^{\perp}}\mathbf{Z}\|_{2,\infty}^2$. Thus, as it is seen from the figure, there is a range of RSNR values ($0 < RSNR < 140$, $p = 10^2$ and $N = 10^5$ in this example), for which the value of ν_k in the presence of a rare-vector lies much higher (between the dot-dashed and dashed lines, representing the min and max values, respectively) than the nearly deterministic value of ν_k in the absence of a rare-vector (heavy horizontal solid line). This phenomenon corresponds to the poor representation of rare-vectors by SVD. This range of RSNR values, however, is of high practical importance in some applications. For instance, in hyperspectral that we examined, characterized by SNR = 100, the observed RSNR satisfies RSNR $\leq$ 30, which means that SVD would most likely fail to appropriately represent rare-vectors in this application. On the other hand, for high RSNR values, the rare-vector contributions becomes stronger in the ℓ_2-sense, compared to the noise contributions. As a result, for high RSNR values, SVD represents well the rare-vectors. This can be seen from the fact that the dot-dashed line in Fig. 2.2 converges to the heavy horizontal solid line.

For clarification, in Fig. 2.2 (b) we show results of the above simulation for an assumed incorrect dimensionality value of $k-1$. As expected, SVD "prefers" to represent abundant vectors. This results in a maximum misrepresentation error that is dictated solely by the norm of the rare-vector for a much wider range of $RSNR$ values. Note that the min and max values are not equal because of the noise added to the rare vector. We also simulated the case of a "wrong" dimensionality of $k+1$ and noticed, as expected, that it produces results close to the case of the correct dimensionality k in Fig. 2.2 (a).



(a)



(b)

Figure 2.2: **Monte-Carlo simulation of SVD-based signal-subspace estimation in the presence of rare-vectors for $p = 10^2$ and $N = 10^5$.** The rare-vector squared norm $\|\mathbf{y}_{rare}\|_2^2$ (solid thin line), the sample-minimum of maximum data-residual squared-norms ν_k *in the presence* of rare-vectors (dashed line), the sample-maximum of the maximum data-residual squared norm ν_k *in the presence* of a rare-vector (dot-dashed line), the sample-mean of maximum noise-residual squared-norms ν_k *in the absence* of rare-vectors (heavy horizontal solid line); a) for correct rank k b) for "wrong" rank $k-1$.

In summary, the above example demonstrates that SVD may poorly represent the rare-vectors for an important range of low RSNR values.

2.1.3 Signal-Subspace determination by $\ell_{2,\infty}$-norm minimization

In the last example we have seen that SVD, being ℓ_2-optimal (2.1), may not be sensitive to rare-vectors, leaving large rare-vector residuals in $\hat{\mathcal{S}}_k^{\perp}$. In order to tackle this problem, we propose using $\ell_{2,\infty}$ instead of ℓ_2, which transforms the optimization problem (2.1) to the following form:

$$\begin{aligned} \hat{\mathcal{S}}_k &= \underset{\mathcal{L}}{\operatorname{argmin}} \ \|\mathcal{P}_{\mathcal{L}^{\perp}}\mathbf{X}\|_{2,\infty}^2 \\ &\quad \text{s.t.} \qquad \operatorname{rank}\mathcal{L} = k. \end{aligned} \tag{2.8}$$

The objective function of this optimization problem is not differentiable and, therefore, is hard to optimize. In order to make the problem differentiable, analogously to the Chebyshev (minimax) approximation problem in [66], the problem of (2.8) can be recast as follows:

$$\begin{aligned} \hat{\mathcal{S}}_k &= \underset{\mathcal{L},\gamma}{\operatorname{argmin}} \ \gamma \\ \text{s.t.} \qquad & \|\mathcal{P}_{\mathcal{L}^{\perp}}\mathbf{x}_j\|_2^2 \leq \gamma \qquad \forall j = 1,\ldots,N, \\ & \operatorname{rank}\mathcal{L} = k, \end{aligned} \tag{2.9}$$

where the additional parameter γ was introduced to bound all residual squared norms $\|\mathcal{P}_{\mathcal{L}^{\perp}}\mathbf{x}_j\|^2$ (including the maximal one) from above. Thus, by minimizing this bound with respect to $\mathcal{L}$, one minimizes the maximum residual norm corresponding to $\|\mathbf{X}\|_{2,\infty}$ of (2.8), which makes problems (2.8) and (2.9) equivalent.

Although the obtained equivalent optimization problem is differentiable, it still seems to be practically intractable because of the large multiplicity of constraints, which is equal to N (the number of data vectors). Therefore, in the next section we propose a suboptimal greedy algorithm that is found to produce good results.

2.2 Signal-Subspace determination by combining SVD with min-max of residual norms (MX-SVD)

In order to make the minimization of (2.8) or (2.9) computationally plausible, we propose to constrain the sought $\hat{\mathbb{S}}_k$ basis to be of the following form:

$$\hat{\mathbb{S}}_k = \text{range } [\mathbf{\Psi}_{k-h}|\mathbf{\Omega}_h], \tag{2.10}$$

where $\mathbf{\Omega}_h$ is a matrix composed of h columns selected from $\mathbf{X}$, and $\mathbf{\Psi}_{k-h}$ is a matrix with $k-h$ orthogonal columns, obtained via SVD of $\mathcal{P}_{\mathbf{\Omega}_h^\perp}\mathbf{X}$. As demonstrated in the previous section, the $\ell_{2,\infty}$ norm of data-vector residuals is governed by the rare-vector miss-representations via SVD (which is ℓ_2 - optimal), whereas abundant vectors can be successfully represented via SVD. Therefore, the main idea of the proposed approach, which we denote as MX-SVD, is to collect rare-vectors into $\mathbf{\Omega}_h$ in order to directly represent the rare-vectors subspace. Since rare-vectors are not necessarily orthogonal to abundant vectors, matrix $\mathbf{\Omega}_h$ also partially represents *abundant vectors*. The residual abundant vector contribution to the null-space of $\mathbf{\Omega}_h^T$ is represented by principal vectors found by applying SVD on $\mathcal{P}_{\mathbf{\Omega}_h^\perp}\mathbf{X}$. As noted above, the columns in $\mathbf{\Omega}_h$ are directly selected from $\{\mathbf{x}_i\}_1^N$, the set of noisy data vectors. Although this makes *range* $\mathbf{\Omega}_h$ a noisy estimation of the pure rare-vectors subspace, it still represents well the noisy rare-vectors in the data, which is, actually, the main objective of MOCA.

The determination of the basis vectors of $\hat{\mathbb{S}}_k$ in terms of $[\mathbf{\Psi}_{k-h}|\mathbf{\Omega}_h]$, for a given value of k, is performed as follows: First, we initialize $[\mathbf{\Psi}_k|\mathbf{\Omega}_0]$, such that

$$\mathbf{\Psi}_k \;=\; [\mathbf{u}_1, \ldots, \mathbf{u}_k]\,;\; \mathbf{\Omega}_0 = [], \tag{2.11}$$

where $\mathbf{u}_1, \ldots, \mathbf{u}_k$ are k principal left singular vectors of $\mathbf{X}$.

Then, a series of matrices $\{[\mathbf{\Psi}_{k-j}|\mathbf{\Omega}_j]\}_{j=0}^k$ is constructed such that

$$\mathbf{\Omega}_{i+1} \;=\; [\mathbf{\Omega}_i|\mathbf{x}_{\omega_i}] \tag{2.12}$$

$$\mathbf{\Psi}_{k-i-1} \;=\; \left[\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_{k-i-1}\right], \tag{2.13}$$

where, for each $i = 0, \ldots, k-1$, ω_i is the index of a data vector $\mathbf{x}_{\omega_j}$ that has the

maximal residual squared norm r_i:

$$\omega_i \triangleq \underset{n=1,\ldots,N}{\operatorname{argmax}} \|\mathcal{P}_{[\boldsymbol{\Psi}_{k-i}|\boldsymbol{\Omega}_i]^\perp}\mathbf{x}_n\|, \tag{2.14}$$

$$r_i \triangleq \|\mathcal{P}_{[\boldsymbol{\Psi}_{k-i}|\boldsymbol{\Omega}_i]^\perp}\mathbf{x}_{\omega_i}\|^2, \tag{2.15}$$

and $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_{k-i-1}$ are $k-i-1$ principal left singular vectors of $\mathcal{P}_{\boldsymbol{\Omega}_{i+1}^\perp}\mathbf{X}$. Thus, the k columns of $[\boldsymbol{\Psi}_{k-j}|\boldsymbol{\Omega}_j]$, for each $j = 0, \ldots, k$, span k-dimensional subspaces, respectively. Each subspace is spanned by a number of data vectors collected in the matrix $\boldsymbol{\Omega}_j$ and by SVD-based vectors that best represent (in ℓ_2 sense) the data residuals in the null-subspace of $\boldsymbol{\Omega}_j$. Moreover, each subspace is characterized by it's maximum-norm data representation error r_j. One of these subspaces is to be selected as $\hat{\mathbb{S}}_k$. In the light of our objective to minimize the worst-case representation error, we choose $\hat{\mathbb{S}}_k = \text{range } [\boldsymbol{\Psi}_{k-h}|\boldsymbol{\Omega}_h]$, with the value of h that minimizes the $\ell_{2,\infty}$-norm of residuals, i.e.,

$$h = \underset{j=0,\ldots,k}{\operatorname{argmin}} r_j. \tag{2.16}$$

This policy combines the ℓ_2-based minimization of abundant vector-residual norms with ℓ_∞-based minimization of rare vector residual norms. As we have seen earlier, the rare-vectors have large residuals with respect to principal subspaces found by SVD. This property would cause them to be selected among columns of $\boldsymbol{\Omega}_h$, whereas the abundant vector projections onto the null-space of $\boldsymbol{\Omega}_h$ would lie in the range $\boldsymbol{\Psi}_{k-h}$. A flowchart summarizing the MX-SVD process is shown in Fig. 2.3.

2.3 MX-SVD vs. SVD - simulation results

In Fig. 2.4, we show empirical pdfs of $\|\mathcal{P}_{\hat{\mathbb{S}}_k^\perp}\mathbf{X}\|_{2,\infty}^2$, obtained via a Monte-Carlo simulation for $k = r_{abund} + r_{rare} = 5+3 = 8$, where r_{abund} is the rank of abundant-vectors subspace and r_{rare} is the number of rare-vectors, which were generated as in the previous example of section 2.1.2 by appending orthogonal vectors of equal norms $\{\mathbf{y}_j\}_{j=1}^{r_{rare}}$, $\mathbf{y}_j \in \textit{null } \mathbf{Y}_{abund}^T$. A $10^2 \times 10^5$ matrix $\mathbf{X}$ was generated 1000 times for RSNR = 10, $\sigma = 1$. The pdfs of $\|\mathcal{P}_{\hat{\mathbb{S}}_k^\perp}\mathbf{X}\|_{2,\infty}^2$ corresponding to

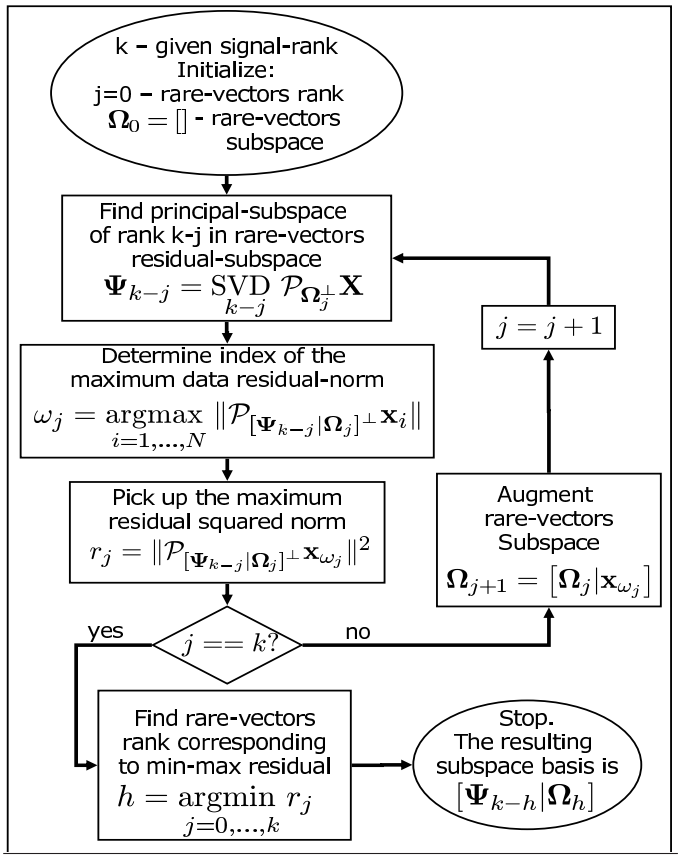


Figure 2.3: **MX-SVD flowchart.** For a given signal subspace rank value k, constructs a signal-subspace basis of the form $\hat{\mathbb{S}}_k = [\mathbf{\Psi}_{k-h}|\mathbf{\Omega}_h]$, $h \in$ *integers* $[0, k]$, that minimizes $\|\mathcal{P}_{\hat{\mathbb{S}}_k^{\perp}}\mathbf{X}\|_{2,\infty}^2$, where $\mathbf{\Omega}_h$ is responsible for representing rare-vectors and $\mathbf{\Psi}_{k-h}$ is responsible for representing the remaining (abundant) vectors in the data.

subspace estimation by MX-SVD (dashed line) and SVD (solid line) are shown in Fig. 2.4(a).

It is clearly seen from the figure that max-norm residuals obtained via MX-SVD have a lower value and have a much narrower pdf, as compared to residuals obtained by SVD. As a matter of fact, the MX-SVD-related pdf is very close to the pdf of $\|\mathcal{P}_{\hat{\mathbb{S}}_k^{\perp}}\mathbf{Z}\|_{2,\infty}^2$, which equals to the squared norm of the maximum-norm noise residual. This fact is supported by Fig. 2.4(b). Here, we plot the empirical pdf of $\|\mathcal{P}_{\hat{\mathbb{S}}_k^{\perp}}\mathbf{X}\|_{2,\infty}^2$ obtained via MX-SVD (dashed line) versus the exact pdf of $\|\mathcal{P}_{\hat{\mathbb{S}}_k^{\perp}}\mathbf{Z}\|_{2,\infty}^2$ (solid line), obtained from a model (with the above parameters) that is based on Extreme Value Theory results, presented in Appendix A.

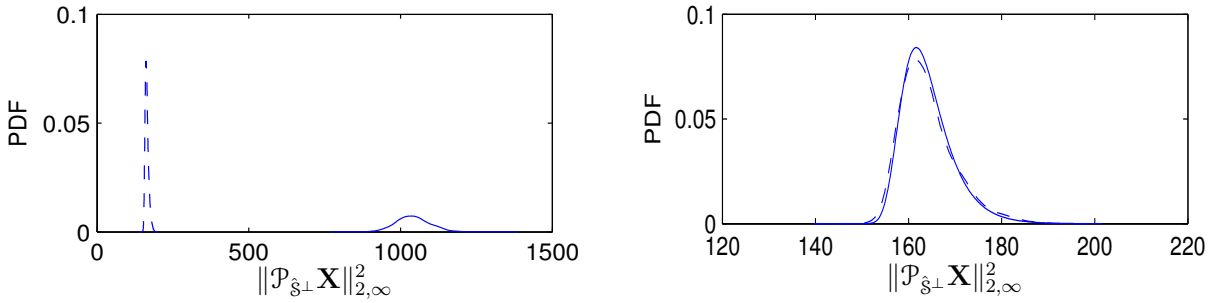


Figure 2.4: **The pdfs of $\|\mathcal{P}_{\hat{\mathbf{S}}_k^{\perp}}\mathbf{X}\|_{2,\infty}^2$, obtained via a Monte-Carlo simulation.** (a) The empirical pdfs of $\|\mathcal{P}_{\hat{\mathbf{S}}_k^{\perp}}\mathbf{X}\|_{2,\infty}^2$ obtained by MX-SVD (dashed line) and SVD (solid line) for RSNR $= 10$, $\sigma = 1, p = 10^2, N = 10^5$, $k = r_{abund} + r_{rare} = 5 + 3 = 8$ (b) The empirical pdf of $\|\mathcal{P}_{\hat{\mathbf{S}}_k^{\perp}}\mathbf{X}\|_{2,\infty}^2$ by MX-SVD (dashed-line) versus the exact pdf of $\|\mathcal{P}_{\hat{\mathbf{S}}^{\perp}}\mathbf{Z}\|_{2,\infty}^2$ (solid line).

In summary, MX-SVD was designed to yield $\|\mathcal{P}_{\hat{\mathbf{S}}_k^{\perp}}\mathbf{X}\|_{2,\infty}^2 \approx \|\mathcal{P}_{\hat{\mathbf{S}}_k^{\perp}}\mathbf{Z}\|_{2,\infty}^2$ for $k \geq r$ in the presence of rare vectors, as opposed to SVD, which produces arbitrary large residuals for a range of low RSNR values. The fact that for $k \geq r$ the maximum-norm residuals are governed by the maximum-norm realization of the noise will be used in the next section for constructing a signal-subspace rank estimator, which is based on statistical properties of $\|\mathcal{P}_{\hat{\mathbf{S}}_k^{\perp}}\mathbf{Z}\|_{2,\infty}^2$.

2.4 Rank Determination

In this section we construct a signal-subspace rank estimator $\hat{r}$ (recall that the signal-subspace basis may include rare-vectors). This rank estimator is based on examining the maximal data residual norms $\|\mathcal{P}_{\hat{\mathbf{S}}_k^{\perp}}\mathbf{X}\|_{2,\infty}^2$, for an increasing sequence of k values. The only thing we know about $\|\mathcal{P}_{\hat{\mathbf{S}}_k^{\perp}}\mathbf{X}\|_{2,\infty}^2$ is that for $k < r$, it could be arbitrarily higher than $\|\mathcal{P}_{\hat{\mathbf{S}}_k^{\perp}}\mathbf{Z}\|_{2,\infty}^2$; whereas for $k \geq r$, due the signal-subspace estimation approach, which minimizes $\ell_{2,\infty}$-norm of residuals, one may assume that the maximum-norm data residual is governed by the maximum-norm noise residual, i.e.,

$$\|\mathcal{P}_{\hat{\mathbf{S}}_k^{\perp}}\mathbf{X}\|_{2,\infty}^2 \approx \|\mathcal{P}_{\hat{\mathbf{S}}_k^{\perp}}\mathbf{Z}\|_{2,\infty}^2. \tag{2.17}$$

Guided by (2.17), we consider a test that determines the rank r as follows in the next section.

2.4.1 Signal and noise hypotheses assessment

We assume that for some k, $r_{abund} \leq k \leq r$, the signal-subspace $\hat{\mathcal{S}}_k$, estimated by MX-SVD described above, is close to the subspace of abundant-vectors. This assumption is plausible due to the SVD-part of the MX-SVD process that is designed to represent well the abundant-vectors subspace, which is of rank $r_{abund} \leq r$. As a result, the abundant-vector residuals in the complementary subspace $\hat{\mathcal{S}}_k^{\perp}$ are governed by the noise contribution, whereas the rare-vector residuals may still include significant signal contributions. Thus, for $k \geq r_{abund}$, the set of all data-vector indices can hypothetically be divided into two subsets according to the properties of data-vector residuals:

$$\begin{aligned} \Gamma_k &\triangleq \{\text{indices } \gamma_j \text{ of abundant-vector residuals}\} \\ \Delta_k &\triangleq \{\text{the remaining data-vector indices } \delta_i\ \}, \end{aligned} \tag{2.18}$$

such that $j = 1, \ldots, \#\Gamma_k, i = 1, \ldots, \#\Delta_k$ and $\#\Gamma_k \gg \#\Delta_k$, where $\#$ denotes cardinality of a set.

Let η_k be the maximum data-residual squared-norm, $\eta_k = \max\limits_{j=1,\ldots,N} \|\mathcal{P}_{\hat{\mathcal{S}}_k^{\perp}} \mathbf{x}_j\|^2$. Given η_k, we formulate the following two hypotheses:

$$H_0 \ : \ \eta_k \text{ belongs to } \Gamma_k, \tag{2.19}$$

$$H_1 \ : \ \eta_k \text{ belongs to } \Delta_k. \tag{2.20}$$

The following notation will help us to perform a statistical analysis of η_k:

$$\begin{aligned} \nu_k &\triangleq \max_{\gamma_j \in \Gamma_k} \|\mathcal{P}_{\hat{\mathcal{S}}_k^{\perp}} \mathbf{x}_{\gamma_j}\|^2 \\ \xi_k &\triangleq \max_{\delta_i \in \Delta_k} \|\mathcal{P}_{\hat{\mathcal{S}}_k^{\perp}} \mathbf{x}_{\delta_i}\|^2. \end{aligned} \tag{2.21}$$

Now, η_k, can be expressed as:

$$\eta_k = \max(\nu_k, \xi_k). \tag{2.22}$$

Due to the assumption leading to (2.18), and according to (2.21), the value of ν_k is governed by the extreme value statistics of maximum-norm noise realizations. Now, we set the rank estimator $\hat{r}$ to be equal to the minimal value of k for which the following condition is satisfied:

$$p(H_0|\eta_k) \geq p(H_1|\eta_k), \tag{2.23}$$

which means that the optimal rank is reached when there is a higher likelihood that the maximum data-residual squared norm η_k is governed by the noise statistics (i.e., it doesn't include significant signal contributions).

In order to evaluate the conditional probabilities $p(H_0|\eta_k)$ and $p(H_1|\eta_k)$, one has to specify pdfs $f_{\nu_k}(\cdot)$ and $f_{\xi_k}(\cdot)$, or, equivalently, cdfs $F_{\nu_k}(\cdot)$ and $F_{\xi_k}(\cdot)$. Whilst the probability of maximum-norm noise realization ν_k can be determined by Extreme Value Theory results, as shown in Appendix A, the pdf of ξ_k is generally unknown. The only thing we know about ξ_k is that at each iteration k, it has to be less or equal η_{k-1}. A possible choice for $f_{\xi_k}(\cdot)$ is therefore,

$$\xi_k \sim U[0, \eta_{k-1}], \tag{2.24}$$

where U denotes a uniform distribution.

Now, it can be shown (see Appendix B for details) that posterior hypotheses probabilities are given by:

$$p(H_0|\eta_k) = \frac{\eta_k f_{\nu_k}(\eta_k)}{\eta_k f_{\nu_k}(\eta_k) + F_{\nu_k}(\eta_k)}, \tag{2.25}$$

$$p(H_1|\eta_k) = \frac{F_{\nu_k}(\eta_k)}{\eta_k f_{\nu_k}(\eta_k) + F_{\nu_k}(\eta_k)}, \tag{2.26}$$

where the expressions above are valid for $0 \leq \eta_k \leq \eta_{k-1}$. It is important to note, however, that the functional form of the posterior conditional probabilities, as given in (2.25) and (2.26), does not depend on η_{k-1}. Moreover, due to a successive application of MX-SVD for an increasing sequence of k values, it is guaranteed that $0 \leq \eta_k \leq \eta_{k-1}$. Therefore, in the forthcoming expressions, we omit explicit mention of the argument boundaries.

Fig. 2.5(a) shows the corresponding graphs of these posterior probabilities for a residual-subspace rank $l = p - k = 10^2$, where p is the dimensionality of the data vectors $\mathbf{x}$, the total number of data vectors $N = 10^5$, and the noise std $\sigma = 1$. It is clearly seen that the transition region between hypotheses is steep and narrow. Actually, its width depends on the form of f_{ν_k} (see Fig. 2.5(b)), which is well-localized, as explained in Appendix A.

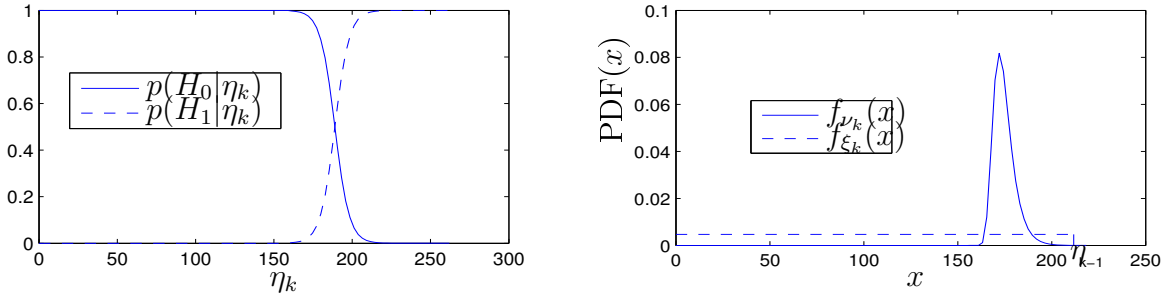


Figure 2.5: **a) posterior conditional hypotheses probabilities $p(H_0|\eta_k)$ and $p(H_1|\eta_k)$ b) distributions of maximum squared-norm of rare (solid line) and abundant (dashed line) vector residuals.** For residual-subspace rank $l = 10^2$, total number of data vectors $N = 10^5$, and the noise std $\sigma = 1$.

In summary, the signal-subspace rank is determined by applying MX-SVD and examining condition (2.23) for an increasing sequence of k values. As the maximum-norm residual becomes low and (2.23) becomes true, it can no longer be confidently associated with the signal contribution, and the procedure is terminated. The estimated rank is equal to the last-examined k value. As was already noted above, this combination of applying MX-SVD and examining condition (2.23) for an increasing sequence of k values, defines what we called Maximum Orthogonal Complement Algorithm (MOCA), and is summarized next.

2.4.2 MOCA summary for combined subspace and rank determination

In this subsection we summarize the proposed approach of signal-subspace and rank determination via MOCA by presenting its major parts in the flowchart of Fig. 2.6.

The algorithm begins with an initial guess for the signal-subspace rank, such as $k = 1$. At each rank value iteration, the signal-subspace basis $\mathbf{\Phi}_k = [\mathbf{\Psi}_{k-h}|\mathbf{\Omega}_h]$ is obtained via MX-SVD of section 2.2, using the conjectured rank k. Then the data maximum residual-norm is calculated in the null space of $\mathbf{\Phi}_k$. This norm is tested in order to decide if it belongs to the noise hypothesis (this decision is performed by evaluating inequality (2.23)). If the noise hypothesis passes, the algorithm is terminated, and the estimated signal-subspace and rank equals to the span of the last obtained $\mathbf{\Phi}_k$, and to the last value of k, respectively. Otherwise, the rank conjecture k is incremented and a new iteration is carried out.

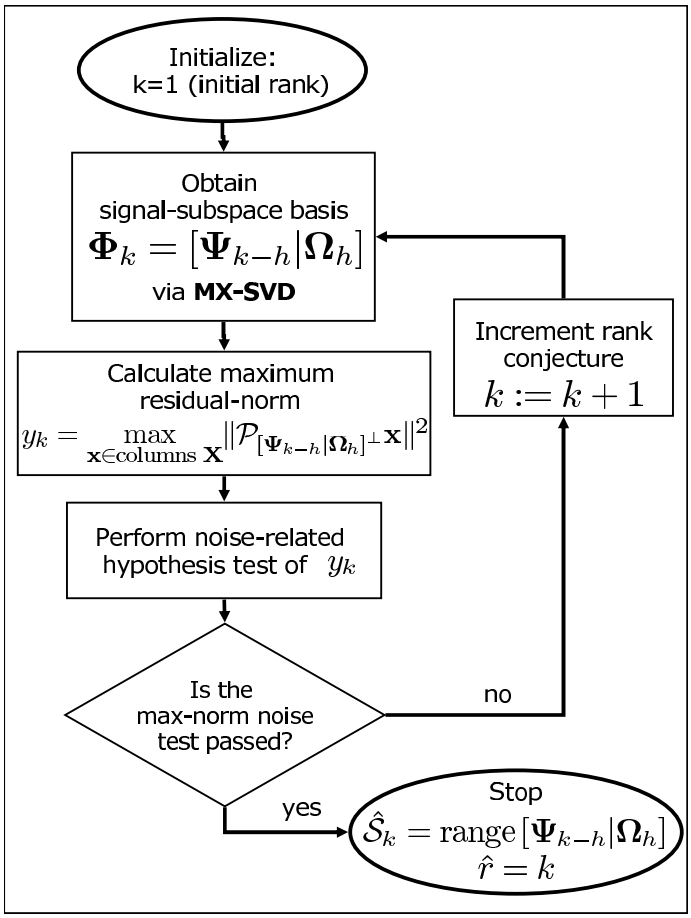


Figure 2.6: **Maximum Orthogonal Complement Algorithm (MOCA) flowchart.**

2.5 Comparison of rank determination by MOCA vs. MDL

In this section we compare the performance of MOCA with that of the Minimum Description Length (MDL) approach for signal-subspace rank determination.

2.5.1 MDL basics

MDL is a widely-used model-order determination criterion, based on coding arguments and the minimum description length principles [25],[26]. The same rule has been also obtained via a rather different approach, based on a Bayesian Information Criterion (BIC) [39]. Thus, in [40] it is proposed to apply the MDL for determining the model-order of (1.1), with $\{\mathbf{s}_i\}$ being an ergodic Gaussian process with a positive definite covariance matrix and the noise variance σ^2 is unknown.

The MDL was also proven in [40] to be consistent in terms of yielding the true signal-subspace rank, with probability one, as the sample size N increases. It is based on minimizing the following criterion with respect to k:

$$\mathrm{MDL}(k) = -\ln f(\mathbf{X}|\hat{\mathbf{\Theta}}(k)) + \frac{1}{2}\eta \log N, \tag{2.27}$$

where $f(\cdot)$ is a family of probability densities parameterized by $\mathbf{\Theta}(k)$, and η denotes the number of model degrees of freedom. In our case, where σ^2 is known, manipulation of the results in [40] gives:

$$\mathrm{MDL}(k) = \sum_{i=1}^{k} \log(\hat{l}_i) + (p-k)\log(\sigma^2) + k + \sum_{i=k+1}^{p} \frac{\hat{l}_i}{\sigma^2} + k(2p-k)\frac{\log(N)}{N}, \tag{2.28}$$

where $\{\hat{}_i\}_1^p$ denote eigenvectors of the data-covariance matrix $\mathbf{R} \triangleq E\{\mathbf{x}\mathbf{x}^T\}$, and σ^2 is the known noise variance.

2.5.2 Simulation of rank determination by MOCA vs. MDL

In this subsection we compare the results of applying MOCA and MDL to simulated examples, in the presence of rare vectors, and assess their performance in terms of rank errors expressed by rank-RMSE defined by $e_{rank} \triangleq \sqrt{E(r-\hat{r})^2}$.

Fig. 2.7 shows the performance of MOCA vs. MDL for $r = r_{abund} + r_{rare} = 5 + 10 = 15$, SNR $= 100$ (the rare vectors were generated as in the example of section 2.1.2); with Fig. 2.7 (a) and (b) corresponding to different sizes of $N = 10^4$ and $N = 10^5$, respectively. MOCA and MDL were tested 50 times for each value of RSNR . Then, the rank-determination errors were calculated and plotted. The error e_{rank} obtained by MDL for a range of low RSNR values, which is a function of N, is equal to 10 (the rare-vectors subspace rank r_{rare}). In other words, the MDL completely fails to determine r at low RSNR values. The dependence of e_{rank} on N is obvious - the larger the sample size N is, the more "blind" becomes MDL to rare-vectors, which have to be much stronger in order to become apparent to MDL. Thus, the correct rank determination by MDL starts only at very high values of RSNR. In contrast to MDL, MOCA performs much better, with low values of e_{rank} obtained already at a very low RSNR value.

It turns out, that the probability of rank determination error by MOCA becomes small and approximately constant already for RSNR as small as 2 (see Appendix C for details). This turning point is marked by a heavy dot-dashed vertical line in Fig. 2.7.

It is important to note that the simulations above were designed to reflect a typical situation seen in hyperspectral images, in which the background process is characterized by $SNR \approx 100$ (20dB), while anomalies are characterized by $RSNR \leq 30$. Hence, the simulations above indicate that MDL is expected to be "blind" to the anomaly subspace rank in typical hyperspectral images, whereas MOCA is expected to succeed in estimating the rank.

A reasonable question that arises is how to identify the transition point, below which one should use MOCA due to its ability to recover the rank at low RSNR values, and above which one could use MDL due to its computational simplicity. We turn to (2.28) and notice that MDL(k) has to accept its minimum at k. That means that the increase in penalty (the last term of (2.28)) has to be smaller than

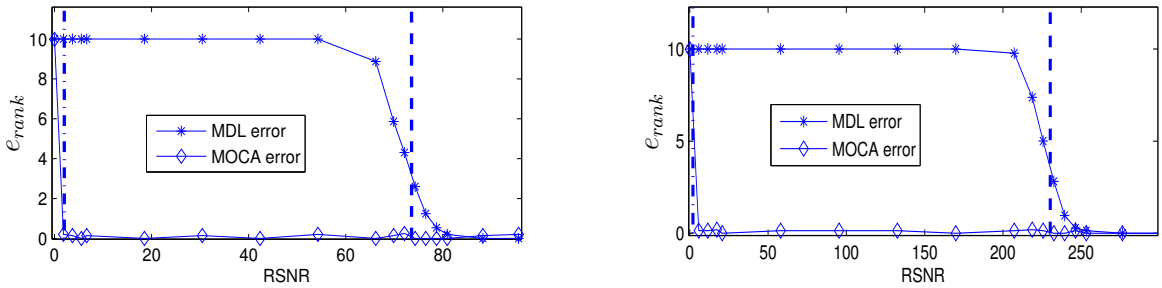
the decrease in minus log-likelihood (the first part of (2.28)) in the transition from $k-1$ to k and, respectively, larger in the transition from k to $k+1$. Now, due to construction of $\mathbf{Y}$ in simulations (see 2.1.2)), the eigenvalue $\hat{l}_k$ stemming from rare vectors is assumed to satisfy:

$$\hat{}_k \approx \sigma^2 + \|\mathbf{y}_{rare}\|^2/N = \sigma^2 + RSNR(p-k)/N. \tag{2.29}$$

By neglecting k with respect to p (since $k \ll p$) and approximating $\hat{l}_i \approx \sigma^2$ for $i > k$, then, with some straightforward manipulations, one obtains that the equilibrium between the change in penalty and change in log-likelihood (when k is changed to $k+1$) is reached when:

$$-\log(\sigma^2 + RSNR\frac{p}{N}) + RSNR\frac{p}{N\sigma^2} = 2p\frac{\log(N)}{N}. \tag{2.30}$$

By numerically solving (2.30) with respect to RSNR, one obtains the turning point in RSNR value below which the MDL is expected to be unreliable in determining contribution of rare-vector to rank. This turning point is marked by a heavy dashed vertical line in Fig. 2.7.



(a) (b)

Figure 2.7: **MOCA vs MDL comparison via Monte Carlo simulations.** The rank estimation error $e_{rank} = \sqrt{E(r-\hat{r})^2}$ in the presence of 10 rare-vectors as a function of RSNR, for (a) $N = 10^4$, (b) $N = 10^5$. The heavy dashed and dot-dashed vertical lines delimit a region in which MOCA is reliable enough and has better performance than MDL.

2.5.3 Comparing MOCA with MDL on real data

In this section we compare results of MOCA and MDL for signal-subspace and rank determination of hyperspectral images. We then compare MOCA and MDL-SVD performances in dimensionality reduction of hyperspectral images by applying MDL and MOCA on a bank of about 50 hyperspectral cubes of size 400×450 with 65 spectral bands. Due to space limitations, results for a typical cube are demonstrated here.

One of hypespectral bands of this cube is shown in Fig. 4.4. Each pixel in this hyperspectral image corresponds to a 65×1 vector. MOCA assumes the noise to be statistically independent between spectral bands. Therefore, in order to make the noise i.i.d., the noise std in each band was estimated and normalized to 1 by scaling the data.

First, MOCA was applied on the upper part of the image shown in Fig. 4.4 that is delimited by horizontal and vertical white lines. According to ground-truth evidence, this part corresponds to a "pure background signal" stemming from agricultural fields radiance. Indeed, the signal-subspace determined by MOCA is given by $\hat{\mathbb{S}}_{\mathrm{I}} = \mathbf{\Psi}_7$, which corresponds to $k = 7, h = 0$; i.e., no rare-vectors were selected in order to represent best the signal-subspace in this subimage. Then, MOCA was applied on the entire image producing $\hat{\mathbb{S}}_{\mathrm{II}} = [\mathbf{\Psi}_6 | \mathbf{\Omega}_4]$, which corresponds to $k = 10, h = 4$. Such a result can be explained by the presence of anomaly pixels (marked by circles) located at the bottom of the image. According to the ground truth, these pixels belong to vehicles, which are anomalous to the natural surroundings in the image. Thus, there are 4 data vector pixels comprising $\mathbf{\Omega}_4$ columns, which represent the anomaly pixels subspace in the data.

It should be stressed that the number of columns in $\mathbf{\Omega}_4$ may be less than the number of anomaly pixels in the data, since the columns of $\mathbf{\Omega}_4$ are intended to span the anomaly pixels subspace, which may be of a rank lower than the number of anomaly pixels. Moreover, since the rare-vector subspace and the background-subspace are not orthogonal to each other, the columns of $\mathbf{\Omega}_4$ may span a subspace close to the background subspace $\mathbf{\Psi}_7$, found initially in $\hat{\mathbb{S}}_{\mathrm{I}}$, which may produce a $\ell_{2,\infty}$-norm of residuals small enough in order to stop at an earlier

MOCA iteration. This explains why the background subspace rank is lower in $\hat{\mathcal{S}}_{\mathrm{II}}$ than in $\hat{\mathcal{S}}_{\mathrm{I}}$ (the pure-background case with no anomalies).

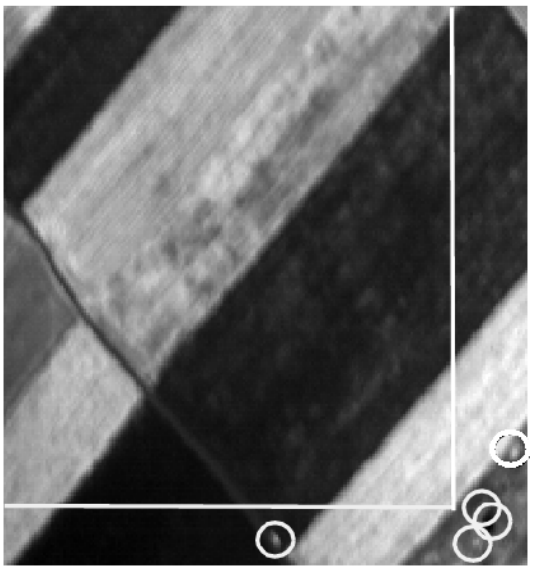


Figure 2.8: **Signal-subspace and rank determination in a hyperspectral image.** MOCA was applied on (i) the subimage above the white lines produces $\hat{\mathcal{S}}_{\mathrm{I}} = \mathbf{\Psi}_{\mathrm{I}}$, (ii) the entire image includes anomalies marked by circles, producing $\hat{\mathcal{S}}_{\mathrm{II}} = [\mathbf{\Psi}_{\mathrm{II}}|\mathbf{\Omega}_{\mathrm{II}}]$. The MDL-estimated rank in both cases is 7.

Turning to the examination of MDL performance, we note that MDL is known to be sensitive to deviations from the white noise assumption [27]. We have found that the noise normalization preprocessing that produced good results with MOCA, isn't sufficient for a proper operation of MDL, since it doesn't compensate for small correlations between noise components in adjacent bands due to a crosstalk between adjacent sensors, and still leaves noise component variances different. We have applied, therefore, the Robust MDL (RMDL) algorithm of [27] (which assumes different diagonal entries $\sigma_1^2, \ldots, \sigma_p^2$), but with a slight modification, to account for correlations between the adjacent noise components. The modification we applied to RMDL is described in Appendix D.

We have applied the modified RMDL algorithm for rank estimation on the above mentioned hypespectral images: the pure-background subimage and the anomaly-containing entire image. In the pure-background subimage case, the MDL has produced a rank of 7, which is in accordance with the result of MOCA.

However, in the case of the entire image, which contains rare vectors, the MDL algorithm misses the contribution of rare-vectors to signal-subspace rank, leaving the rank value at 7, whereas MOCA manages to detect the contribution of anomalies to the signal-subspace and rank producing a higher rank of 10 corresponding to both the background and rare-vector pixels.

Now, all hyperspectal pixels were projected onto the subspace found by SVD, of rank found by MDL, as well as onto the signal-subspace basis $\hat{S}_{II}$ found by MOCA. In Fig. 2.9 we show squared norms of residuals corresponding to (a) MDL-SVD, and (b) MOCA based subspaces. It is clearly seen that MOCA-based dimensionality reduction better represents all pixels in the image including the anomalies, compared to MDL-SVD based dimensionality reduction, which misrepresents anomaly pixels producing high-intensity residuals (white blobs in Fig. 2.9 (a)) at their location.

(a) (b)

Figure 2.9: **Squared norms of residuals corresponding to (a) MDL-SVD, and (b) MOCA based subspaces.**

2.6 Summary

In conclusion, in this chapter we have proposed an algorithm for redundancy reduction of high-dimensional noisy signals, named MOCA, which is designed for applications where a good representation of *both* the abundant and the rare

vectors is essential. The combined subspace of rare and abundant vectors is obtained by using the proposed $\ell_{2,\infty}$-norm that penalizes individual data-vector miss-representations. Since this criterion is hard to optimize, a sub-optimal greedy algorithm is proposed. It uses a combination of SVD and direct selection of vectors from the data to form the signal-subspace basis. The rank is determined by applying Extreme Value Theory results to model the distribution of the maximal noise-residual ℓ_2-norms. In simulations, conducted for various rare-vectors signal-to-noise conditions, the proposed approach is shown to yield good results for practically-significant RSNR values (RSNR essentially measures the SNR of rare-vectors with respect to noise), for which the classical methods of SVD and MDL fail to determine correctly the signal-subspace and rank, respectively, of high dimensional signals composed of abundant and rare vectors.

The proposed approach was also applied for the signal-subspace and rank determination of a hyperspectral image with and without anomaly pixels. The results of MOCA were found to be equal to those of MDL (or when necessary RMDL) for the pure-background subimage, whereas in the presence of anomalies, MOCA has detected a higher rank than MDL, while MDL produced the same rank as in the pure-background case. This indicates that MDL failed to determine correctly the signal-subspace rank of a hyperspectral image composed of both abundant and rare vectors, whereas MOCA succeeded in representing it well.

Chapter 3

Anomaly Extraction and Discrimination Algorithm (AXDA)

In this chapter we propose an Anomaly Extraction and Discrimination Algorithm (AXDA) that employs signal-subspace and rank estimation results obtained by the above described MOCA.

Let's recall that the signal-subspace basis $\mathbf{\Phi}$ produced by MOCA admits the following form:

$$\mathbf{\Phi}_{\hat{r}} = [\mathbf{\Psi}_{\hat{r}-h} | \mathbf{\Omega}_h] \,, \tag{3.1}$$

where $\hat{r}$ is the estimated signal-subspace rank, the sub-matrix $\mathbf{\Omega}_h$ consists of h linearly independent columns selected from the data matrix $\mathbf{X}$ and the sub-matrix $\mathbf{\Psi}_{\hat{r}-h}$ consists of $\hat{r} - h$ principal components of $\mathcal{P}_{\mathbf{\Omega}^\perp}\mathbf{X}$.

As it was already noted above, the matrix $\mathbf{\Omega}_h$ represents the anomaly vectors subspace. First of all, given the matrix $\mathbf{\Omega}_h$, one can mark anomaly vectors by locating indices of $\mathbf{\Omega}_h$ columns in the original data matrix $\mathbf{X}$. However, this straightforward method does not enable us to find all the pixels in the data that belong to the anomaly subspace, since not all anomalies are guaranteed to be within the columns of $\mathbf{\Omega}_h$. For example, in a case where there are number of vehicles having the same anomalous reflected spectrum, only one pixel representing all vehicle pixels would be collected by MOCA into $\mathbf{\Omega}_h$. Therefore, neither all vehicles, nor all vehicle pixels would be marked by this straightforward method. It was experimentally observed that simple approaches such as looking for data vec-

tors lying close enough to $\mathbf{\Omega}_h$ columns (or, alternatively, to the subspace spanned by $\mathbf{\Omega}_h$ columns) are of a low practical value due to the need for a threshold and due to a high false-alarm rate caused by background interference.

In order to detect and discriminate all anomaly pixels, we propose a new algorithm that extracts all anomalies in the data and associates them with the $\mathbf{\Omega}_h$ columns found by MOCA. As stated earlier, the proposed algorithm is denoted as Anomaly Extraction and Discrimination Algorithm (AXDA). For the sake of clarity, we first present a concise outline of AXDA in Fig. 3.1.

3.1 Concise outline of AXDA

The main idea of the algorithm is to iteratively reduce the anomaly vector subspace-rank by dropping columns of $\mathbf{\Omega}_h$, producing submatrices $\{\mathbf{\Omega}_j\}_{j=0}^{h-1}$. Since for a given rank $\hat{r}$, the matrix $[\mathbf{\Psi}_{\hat{r}-h}|\mathbf{\Omega}_h]$ minimizes the $\ell_{2,\infty}$ of data residuals in the $(range\ [\mathbf{\Psi}_{\hat{r}-h}|\mathbf{\Omega}_h])^{\perp}$ (as noted above), dropping columns from $\mathbf{\Omega}_h$ increases the $\ell_{2,\infty}$-norm of data residuals. Obviously, this change in residual norms occurs in pixels that are well-represented by the dropped column, including the residual norm of the dropped column itself. Therefore, this operation reveals anomaly vectors in the data that belong to the dropped column by increasing their residual norms. The increased residual norms are compared to the $\ell_{2,\infty}$-norm of data residuals from the previous iteration, which are determined by the test in (2.23) as stemming from noise. If the increased norms exceed the $\ell_{2,\infty}$-norm of data residuals from the previous iteration, the corresponding pixels are marked as belonging to the dropped column and are depleted from data. Depletion of such pixels makes the $\ell_{2,\infty}$-norm of data residuals in the current iteration to pass again the noise hypothesis in (2.23). All operations in this paragraph are performed in block (2) of Fig. 3.1.

There are two indices, j and s that keep track of anomaly subspace and *total* signal subspace dimensionality, respectively, at each iteration. The index j, which is initialized as $j = h$, denotes the anomaly subspace rank throughout the AXDA iterations. It is decremented by one at each iteration. The index s (initialized as $s = \hat{r}$), denotes the total signal-subspace rank throughout the AXDA iterations. The initialization is depicted in block (1) of Fig. 3.1. Since the depletion

of anomaly vectors is supposed to decrease the anomaly subspace dimensionality in the data by one (see block (3)), one expects the total signal-subspace rank s to decrease by one as well. However, this is not always the case. For example, in cases where the dropped anomaly vector is highly correlated with the background subspace, dropping it from $\mathbf{\Omega}_h$, impairs the ability of $[\mathbf{\Psi}_{\hat{r}-h}|\mathbf{\Omega}_{h-1}]$ (here $s = \hat{r} - 1$) to represent well the background subspace. In order to sequentially deplete anomaly vectors at each iteration, one needs to maintain the $\ell_{2,\infty}$-norm of data residuals to be low enough to admit the noise hypothesis in (2.23) at each iteration (see block (5)). Therefore, in this example, we need to increase the background dimensionality by one, which is performed by retaining s unchanged. Therefore, the decision to decrement the total signal-subspace s rank (see block (7)) is taken only if the reduced-rank subspace meets the maximum-norm noise residual hypothesis (see block (6)).

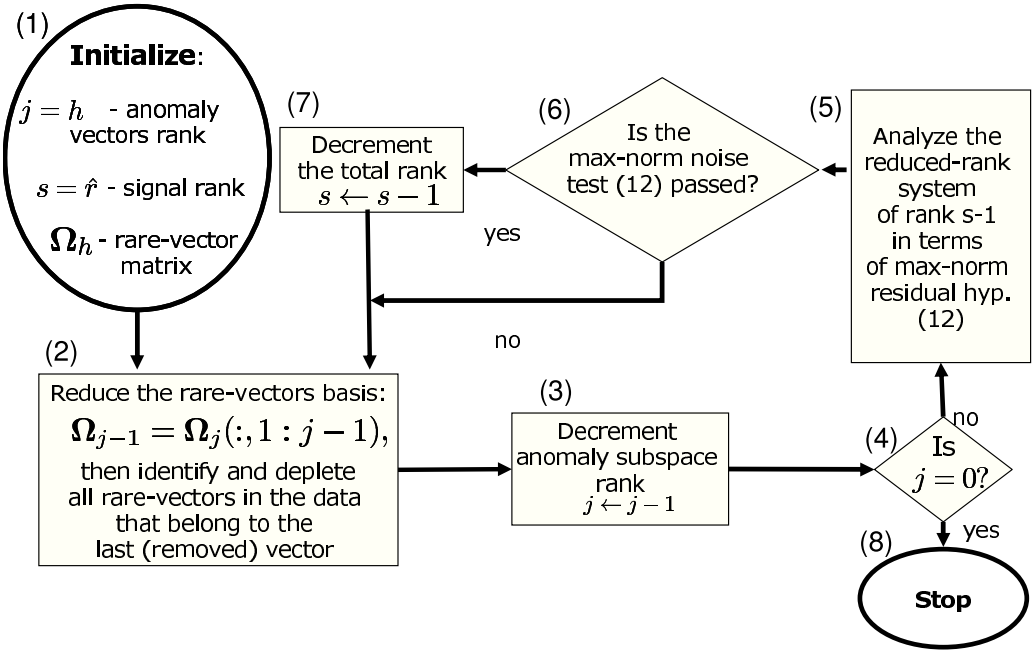


Figure 3.1: **A concise outline of Anomaly Extraction and Discrimination Algorithm (AXDA).** The notation in block (2) is MATLAB® notation.

3.2 Detailed description of AXDA

At this point, we are ready to describe the AXDA algorithm in detail as shown in Fig. 3.2, where we mainly introduce details of block (2) in Fig. 3.1. The numbering of the following items correspond to the block numbers in Fig. 3.2.

1. **Initialization**

   The AXDA algorithm starts by initializing $j = h$, the number of anomaly vectors in $\mathbf{\Omega}_h$ and $s = \hat{r}$, the determined rank of signal subspace *range* $[\mathbf{\Psi}_{s-j}|\mathbf{\Omega}_j]$, and the maximum-residual norm denoted by η_s (see (2.21)), all as obtained from MOCA.

2. **Reduction of anomaly subspace basis**

   It is important to note that initially, the number $j = h$ of anomaly vectors in the partition $\mathbf{\Phi}_{\hat{r}} = [\mathbf{\Psi}_{\hat{r}-h}|\mathbf{\Omega}_h]$, obtained by MOCA, is optimal for the given signal-subspace rank, i.e., a decrease of h for a given $\hat{r}$ would result in an increased maximum-residual norm and, possibly, of other residual norms of data-vectors.

   We intentionally alter this optimality by dropping the last column of $\mathbf{\Omega}_j$, providing $\mathbf{\Omega}_{j-1}$. This operation is designed to detect anomalies related to the last column of $\mathbf{\Omega}_j$.

3. **Calculation of a new background vector representation basis $\mathbf{\Psi}_{s-j+1}$, corresponding to the new anomaly subspace basis $\mathbf{\Omega}_{j-1}$**

   In order to retain the total signal-subspace rank s, a new background vector representation basis $\mathbf{\Psi}_{s-j+1}$ is calculated by applying SVD on $\mathcal{P}_{\mathbf{\Omega}_{j-1}^{\perp}}\mathbf{X}$ that matches the reduced-rank matrix $\mathbf{\Omega}_{j-1}$.

4. **Calculation of data residual-norms in the obtained residual-subspace**

$$r_i = \|\mathcal{P}_{[\mathbf{\Psi}_{s-j+1}|\mathbf{\Omega}_{j-1}]^{\perp}}\mathbf{x}_i\|^2 \tag{3.2}$$

5. **Detection of anomaly vectors belonging to the dropped column j**

In this block we identify indices of all anomaly vector residuals that exceeded the noise level η_s, which is equal to the maximum residual-norm initially obtained from MOCA.

6. **Decision about the next operation, based on previous block results**

   In this block we decide about the next operation based on whether anomaly vectors were found in the previous block. If there are such indices, then we perform the inner loop, in which we deplete the found anomaly vectors, recalculate $\mathbf{\Psi}_{s-j+1}$, and try to detect more anomaly vectors. Otherwise, the depletion of anomaly vectors in this iteration is completed and other operations of current iteration are performed.

7. **Association of found anomaly vectors to j-th column of $\mathbf{\Omega}_j$**

   This block belongs to the inner loop of anomaly vectors depletion. We associate all data vectors indices (found in the block (5)) to the dropped column j and store them. Therefore, the corresponding anomaly vectors are denoted as *j-associated anomaly vectors.*

8. **Depletion of found j-associated anomaly vectors from input data and recalculation of $\mathbf{\Psi}_{s-j+1}$**

   Since $(range\ \mathbf{\Omega}_{j-1})^{\perp}$ contains j-associated anomaly vector contributions, the subspace corresponding to $\mathbf{\Psi}_{s-j+1}$ (obtained earlier via SVD in Block (3)) is expected to be diverted in a way that aims to reduce these contributions along with the background vector residual-norms. As a result, not all anomaly vectors corresponding to the dropped j-th column of $\mathbf{\Omega}_j$ may be detected via tresholding their corresponding norms by η_s in Block (5). In order to remedy this problem, we deplete the j-associated anomaly vectors in the data (detected in Block (5)) and perform operations of blocks (3) - (5) again in order to obtain a more precise estimation of the background vectors subspace $\mathbf{\Psi}_{s-j+1}$, which is not diverted by the j-associated anomaly vectors found in Block (5).

9. **Decrementing of anomaly subspace rank**

Once all j-associated anomaly vectors are depleted, the rank of the anomaly subspace can be reduced by one. However, this does not necessarily mean that the total signal representation rank should also drop by one. This can be explained as follows: As it was already noted earlier, the background and anomaly subspaces in the hyperspectral images are not orthogonal. Therefore, if one reduces the rank of $[\mathbf{\Psi}_{s-j}|\mathbf{\Omega}_j]$ by removing a column from $\mathbf{\Omega}_j$, one might transfer a significant amount of background contribution to the complementary subspace $(range\ [\mathbf{\Psi}_{s-j}|\mathbf{\Omega}_{j-1}])^{\perp}$, which means that the reduced-rank subspace basis $[\mathbf{\Psi}_{s-j}|\mathbf{\Omega}_{j-1}]$ might not represent well the signal-subspace of the data after the j-associated anomaly-vectors depletion.

Therefore, the decrementing of anomaly subspace rank j does not necessarily entails decrementing the total signal-basis rank s. Thus, to decide if the total signal-basis rank s should be also decremented, we again employ, in the next blocks, the maximum-norm hypothesis testing (2.23). Due to the algorithm construction (see blocks (4),(5),(6)), it is guaranteed that at the input to this block, the subspace $(range\ [\mathbf{\Psi}_{s-j+1}|\mathbf{\Omega}_{j-1}])^{\perp}$ doesn't contain signal contributions. It is left to determine if the same holds true for the subspace $(range\ [\mathbf{\Psi}_{s-j}|\mathbf{\Omega}_{j-1}])^{\perp}$, which corresponds to the reduced total signal rank $s-1$.

We start by setting $j \leftarrow j-1$. In the next blocks we perform steps necessary for deciding if to decrement also the total signal-subspace rank s.

10. **Termination condition block**

    If the anomaly subspace rank has reached 0, then terminate. Otherwise, continue.

11. **Calculation of $\mathbf{\Psi}_{s-1-j}$ corresponding to a reduced-rank signal-subspace**

    In order to decide if decrementing j should also entail the decrementing of the total signal-subspace s, one has to obtain the reduced-rank subspace $[\mathbf{\Psi}_{s-1-j}|\mathbf{\Omega}_j]$ and test the corresponding data residuals. Therefore, in this block, we calculate $\mathbf{\Psi}_{s-1-j}$ by:

$$\mathbf{\Psi}_{s-1-j} = \underset{s-1-j}{\mathrm{SVD}}\, \mathcal{P}_{\mathbf{\Omega}_j^{\perp}}\mathbf{X}. \tag{3.3}$$

12. **Calculation of maximum data residual-norm η_{s-1} in the obtained residual-subspace**

$$\eta_{s-1} = \max_{\mathbf{x}\in\text{cols }\mathbf{X}} \|\mathcal{P}_{[\mathbf{\Psi}_{s-1-j}|\mathbf{\Omega}_j]^{\perp}}\mathbf{x}\|^2 \tag{3.4}$$

13. **Performing noise-related hypothesis testing of η_{s-1}**

    In this block we assess if η_{s-1} contains signal-contribution. For this purpose we apply the test of equation (2.23).

14. **Decision if to reduce the total signal-subspace rank s**

    If η_{s-1} meets the noise-hypothesis, meaning that the subspace $(range\ [\mathbf{\Psi}_{s-1-j}|\mathbf{\Omega}_j])^{\perp}$ doesn't contain signal contributions (i.e., the basis $[\mathbf{\Psi}_{s-1-j}|\mathbf{\Omega}_j]$ represents well the signal-subspace), then s should be decremented. Otherwise, leave s intact and continue to a new iteration.

15. **Decrementing the total signal-subspace rank s**

$$s \leftarrow s - 1, \tag{3.5}$$

    and continue to a new iteration at block (2).

Comments

1. Once the new value of s is determined, we approach a nominal state (at block (2)), where the anomaly vectors matrix rank is decremented by 1, and the signal-subspace basis $[\mathbf{\Psi}_{s-j}|\mathbf{\Omega}_j]$ (with the updated values of j and s) is "MOCA-optimal" with respect to the modified data-matrix $\mathbf{X}$. In order to extract other anomaly vectors, corresponding to the rest of $\mathbf{\Omega}_j$ columns, until the complete depletion of all anomaly vectors, steps 2 - 15 are repeated. The iterations stop when there are no more columns in the anomaly-basis matrix $\mathbf{\Omega}_j$, i.e., $j = 0$.

2. It is important to note that at the end of the AXDA procedure, the signal-subspace basis is composed solely of $\mathbf{\Psi}_s$ ($s \leq \hat{r}$), which constitutes the

MOCA-optimal basis of the background vectors. So the AXDA algorithm equips us also with a anomaly-free (in other words "robust") estimated background-subspace and rank.

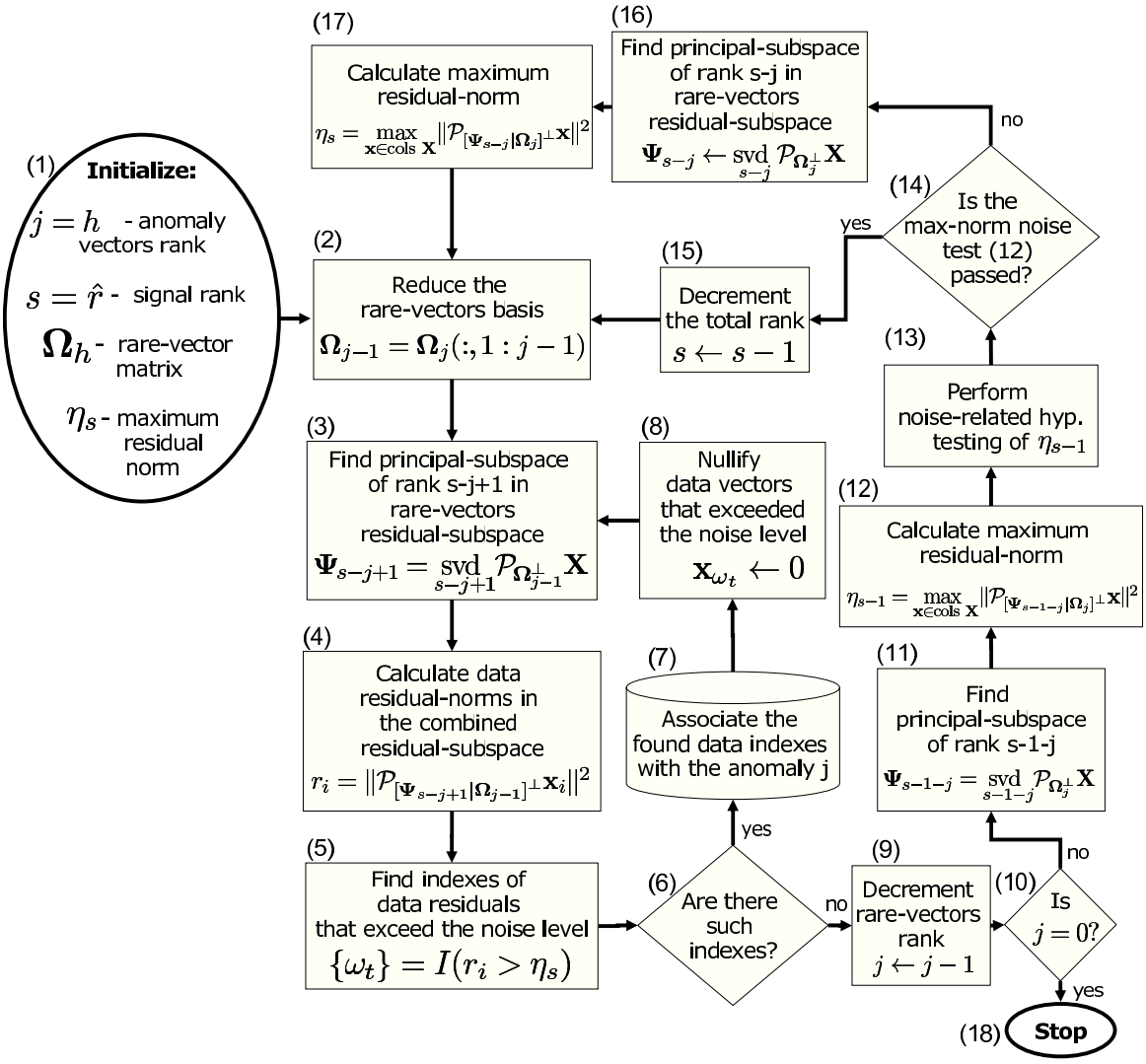


Figure 3.2: **Detailed description of Anomaly Extraction and Discrimination Algorithm (AXDA).** The notation in block (2) is MATLAB® notation.

3.3 Experiments with Real Hyperspectral Data

In this section we evaluate performance of the MOCA algorithm followed by AXDA postprocessing applying them on real hyperspectral data. For an analysis of the effect of noise on MOCA, using self designed synthetic data experiments with different signal to noise ratios, the reader is referred to chapter 2.

To demonstrate the results, the proposed approach was applied to 6 real hyperspectral image cubes collected by an AISA airborne sensor configured to 65 spectral bands, uniformly covering VNIR range of $400nm$ - $1000nm$ wavelengths. At 4 km altitude pixel resolution corresponds to $(0.8m)^2$. The obtained image cubes are $b \times r \times c = 65 \times 300 \times 479$ hyperspectral images, where b,r and c denote the number of hyperspectral bands, the number of rows and the number of columns in the image, respectively.

In Fig. 5.2 one can see results of anomaly detection and discrimination. Shown are images containing the 30th-band of 4 different hyperspectral cubes with different terrain types. The 5th and 6th images are not shown here just because of convenience of placing an even number of images in the figure. The left 4 images contain ground-truth anomalies (marked in white and encircled by red ellipses), which were manually identified using side information collected from high resolution RGB images of the corresponding scenes. In Fig. 3.4, we show one of RGB images used for identifying the ground-truth anomalies. The right 4 images contain anomalies (marked in color) detected by AXDA, overlayed on the white ground-truth pixels. All anomaly pixels of the same type are marked by the same color. There are no missed anomalies in the presented 4 images. The corresponding dimensionality results obtained by MOCA and AXDA are separately summarized in Table. 3.1, where $\hat{r}$ is the signal subspace rank determined by MOCA, h is the anomaly dimensionality, s is the dimensionality of anomaly-free background. Note, that according to the discussion in step 9 of the AXDA algorithm presented in the section 3.2, it is possible that $s \geq \hat{r} - h$. Thus, AXDA allows discrimination of anomalies according to corresponding anomaly endmembers (constituent materials spectra) found by MOCA

As it was noted above, in Fig. 3.4, we show one of the RGB images used for identifying the ground-truth anomalies. The scene under consideration is shown

Table 3.1:

No. image	$\hat{r}$	h	s
1	10	2	10
2	15	9	11
3	10	5	8
4	16	8	12
5	15	7	13
6	11	4	10

in a high-resolution (2672×4000) color RGB-image. The ground-truth anomalies are encircled by red ellipses. As it can be seen, the detected anomalies correspond to vehicles and small agriculture facilities, which occupy a few pixel segments.

In Fig. 3.5, we compare between GMRX [38], MSD [30] and the proposed AXDA in terms of Receiver Operation Characteristic (ROC) curves. For the purpose of ROC curves generation, 6 hyperspectral images were used, in which the total number of anomaly segments count is 25.

An anomaly is considered as detected if at least one of the detected pixels hits the corresponding marked segment. All pixels detected by the algorithms were grouped into connected objects using 8-connected object labelling. If an object doesn't intersect a marked anomaly, it is considered a false alarm object. This kind of anomaly detection/miss criteria is particularly suitable for applications that aim to *alert* the user on all anomalies of all sizes. Therefore, it is more important to detect at least one pixel on each anomaly, rather than many pixels on only some of the anomalies.

In order to obtain multiple operating points for AXDA, an additional parameter should be introduced to the proposed algorithm. A reasonable place for such a parameter is in the noise hypothesis relation in (2.23). However, due to special characteristics of maximum-norm noise distribution, which is very narrow - almost deterministic (see chapter 2), any factor introduced to this relation would result in almost the same decision. Thus, AXDA has naturally a single(nominal) operating point dictated by the noise statistical properties.

Yet, for the sake of comparison, we've introduced a rather compelling param-

eter γ to the equation of block (5) in Fig. 3.2, which now reads as:

$$\{\omega_t\} = I(r_i > \gamma\eta_s). \tag{3.6}$$

In words, the noise-related threshold value η_s (measured in a previous iteration) is multiplied by the factor γ in order to produce a new threshold value. The lower the factor γ is, the more data vectors will be treated as anomaly-vectors and be associated to the dropped column j of $\mathbf{\Omega}_j$. In our simulation, we have used 30 values of γ, which were uniformly sampled from $[0.8, 1.2]$. The position of nominal operating point of AXDA (for $\gamma = 1$) is pointed out by a red arrow. As can be seen from the figure, the nominal operating point provides a high detection rate (24 detected anomalies) with a significantly low false alarm rate (6 false alarm segments).

The GMRX algorithm was initialized by an excessive number of Gaussians using the k-means algorithm for initializing the Gaussian parameters. During the EM iterations of the GMRX, too small clusters, and hence unreliable, were eliminated. In Fig. 3.6 one can see results of the GMRX algorithm, applied to the same 4 hyperspectral cubes as AXDA, with a GLR parameter producing the same false alarm rate as AXDA at the nominal operating point (which equals to 6 false alarm segments). As in Fig. 5.2, The left 4 images contain manually identified ground-truth anomalies (marked in white and encircled by red ellipses), whereas the right 4 images contain anomalies (marked in red), detected by GMRX, overlayed on the white ground-truth pixels. The missed targets are encircled by cyan ellipses.

The MSD algorithm was provided an anomaly-free estimation of the background basis $\mathbf{\Psi}_s$ estimated by AXDA, which uses the anomaly subspace basis $\mathbf{\Omega}_h$ provided by MOCA, since MOCA and AXDA combined are unique in their ability to perform an unsupervised determination of both anomaly and background subspaces and their ranks.

Fig. 3.5 clearly shows that for the examined images AXDA has a better performance than GMRX and MSD, in most of the range of the tested parameters. It is also important to note, that in contrast to MSD and GMRX, AXDA allows an unsupervised determination of the nominal operating point, determined by

maximum-norm noise statistical properties. Moreover, AXDA has an ability to discriminate between different types of anomalies.

3.4 Summary

In this chapter we have proposed an algorithm for anomaly detection, discrimination and population estimation of anomalies of the same type, called AXDA. The algorithm is based on a signal-subspace and rank estimation provided by MOCA chapter 2. By its construction, the signal basis consists of two groups of basis vectors. One group spans the subspace of anomalies. The second group is designed to represent background pixel residuals belonging to the subspace that is complementary to the subspace of the anomalies. The proposed AXDA extracts anomaly pixels by removing an anomaly basis vector from the anomaly vectors group and compensating for its removal by augmenting the background vectors related subspace. This operation causes a violation of the noise hypothesis condition in vectors that are highly correlated with the removed anomaly basis vector. Such vectors are detected, associated with the removed basis vector, and depleted from the data. This way we obtain groups of data vectors associated with each one of the anomaly basis vectors.

In experiments with real hyperspectral image cubes AXDA was shown to have a better performance than GMRX and MSD, in most of the range of the tested parameters. Since the anomaly and background subspaces are unknown in advance, the MSD algorithm was provided the anomaly-free estimation of the background basis $\mathbf{\Psi}_s$ obtained from AXDA and the anomaly subspace obtained from MOCA. This provides MSD subspace-related information that is (at least) as good as AXDA has for the detection of anomalies. It is also important to note, that in contrast to MSD and GMRX, AXDA is equipped with an unsupervised determination of the nominal operating point. AXDA also has a capability to discriminate between different types of anomalies, though the accuracy of this discrimination, as well as the accuracy of population estimation of anomalies of the same type, are not evaluated in this research and may be a subject for future research. Moreover, AXDA allows also an anomaly-free (robust) estimation of the background-subspace and rank.

It turns out now that MOCA in combination with AXDA provide means to meet a wide range of signal-subspace estimation scenarios:

1. *Estimation of a signal-subspace that includes anomaly-vectors.*
2. *Detection of anomaly-vectors and determination of their subspace.*
3. *Providing a natural (nominal) operating point for anomaly detection.*
4. *Estimation of a pure (free of outliers) background-subspace.*

Figure 3.3: **AXDA results at the *nominal* operating point.** The left 4 images contain manually identified ground-truth anomalies (marked in white and encircled by red ellipses). The right 4 images contain anomalies (marked in color) detected by AXDA, overlayed on the white ground-truth pixels. There are no missed anomalies in the presented 4 images. All anomaly pixels of the same type are marked by the same color.

Figure 3.4: **High resolution RGB image of the analyzed scene,** used as a ground-truth indication for AXDA results verification. The ground-truth anomalies are encircled by red ellipses.

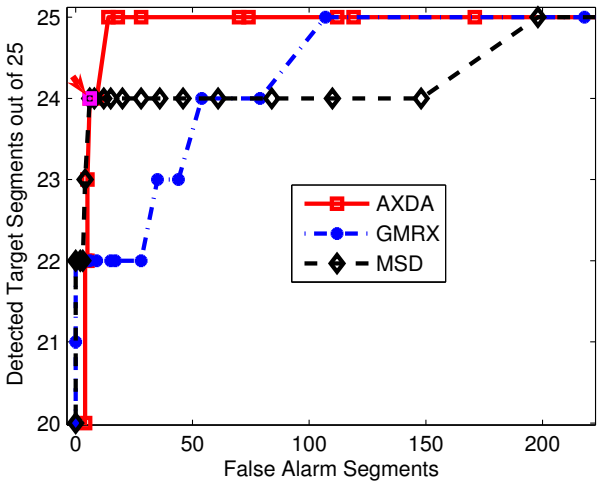


Figure 3.5: **ROC curves corresponding to GMRX, MSD and AXDA.** The nominal operating point of AXDA is marked in magenta color and is pointed out by the arrow. This point corresponds to 24 detected anomalies and 6 false alarm segments.

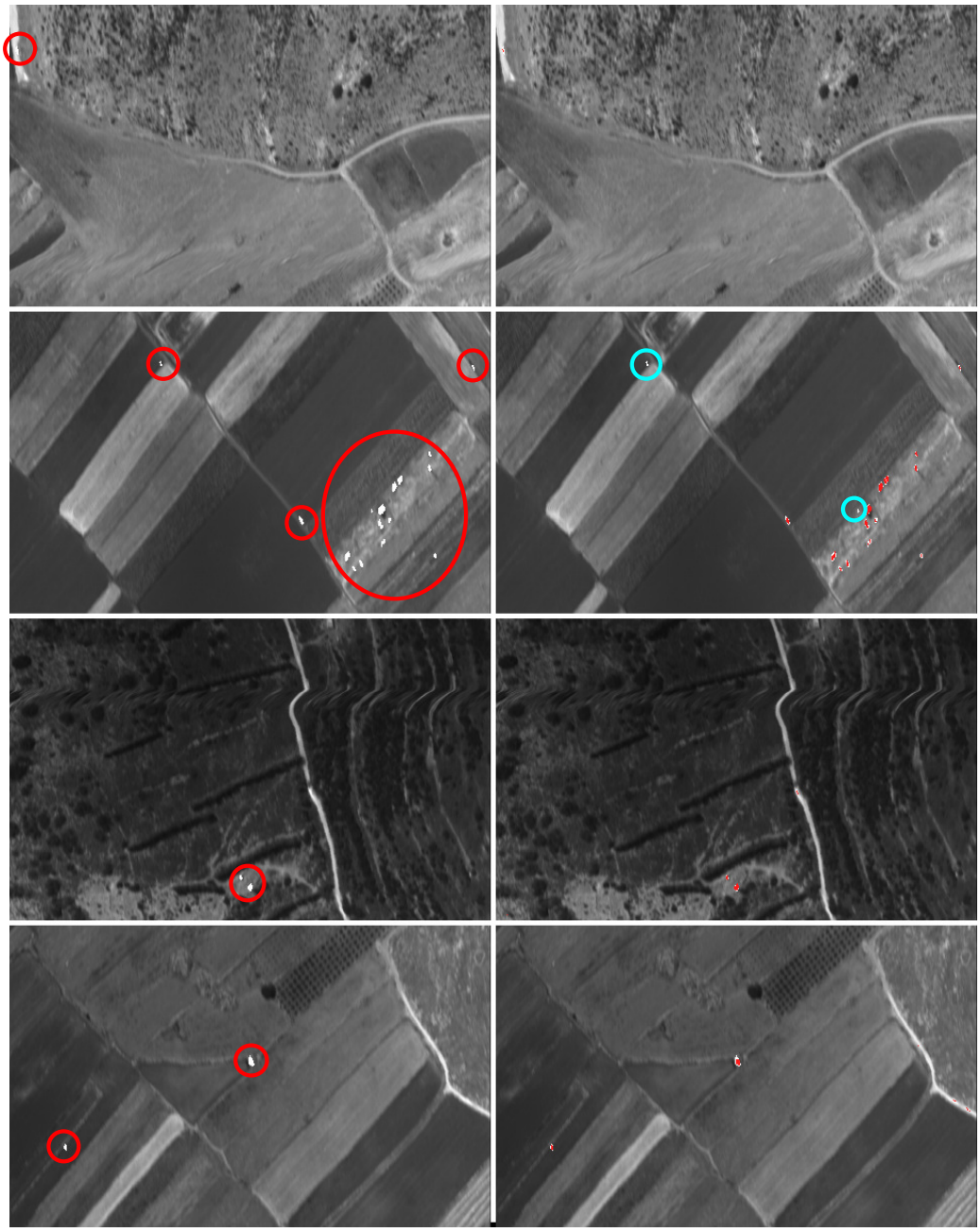


Figure 3.6: **GMRX Anomaly Detection Results for GLRT parameter producing the same false alarm rate as AXDA at its nominal operating point.** The left 4 images contain manually identified ground-truth anomalies (marked in white and encircled by red ellipses). The right 4 images contain anomalies (marked in red) detected by GMRX, overlayed on the white ground-truth pixels. Missed anomalies are encircled by cyan ellipses.

Chapter 4

$\ell_{2,\infty}$-Optimal Subspace Estimation

In this chapter we propose an optimal algorithm for the signal-subspace estimation that utilizes a natural conjugate gradient learning approach proposed in [62] to minimize $\ell_{2,\infty}$-norm of the misrepresentation residuals. During the minimization process, the signal-subspace basis matrix is constrained to the Grassmann manifold defined as the set of all n dimensional subspaces in $\mathbb{R}^m$, $n \leq m$ [62]. Since $\ell_{2,\infty}$-norm of the misrepresentation residuals can be also referenced as the maximum orthogonal complement norm, we denote the proposed algorithm as Maximum of *Orthogonal complements Optimal Subspace Estimation* (MOOSE).

4.1 Minimizing $\ell_{2,\infty}$-norm on the Grassmann manifold

4.1.1 Problem formulation

Generally, the problem stated in (2.8) can be recast as

$$\hat{\mathbb{S}} = \underset{[\mathbf{W}]}{\operatorname{argmin}}\, F([\mathbf{W}]), \tag{4.1}$$

where the objective function $F([\mathbf{W}])$ is defines as

$$F([\mathbf{W}]) \triangleq \|\mathbf{W}^\top \mathbf{X}\|_{2,\infty}^2 \tag{4.2}$$

and $[\mathbf{W}]$ is an equivalence class of all $p\times(p-k)$ orthogonal matrices whose columns span the same subspace in $\mathbb{R}^p$ as $\mathbf{W}$. Here $[\mathbf{W}]$ represents the orthogonal complement subspace to the sought signal-subspace $\mathcal{S}_k$. The set of all n-dimensional subspaces in $\mathbb{R}^m$, denoted by $G_{m,n}$, is called the Grassmann manifold [62]. The geometrical structure of the Grassmann manifold allows a continuous choice of subspaces, which is essential for constructing a local minimization procedure. Without loss of generality, by necessity, we must pick a representative of the equivalence class $[\mathbf{W}]$, say $\mathbf{W}$, in order to be able to work with $[\mathbf{W}]$ on the computer. Thus, by smoothly changing $\mathbf{W}$, such that $[\mathbf{W}] \in G_{p,p-k}$ we would be able to continuously move from one subspace to another and iteratively improve the objective function in a maner similar to well known unconstrained gradient-based algorithms such as steepest descent and conjugate gradient [66].

4.1.2 Grassmann manifold geometry

As stated in [62], the benefits of using gradient-based algorithms for the unconstrained minimization of an objective function can be carried over to a minimization constrained to the Grassmann manifold. The familiar operations employed by unconstrained minimization in the Euclidean space (plain space) such as computing gradients, performing line searches, etc., can be traslated into their covariant versions on the Grassmann manifold (curved space).

In the following we briefly outline basic results from [62] used in this work for calculating gradients of an objective function and performing a line search along a search direction on the Grassmann manifold. Then, we develop a technique for minimizing $F([\mathbf{W}])$ of (4.2).

4.1.2.1 Gradient on Grassmann

The gradient of the objective function $F([\mathbf{W}])$ on the Grassmann manifold is defined to be a matrix $\nabla F \in T_{[\mathbf{W}]}$, where $T_{[\mathbf{W}]}$ is the tangent space at $[\mathbf{W}]$, such that for all $\mathbf{T} \in T_{[\mathbf{W}]}$, the following holds:

$$\langle F_{\mathbf{W}}, \mathbf{T}\rangle = \langle \nabla F, \mathbf{T}\rangle\,, \tag{4.3}$$

where $F_{\mathbf{W}}$ is the $p \times (p-k)$ matrix of partial derivatives of F with respect to the elements of $\mathbf{W}$; $\langle \cdot, \cdot \rangle$ denotes the standard inner product in $p \times (p-k)$ - dimensional Euclidean space defined as

$$\langle \triangle_1, \triangle_2 \rangle \triangleq \mathrm{tr}(\triangle_1^\top \triangle_2). \tag{4.4}$$

In words, the relation in (4.3) states that the gradient of $F([\mathbf{W}])$ on the Grassmann manifold is the projection of $F_{\mathbf{W}}$ onto $T_{[\mathbf{W}]}$. Since $T_{[\mathbf{W}]}$ is the set of subspaces spanned by the columns of matrices of the form

$$\mathbf{T} = \mathbf{W}_\perp \mathbf{B}, \tag{4.5}$$

where $\mathbf{B}$ are arbitrary $k \times k$ matrices and $\mathbf{W}_\perp$ is a $p \times k$ orthogonal matrix satisfying

$$\mathbf{W}\mathbf{W}^\top + \mathbf{W}_\perp \mathbf{W}_\perp^\top = \mathbf{I}, \tag{4.6}$$

one obtains

$$\nabla F = F_{\mathbf{W}} - \mathbf{W}\mathbf{W}^\top F_{\mathbf{W}}. \tag{4.7}$$

A more rigorous treatment of these intuitive concepts is given in [62] where a solid foundation framework for the optimization algorithms involving orthogonality constraints is developed.

4.1.2.2 Line search

The line search in the Grassmann manifold is defined to be the minimization of $F([\mathbf{W}])$ along a geodesic, which is the curve of shortest length between two points in a manifold. By noticing that the geodesic equation is a second-order ODE, it follows from the local existence and uniqueness theorem that for any point $\mathbf{p}$ in a manifold and for any vector $\mathbf{v}$ in the tangent space at $\mathbf{p}$, there exists a unique geodesic curve passing through $\mathbf{p}$ in the direction $\mathbf{v}$ [63]. This observation makes the generalization of local optimization methods straightforward: given a descent direction $\mathbf{H} \in T_{[\mathbf{W}]}$ (for example, $\mathbf{H} = -\nabla F$), the objective function $F([\mathbf{W}])$ is minimized by the line search along the geodesic passing through $[\mathbf{W}]$ in the direction $\mathbf{H}$. An easy to compute formula for geodesics on the Grassmann

manifold proposed in [62] reads as:

$$\mathbf{W}(t) = (\mathbf{W}\mathbf{V} \;\; \mathbf{U}) \begin{pmatrix} \cos(t\mathbf{\Sigma}) \\ \sin(t\mathbf{\Sigma}) \end{pmatrix} \mathbf{V}^\top, \tag{4.8}$$

where t is a geodesic curve traversing parameter and $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ is the compact singular value decomposition (SVD) of $\mathbf{H}$. Compact SVD here means that the zero singuar values are discarded along with the respective columns in $\mathbf{U}$ and $\mathbf{V}$, and the signalur values are set in a decreasing order in $\mathbf{\Sigma}$. It can be easily verified that the diagonal elements of the matrix $t\mathbf{\Sigma}$ traverse Principal angles [69] between the column spaces $[\mathbf{W}(t)]$ and $[\mathbf{W}]$. Thus, for $t = 0$, one obtains the original subspace $[\mathbf{W}]$ that is rotated by the angles $t\mathbf{\Sigma}$ when t increases. Moreover, the geodesic distance between $[\mathbf{W}(t)]$ and $[\mathbf{W}]$ on the Grassmann manifold denoted by $d([\mathbf{W}(t)], [\mathbf{W}])$ satisfies [62]:

$$d([\mathbf{W}(t)], [\mathbf{W}]) = t\sqrt{\mathrm{tr}(\mathbf{\Sigma}^2)}. \tag{4.9}$$

It should be noted that for large t values, the distance $d([\mathbf{W}(t)], [\mathbf{W}])$ is not the shortest one between $[\mathbf{W}]$ and $[\mathbf{W}(t)]$, since for large t, $[\mathbf{W}(t)]$ may complete one or more full circles in terms of the angles on the diagonal of $t\mathbf{\Sigma}$. However, it is still true that locally, for small t increments, $[\mathbf{W}(t)]$ is the shortest path on the Grassmann manifold connecting points on it. Moreover, the relation (4.9) implies that the rotation velocity, when one traverses the geodesic $[\mathbf{W}(t)]$ by changing t, equals to $\sqrt{\mathrm{tr}(\mathbf{\Sigma}^2)}$ and, therefore, may change from iteration to iteration. In order to make it constant during the line search for all iterations, the matrix $\mathbf{\Sigma}$ is normalized:

$$\tilde{\mathbf{\Sigma}} \triangleq \mathbf{\Sigma} \big/ \sqrt{\mathrm{tr}(\mathbf{\Sigma}^2)}. \tag{4.10}$$

Now, the line search is perfomed by looking for t that corresponds to a “significant reduction" of the objective function along a geodesic $[\mathbf{W}(t)]$. The notion of “a significant reduction" means that, on one hand, t should be low enough to ensure reduction of the objective function value; on the other hand, the search step t should be large enough for fast algorithm convergence. For this purpose, we use the Backtracking-Armijo linesearch method [66], [64] summarized in Algorithm 1.

Algorithm 1 *Backtracking-Armijo line search.*

Given a geodesic $[\mathbf{W}(t)]$ in a descending direction $\mathbf{H}$, $\alpha \in (0, 0.5)$, $\beta > 1$, $t := t_0$
Backtracking:
while $(F([\mathbf{W}(t)]) > F([\mathbf{W}]) + \alpha\, t\, \langle \nabla F, \mathbf{H} \rangle)$, $t := t/\beta$
Armijo:
while $(F([\mathbf{W}(t)]) \leq F([\mathbf{W}]) + \alpha\, t\, \langle \nabla F, \mathbf{H} \rangle)$ **and** $(F([\mathbf{W}(\beta\, t)]) < F([\mathbf{W}]) + \alpha\, \beta\, t\, \langle \nabla F, \mathbf{H} \rangle)$, $t := \beta t$

In words, if the value of t is too large, it is iteratively decreased by dividing it by β in the Backtracking "while" stage, until the following condition holds:

$$F([\mathbf{W}(t)]) \leq F([\mathbf{W}]) + \alpha\, t\, \langle \nabla F, \mathbf{H} \rangle\,. \tag{4.11}$$

Since $\mathbf{H}$ is a descent direction and $\alpha < 1$, we have $\langle \nabla F, \mathbf{H} \rangle < 0$, so for small enough t, the following holds:

$$\begin{aligned} &F([\mathbf{W}(t)]) \approx F([\mathbf{W}]) + t\, \langle \nabla F, \mathbf{H} \rangle \leq \\ &\leq F([\mathbf{W}]) + \alpha\, t\, \langle \nabla F, \mathbf{H} \rangle \leq \\ &\leq F([\mathbf{W}]), \end{aligned} \tag{4.12}$$

which shows that the Backtracking "while" expression eventually terminates and that t is small enough to cause a decrease of the objective function value.

If the value of t is too small, it is iteratively increased by mutiplying it by β in the Armijo "while" stage, until the condition (4.11) is concurrently satisfied with:

$$F([\mathbf{W}(\beta\, t)]) \geq F([\mathbf{W}]) + \alpha\, \beta\, t\, \langle \nabla F, \mathbf{H} \rangle\,. \tag{4.13}$$

In words, t is increased until it reaches a point in which it is still small enough to satisfy condition (4.11), but already large enough so that it is no longer satisfied in the next iteration, i.e., when βt replaces t (see (4.13)).

4.1.3 Minimization of $F([\mathbf{W}])$ on the Grassmann manifold.

In this subsection we develop a technique for solving (4.1) for $F([\mathbf{W}])$ of (4.2) on the Grassman manifold. A natural choice for the search direction is the negative gradient $\mathbf{H} = -\nabla F$ [65]. The calculation of ∇F involves the calculation of $F_{\mathbf{W}}$ (see (4.7)). For the calculation of $F_{\mathbf{W}}$ we consider here two cases: One case is when the maximum is obtained for only one data vector, while the other case is when the maximum is obtained for more than one data vector.

Case 1. If the maximum is obtained for only one vector at each $\mathbf{W}$ througout the minization, the calculation of $F_{\mathbf{W}}$ becomes straightforward:

$$F_{\mathbf{W}} = \mathbf{x}_j \mathbf{x}_j^\top \mathbf{W}, \tag{4.14}$$

where $\mathbf{x}_j$ is the vector for which $\max_{i=1,\ldots,N} \|\mathbf{W}^\top \mathbf{x}_i\|_2$ is obtained.

Case 2. If the maximum is obtained for a set of indices J that containes more than one index, then the gradient direction $\hat{\mathbf{G}} = F_{\mathbf{W}} / \|F_{\mathbf{W}}\|_2$ is given by solving the following problem:

$$\begin{aligned} \hat{\mathbf{G}} \quad = \quad & \max_{\mathbf{G}} \min_{j \in J} \left\langle \mathbf{G}, \mathbf{x}_j \mathbf{x}_j^\top \mathbf{W} \right\rangle \\ \text{s.t.} \quad & \left\langle \mathbf{G}, \mathbf{x}_j \mathbf{x}_j^\top \mathbf{W} \right\rangle > 0 \ \ \forall j \in J \\ & \langle \mathbf{G}, \mathbf{G} \rangle = 1, \end{aligned} \tag{4.15}$$

with $\langle \cdot, \cdot \rangle$ being defined in (4.4). In words, it is a unit-norm matrix that maximizes the minimal projection norm onto gradients obtained individually for each $\mathbf{x}_j$, $j \in J$ (as in (4.14)). If the problem (4.15) is feasible, then the direction $-\hat{\mathbf{G}}$ is guaranteed to be a descent direction for all maximal residual norms $\|\mathbf{W}^\top \mathbf{x}_j\|$, $j \in J$, since all projections are constrained to be positive. Moreover, it is the steepest descent direction of the objective function $F([\mathbf{W}])$, because the descent rate of $F([\mathbf{W}])$ is determined by the lowest descent rate of the maximal residual norm $\|\mathbf{W}^\top \mathbf{x}_j\|$, for some $j \in J$, which is maximized (see the problem formulation in (4.15)). If the problem is infeasible, then $[\mathbf{W}]$ is a local minimum of the objective function $F([\mathbf{W}])$, since there is no search direction that concurrently minimizes all maximal residual norms. The problem (4.15) can be efficiently solved by Second-Order Cone Programming (SOCP) [66]. The norm of the derivative matrix $\|F_{\mathbf{W}}\|$

is given by

$$\|F_{\mathbf{W}}\| = \min_{j \in J} \left\langle \hat{\mathbf{G}}, \mathbf{x}_j \mathbf{x}_j^\top \mathbf{W} \right\rangle, \tag{4.16}$$

I.e., it equals to the lowest descent rate of the the maximal residual norms, or equivalently, to the descent rate of $F([\mathbf{W}])$ in the direction $\hat{\mathbf{G}}$.

Practically, we have observed that in real data distributions the maximum is obtained for only one vector with probability close to one. Therefore, using (4.14) is good enough (practically) for obtaining a steep descent direction as we did in our simulations.

In order to better cope with the complex nature of the cost function $F([\mathbf{W}])$, we propose to use the conjugate gradient method. According to this method, the conjugate search direction is a combination of the previous search direction and the new gradient

$$\mathbf{H}_s = -\nabla F_s + \gamma_s \tilde{\mathbf{H}}_{s-1}, \tag{4.17}$$

where s denotes the iteration index, $\tilde{\mathbf{H}}_{s-1}$ is the parallel translation of the previous search direction $\mathbf{H}_{s-1}$ from the point $[\mathbf{W}_{s-1}]$ to $[\mathbf{W}_s]$ by removing its normal component to the tangent space $\mathbf{T}_{\mathbf{W}s+1}$, as schematically shown in Fig. 4.1; and γ_s is obtained via Polak Ribiére conjugacy condition formula [62]

$$\gamma_s = \left\langle \nabla F_s - \tilde{\nabla} F_{s-1}, \nabla F_s \right\rangle \Big/ \left\langle \nabla F_{s-1}, \nabla F_{s-1} \right\rangle, \tag{4.18}$$

where $\tilde{\nabla} F_{s-1}$ is the parallel translation of ∇F_{s-1} obtained in the same way as $\tilde{\mathbf{H}}_{s-1}$. The parallel translation is needed in order to keep all directions within the tangent space at each iteration. The formula for obtaining $\tilde{\nabla} F_{s-1}$ and $\tilde{\mathbf{H}}_{s-1}$ is [62]:

$$\begin{aligned} \tilde{\mathbf{H}}_{s-1} &= (-\mathbf{W}_{s-1}\mathbf{V}\sin(t\tilde{\mathbf{\Sigma}}) + \mathbf{U}\cos(t\tilde{\mathbf{\Sigma}}))\mathbf{\Sigma}\mathbf{V}^\top \\ \tilde{\nabla} F_{s-1} &= \nabla F_{s-1} - \\ &\quad (\mathbf{W}_{s-1}\mathbf{V}\sin(t\tilde{\mathbf{\Sigma}}) + \mathbf{U}(\mathbf{I} - \cos(t\tilde{\mathbf{\Sigma}}))\mathbf{U}^\top \nabla F_{s-1}. \end{aligned} \tag{4.19}$$

The conjugate gradient construction offers a good compromise between convergence speed and computational complexity [67]. If the objective function is nondegenerate, then the algorithm is guaranteed to converge quadratically [68].

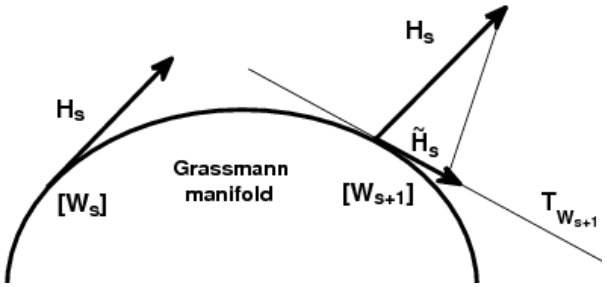


Figure 4.1: **Parallel transport on Grassman manifold.**

In our problem, the contribution of the previous search direction in each iteration, also helps the procedure to employ previous information carried in maximal norms obtained earlier (for possibly different data vectors). This prevents algorithm slow down due to the alternation of the maximum-norm data vectors.

As any local minimization of a non-convex objective function, the proposed algorithm is prone to getting trapped in a local minimum. Therefore, a proper initialization may be crucial for obtaining a good solution. Since MX-SVD finds a suboptimal solution using global principles, it provides a good initial point, which is close to the global minimum. Therefore, in our simulations we use the subspace obtained by MX-SVD as an initial point for the proposed approach.

The proposed approach for minimizing $F([\mathbf{W}])$ is summarized in Algorithm 2.

4.2 Synthetic data simulation results

In this section we compare the results of applying SVD, MX-SVD and MOOSE to simulated examples in the presence of anomaly vectors. For this purpose the input data is constructed as follows:

$$\mathbf{X} = \mathbf{Y} + \mathbf{Z}, \tag{4.20}$$

with

$$\mathbf{Y} = \left[\sqrt{SNR_b}\,\mathbf{B}\mathbf{S}_b \mid \sqrt{SNR_a}\,\mathbf{A}\mathbf{S}_a\right], \tag{4.21}$$

where $\mathbf{B}$ is a $p \times r_b$ matrix with orthogonal unit-norm columns spanning the background subspace; $\mathbf{A}$ is a $p \times r_a$ matrix with orthogonal unit-norm columns spanning the subspace of anomalies; $\mathbf{S}_b$ is a $r_b \times N_b$ matrix of background vector coefficients with columns drawn randomly from a Gaussian distribution with covariance matrix $\mathbf{C}_b = \mathbf{I}/r_b$; $\mathbf{S}_a$ is a $r_a \times N_a$ matrix of anomaly vector coefficients with columns drawn randomly from a Gaussian distribution and *normalized to have unit-norm*; and $\mathbf{Z}$ is a $p \times (N_a + N_b)$ matrix containing white Gaussian noise with variance equal to $1/p$.

For SNR defined as

$$SNR \triangleq E\{\|\mathbf{y}\|^2\}/E\{\|\mathbf{z}\|^2\}, \tag{4.22}$$

one can easily verify that background vectors have $SNR = SNR_b$, whereas the anomaly vectors have $SNR = SNR_a$. Moreover, due to the structure of the anomaly vector coefficient matrix $\mathbf{S}_a$, the norms of noise-free anomaly vectors are equal. This construction is designed to produce anomaly vectors that are equally significant.

Obviously, anomaly vectors are characterized by their low number compared to the number of background vectors, i.e., $N_a \ll N_b$. However, their number is allowed to be higher than the anomaly subspace dimension that they belong to, i.e., $N_a \geq r_a$. The extent of anomaly subspace population (loading) can be characterized by the loading ratio defined as follows:

$$R_a \triangleq N_a/r_a, \tag{4.23}$$

Thus, the minimal loading ratio $R_a = 1$ corresponds to the case where the number of anomalies is equal to the anomaly subspace rank. The larger the value of R_a is, the more anomaly vectors populate the anomaly subspace.

In our simulations we used the parameters shown in Table 4.1. It is important

Table 4.1: Maximum residual-norm simulation parameters

p	r_b	r_a	N_b	N_a	SNR_b	SNR_a
100	5	5	10^5	10	100	10

to note that all parameters were selected to reflect a typical situation in hyperspectral images. Thus, SNR_a and SNR_b were selected to satisfy $SNR_a < SNR_b$ since the anomaly and the background subspaces in hyperspectral images are not orthogonal and, therefore, the anomaly vectors have weak orthogonal components to the subspace of background vectors.

In Fig. 4.2 one can see empirical pdfs of the maximum-residual norm $\|\mathbf{W}^\top\mathbf{X}\|^2_{2,\infty}$ obtained via a Monte-Carlo simulation, where $\mathbf{X}$ was generated 1000 times. As mentioned in chapter 2, the estimated subspace by SVD may be skewed by noise in a way that completely misrepresents the anomaly vectors, since SVD uses ℓ_2 norm for penalizing the data misrepresenation, which is not sensitive to the anomaly-vector contributions. Hence, as clearly seen from the figure, the max-norm data residuals obtained by SVD (thick solid line) have high values which correspond to a poor representation of the anomaly vectors. It is also demonstrated in chapter 2 that for $R_a = 1$ MX-SVD yields

$$\|\mathbf{W}^\top\mathbf{X}\|^2_{2,\infty} \approx \|\mathbf{W}^\top\mathbf{Z}\|^2_{2,\infty}. \tag{4.24}$$

In words, the empirical distribution of the maximum data residual norm $\|\mathbf{W}^\top\mathbf{X}\|^2_{2,\infty}$ for $R_a = 1$ is very close to the distribution of the maximum residual norm of noise $\|\mathbf{W}^\top\mathbf{Z}\|^2_{2,\infty}$, which has a limiting distribution known as the Gumbel distribution [54] (plotted in thin solid line in Fig. 4.2). However, as seen in that figure, for $R_a > 1$ (in this simulation $R_a = 2$), MX-SVD produces max-norm data residuals (whose pdf is plotted in dashed line) that are higher than the max-norm noise residuals. This happens since MX-SVD estimates the anomaly subspace by directly selecting r_b anomalous vectors from the data that contain noise, which skews the resulting subspace. The result is significantly improved by applying the optimal approach which produces max-norm data residuals (whose pdf is plotted in dot-dashed line) with values that are even lower than one would obtain from the Gumbel distribution.

The paradox of such a “super-efficiency" of the optimal approach is explained as follows: On one hand, the Gumbel distribution approximation is valid for max-norm realizations of data vectors drawn from Gaussian distribution. On the other hand, the max-norm data residuals obtained by MOOSE stem no longer from a

Gaussian distribution, since they are minimized by MOOSE and, as a result, become lower than if the corresponding data vectors where randomly sampled from a Gaussian distribution.

In Fig. 4.3 we compare SVD, MX-SVD and the proposed MOOSE algorithm in terms of subspace estimation error. The subspace error used here is defined to be the largest principal angle $\angle\{\hat{\mathcal{S}}, \mathcal{S}\}$ defined as follows [70]:

$$\angle\{\hat{\mathcal{S}}, \mathcal{S}\} = \max_{\mathbf{u}\in\hat{\mathcal{S}}} \min_{\mathbf{v}\in\mathcal{S}} \angle\{\mathbf{u}, \mathbf{v}\}, \quad \mathbf{u} \neq 0, \mathbf{v} \neq 0, \tag{4.25}$$

where $\hat{\mathcal{S}}$ and $\mathcal{S}$ denote the estimated subspace and the original subspace used for the data generation, respectively. In our simulations, for each R_a value $\mathbf{X}$ was generated 50 times. The considered R_a values were sampled logarithmically in $[1, 40]$ as shown in Fig. 4.3. For each R_a value we plot the mean of the subspace estimation error values obtained by SVD (line with star marks), MX-SVD (line with circle marks) and the proposed approach (line with diamond marks). As clearly seen from the figure, the proposed approach corresponds to the lowest mean subspace estimation error for all R_a values. The MX-SVD and the proposed approach perform much better than SVD for a wide range of R_a values. For R_a values high enough SVD manages to catch up with the other two $\ell_{2,\infty}$-norm based approaches, since then the anomalies become significant in terms of the ℓ_2-norm.

4.3 Real data simulation results

In this section we compare the performance of SVD, MX-SVD and MOOSE when applied to 4 hyperspectral image cubes. The images were collected by an AISA airborne sensor [31] configured to 65 spectral bands, uniformly covering VNIR range of $400nm$ - $1000nm$ wavelengths. The obtained image cubes are $b \times r \times c = 65 \times 300 \times 479$ hyperspectral images, where b, r and c denote the number of hyperspectral bands, the number of rows and the number of columns in the image cube, respectively.

The assumed signal-subspace rank is $k = 10$. The only ground-truth information available for this evaluation were locations of man-made objects. In Fig. 4.4

are shown images of the 30th-band of each of the 4 image cubes used for the evaluation. The ground-truth anomalies, which are marked in white and encircled by red ellipses, were manually identified using side information collected from high resolution RGB images of the corresponding scenes. The ground truth anomalies consist of vehicles and small agriculture facilities, which occupy few-pixel segments.

Since the man-made objects are anomalous in these images, it is difficult to represent them with low error by employing the classical ℓ_2-norm based methods, we evaluate the anomaly-preserving algorithm performances in terms of the maximum residual norms obtained on the ground-truth anomalies. That is, the best algorithm should have the following property: once applied on a whole image cube, the $\ell_{2,\infty}$-norm of the ground-truth anomaly residuals and the $\ell_{2,\infty}$-norm of the whole image should be the lowest compared to the other algorithm results obtained in all image cubes. In other words, the better algorithm represents better not only all image pixels, but also the anomalous ones.

Thus, in Table 4.2 one can see that MOOSE has the lowest $\ell_{2,\infty}$-norm of image residuals and the lowest $\ell_{2,\infty}$-norm of the ground-truth anomalies in all examined images. SVD has the highest $\ell_{2,\infty}$-norms of image residuals and anomaly residuals that are equal in all images, which means that it poorly represents anomalies and that the worst-case error obtained by SVD in the whole image is on anomalies. The $\ell_{2,\infty}$-norms of image residuals and anomaly residuals obtained by MOOSE are different, meaning that the $\ell_{2,\infty}$-norms of image residuals are obtained on the background, i.e., the anomalies were represented even better than the background. The results of MX-SVD are much better than those of SVD and comparable to those of MOOSE meaning that practically, the greedy MX-SVD algorithm is a good choice, since it is more computationally efficient.

Table 4.2: Subspace estimation methods in terms of max. error norm

	Global $\ell_{2,\infty}$-norm of residuals			Anomaly $\ell_{2,\infty}$-norm of residuals		
Cube	SVD	MX-SVD	MOOSE	SVD	MX-SVD	MOOSE
1	200.6	98.3	97.3	200.6	82.7	81.7
2	1880.8	312.5	282.0	1880.8	312.5	207.8
3	453.0	98.5	73.6	453.0	84.1	70.9
4	749.6	445.6	401.2	749.6	445.6	383.6

4.4 Summary

In this chapter we have proposed an algorithm for dimensionality reduction of high-dimensional noisy data that preserves rare-vectors. The proposed algorithm is optimal in the sense that the estimated subspace (locally) minimizes the maximal-norm of misrepresentaion residuals. The optimization is performed via a natural conjugate gradient learning approach carried out on the set of n dimensional subspaces in $\mathbb{R}^m$, $m > n$, known as the Grassmann manifold. The proposed algorithm is denoted as *Maximum of Orthogonal complements Optimal Subspace Estimation* (MOOSE) and is the optimal version of a recently proposed greedy algorithm named *Min-Max-SVD* (MX-SVD). As any local minimization of a non-convex objective function, MOOSE is prone to getting trapped in a local minimum. Therefore, a proper initialization is crucial and is obtained by employing MX-SVD that uses global principles to find a suboptimal solution that is close to the global minimum. The results of MOOSE were compared to the results of SVD and MX-SVD by applying them both on simulated data and on real hyperspectral images. It was demonstrated that the results of MOOSE and MX-SVD are much better than those of SVD in terms of max-norm residual error, obtained in both simulated and real data, and in terms of the subspace estimation error obtained for simulated data. Although MX-SVD exhibits results inferior to those of MOOSE, the results of MX-SVD are quite comparable to those of MOOSE meaning that practically, the greedy MX-SVD algorithm is a good choice, since it is more computationally efficient.

Algorithm 2 *Conjugate gradient algorithm for minimizing $F([\mathbf{W}])$ on the Grassmann manifold.*

1 Given $\mathbf{W}_0$, such that $\mathbf{W}_0^\top \mathbf{W}_0 = \mathbf{I}_{p-k}$ and column space that coincides with the subspace obtained by MX-SVD, compute
$F_{\mathbf{W}_0} = \mathbf{x}_j \mathbf{x}_j^\top \mathbf{W}_0$, with j satisfying $\|\mathbf{W}_0^\top \mathbf{x}_j\|^2 = \|\mathbf{W}_0^\top \mathbf{X}\|_{2,\infty}^2$
$\nabla F_0 = F_{\mathbf{W}0} - \mathbf{W}_0 \mathbf{W}_0^\top F_{\mathbf{W}0}$ and set $\mathbf{H}_0 = -\nabla F_0$

2 For $s = 0, 1, \ldots,$

2.1 Obtain the compact decomposition of $\mathbf{H}_s$, $\mathbf{H}_s = \mathbf{U}\Sigma\mathbf{V}^\top$

2.2 Normalize the principal angles $\tilde{\mathbf{\Sigma}} = \mathbf{\Sigma} \diagup \sqrt{\mathrm{tr}\mathbf{\Sigma}^2}$

2.3 Perform Backtracking-Armijo line search (see Algorithm 1) along the geodesic
$\mathbf{W}(t) = \mathbf{W}_s \mathbf{V} \cos(t\tilde{\mathbf{\Sigma}}) \mathbf{V}^\top + \mathbf{U} \sin(t\tilde{\mathbf{\Sigma}}) \mathbf{V}^\top$

2.4 Update the subspace $\mathbf{W}_{s+1} = \mathbf{W}(t)$

2.5 Parallel transport the tangent vectors $\mathbf{H}_s$ and ∇F_s to the point $[\mathbf{W}_{s+1}]$
$\tilde{\mathbf{H}}_s = \left(-\mathbf{W}_s \mathbf{V} \sin(t\tilde{\mathbf{\Sigma}}) + \mathbf{U} \cos(t\tilde{\mathbf{\Sigma}})\right) \mathbf{\Sigma} \mathbf{V}^\top$
$\tilde{\nabla} F_s = \nabla F_s - \left(\mathbf{W}_s \mathbf{V} \sin(t\tilde{\mathbf{\Sigma}}) + \mathbf{U}(\mathbf{I} - \cos(t\tilde{\mathbf{\Sigma}})\right) \mathbf{U}^\top \nabla F_s$

2.6 Compute the new gradients
Euclidean: $F_{\mathbf{W}_{s+1}} = \mathbf{x}_j \mathbf{x}_j^\top \mathbf{W}_{s+1}$, with j satisfying $\|\mathbf{W}_{s+1}^\top \mathbf{x}_j\|^2 = \|\mathbf{W}_{s+1}^\top \mathbf{X}\|_{2,\infty}^2$
Grassmann: $\nabla F_{s+1} = F_{\mathbf{W}s+1} - \mathbf{W}_{s+1} \mathbf{W}_{s+1}^\top F_{\mathbf{W}s+1}$

2.7 Compute the new search direction via Polak Ribiére conjugacy condition formula
$\mathbf{H}_{s+1} = -\nabla F_{s+1} + \gamma_s \tilde{\mathbf{H}}_s$, where $\gamma_s = \left\langle \nabla F_{s+1} - \tilde{\nabla} F_s, \nabla F_{s+1} \right\rangle \diagup \langle \nabla F_s, \nabla F_s \rangle$

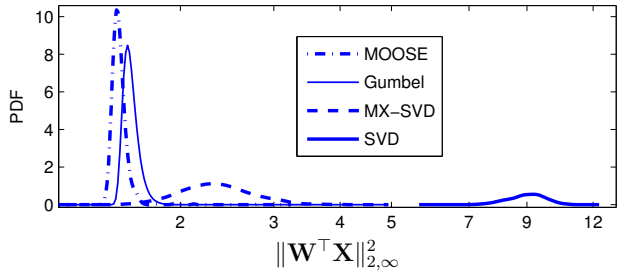


Figure 4.2: **The pdfs of $\|\mathbf{W}^\top\mathbf{X}\|^2_{2,\infty}$ obtained via Monte-Carlo simulation.** The empirical pdfs of $\|\mathbf{W}^\top\mathbf{X}\|^2_{2,\infty}$ obtained by SVD (thick solid line), MX-SVD (dashed line), MOOSE (dot-dashed line) and the limiting Gumbel distribution approximating maximum residual norm of noise (thin solid line).

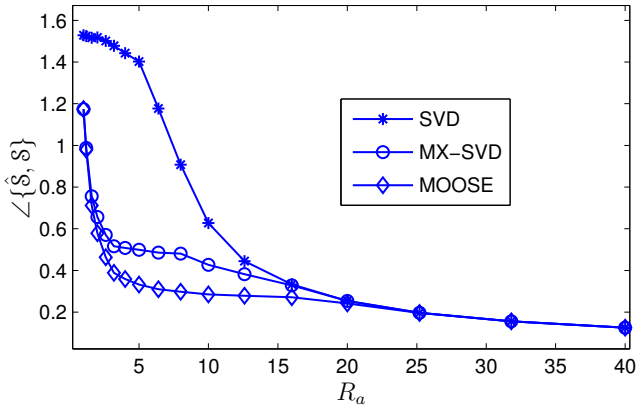


Figure 4.3: **Mean subspace error vs. anomaly loading ratio R_a for parameters of Table 4.1.** Mean-sample of the subspace error as a function of R_a obtained via a Monte-Carlo simulation using SVD (line with star marks), MX-SVD (line with circle marks), and MOOSE approach (line with diamond marks).

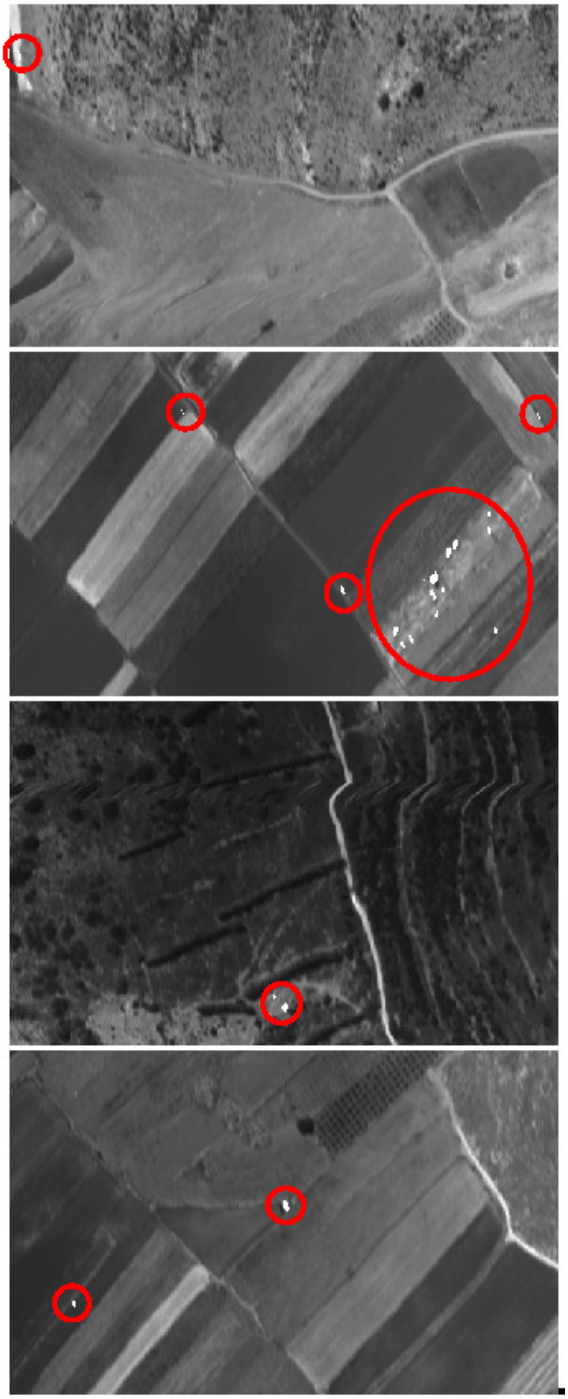


Figure 4.4: **Ground truth.** A 30th-band of each one of 4 image cubes used for evaluation. The ground-truth anomalies were manually identified, marked in white and encircled in red.

Chapter 5

Multispectral Filter Design for Anomaly Detection

In this chapter we propose a novel unsupervised technique for Designing Multispectral Filters that facilitates a better performance of local anomaly detection algorithms. The proposed approach is based on processing a sample hyperspectral image of a typical scene that is likely to be faced by anomaly detection algorithms. The sample image is not neccessarily required to include anomalies. Eventually, the problem of Multispectral Filters design may be formulated as a problem of Redundancy Reduction in Hyperspectral Channel channels, which is performed by replacing adjacent spectral bands by their means. This is a real-world Redundancy Reduction problem that requires preserving anomalies.

A common problem of local anomaly detection algorithms is so-called *Hughes phenomenon* [73], according to which the perfomance of anomaly detection algorithms significantly deteriorates when the number of pixels is severely limited for an accurate learning of the local background models. In order to alleviate the effect of *Hughes phenomenon*, one has to reduce the number of hyperspectral bands, since the complexity of background models is proportional to the hyperspectral data dimensionality.

The novel approach proposed here is based on a new criterion that is designed to retain spectral channels containing valuable anomaly-related information for anomaly detection algorithms. The optimal partition of the spectrum is obtained by *Minimizing the Maximal Mahalanobis Norm* of errors, obtained due to

the misrepresentaion of spectral intervals by constants. Therfore, we denote the proposed technique as Min-Max MN or, in short, MXMN. By minimizing the MXMN of errors, one reduces the anomaly contribution to the errors, which allows to retain more anomaly-related information in the reduced channels, if there are anomalies in the sample image. In the case that the sample scene does not contain anomalies, minimizing the MXMN of errors allows smoothing out spectral bands containing background clutter, which are unfavorable for the anomaly detection since they are likely to mask possible subtle anomaly contributions to other bands.

5.1 Anomaly Preserving Piecewise Constant Representation

5.1.1 Problem statement

Let $x_{i,j}$ denote the ith hyperspectral band of an observed hyperspectral pixel j, where $i = 1, \ldots, M$ and $j = 1, \ldots N$. The piecewise constant representaion model consists of a vector of $K < M$ breakpoints,

$$\mathbf{b}_K \triangleq \{b_1, \ldots, b_K\}, \tag{5.1}$$

corresponding to $K - 1$ contiguous intervals

$$I_k = [b_k, b_{k+1}), \qquad k = 1, \ldots, K - 1. \tag{5.2}$$

Each observed hyperspectral pixel $\mathbf{x}_j$ is approximated by a set of constants $\{\mu_{k,j}\}_{k=1}^{K-1}$ obtained by averaging its values in corresponding spectral intervals as follows:

$$\mu_{k,j} = \frac{1}{|I_k|} \sum_{i \in I_k} x_{i,j}, \tag{5.3}$$

where $|I_k|$ denotes the cardinality of the interval I_k. As a matter of fact, the constants $\{\mu_{k,j}\}$ minimize the mean squared error $S_{k,j}$ in each interval k defined

as follows:

$$S_{k,j} = \sum_{i \in I_k} (x_{i,j} - \mu_k)^2. \tag{5.4}$$

Thus, the partition of spectral bands into $K-1$ intervals by the breakpoints $\mathbf{b}_K$ uniquely determines the piecewise constant representation/approximation of each pixel. The goal is to determine a partition that facilitates good performance of anomaly detection algorithms when applied to the obtained constants $\{\mu_{k,j}\}$.

5.1.2 Objective function

The general idea of the proposed anomaly preserving channel reduction algorithm is to minimize an objective function $J(\mathbf{b}_K)$ that penalizes partitions which may potentially lead to the loss of anomalies during the channel reduction process. We choose the function $J(\mathbf{b}_K)$ to be of the following form:

$$J(\mathbf{b}_K) = \max_{k=1}^{K-1} D_k, \tag{5.5}$$

where by D_k we denote the Potential Anomaly Loss (PAL) measure correspoding to the interval I_k. Thus, by minimizing $J(\mathbf{b}_K)$, one minimizes the worst case PAL measure.

In order to properly define the PAL measure, D_k, let's explore statistical properties of the errors $e_{i,j,k}$ obtained due to the misrepresentation of hyperspectral pixel entries belonging to the interval I_k:

$$e_{i,j,k} = x_{i,j} - \mu_{k,j}, \qquad i \in I_k. \tag{5.6}$$

Denoting all error entries that belong to the same pixel j and correspond to an interval I_k, ordered in a vector form, by $\mathbf{e}_{j,k}$, we assume that all random vectors $\mathbf{e}_{j,k}$ corresponding to the non-anomalous (background) vectors are i.i.d. At this point, we observe that anomaly manifestations in an interval k, which were not represented well by the corresponding constants $\mu_{k,j}$, are likely to produce anomalous error realizations. Eventually, anomalous error realizations are those that do not agree well with the pdf of the background-related errors $\mathbf{e}_{j,k}$. Therefore, D_k, as a PAL measure, should measure the deviation of the obtained error statistics

from a background statistical model. Now, if one models the background-related errors $\mathbf{e}_{j,k}$ by a zero-mean Gaussian pdf, then D_k can be obtained by measuring the deviation of error realizations from the Gaussian model. This approach is quite reasonable, since the larger is the deviation of the error statistics from being Gaussian, the more signal structure is absorbed by the error and the larger is the likelihood that some important information is lost by channel reduction. A widely used criterion for anomaly detection is the Mahalanobis distance between a tested pixel and the background mean vector [72], [32]. This criterion has also been extensively used for assessing multivariate normality [78]. For a *zero mean* Gaussian random vector $\mathbf{e}$, the Mahalanobis distance or, equivalently, the Mahalanobis norm is defined as:

$$G(\mathbf{e}) \triangleq \sqrt{\mathbf{e}^\top \boldsymbol{\Sigma}^{-1} \mathbf{e}}, \tag{5.7}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of the random vector $\mathbf{e}$. Intuitively, the Mahalanobis norm of vectors $\mathbf{e}$ that contain outlying signal contributions and, therefore, are not properly normalized by $\boldsymbol{\Sigma}^{-1}$ in (5.7), is expected to be larger than obtained for vectors that obey the Gaussian paradigm. Thus, in the Reed-Xiaoli (RX) algorithm [32], a benchmark anomaly detector for hyperspectral imagery, the Mahalanobis distance is used to detect anomalies by comparing it to a threshold. It turns out that if the realizations $\mathbf{e}_j$ are contaminated by anomaly or other Non-Gaussian signal contributions, they are likely to produce large Mahalanobis norms. Therefore, we define D_k as follows:

$$D_k \triangleq \max_{j=1}^{N} G(\mathbf{e}_{j,k}) \tag{5.8}$$

This completes the definition of the objective function $J(\mathbf{b}_K)$ in (5.5) that penalizes partitions that may cause a PAL.

By evaluating the proposed approach with real data, we have observed that for obtaining a good partition, one does not nessecarily need to minimize $J(\mathbf{b}_K)$ over data *containing anomalies*. This important observation is further discussed in the section 5.2.

5.1.3 Minimizing the objective function

In order to minimize $J(\mathbf{b}_K)$, over the set of breakpoints $\{b_1, \ldots, b_K\}$, we apply a dynamic programming algorithm based on [77] and [79]. Let's redefine D_k as $D_{[g,h]}$, where g and h are interval boundaries which can be equivalently used to specify intervals instead of using their corresponding indices $\{k\}$. Throughout the minimization process, we iteratively calculate $J(k,p)$, where $J(k,p)$ is the objective function defined using only the first $1 < k \leq K$ breakpoints $\{b_1, \ldots, b_k\}$ and the first $(k-1) \leq p \leq (M-K+k)$ spectral bands.

Initially, we set

$$J(2,p) = D_{[1,p]}, \qquad p = 1, \ldots, (M-K+2). \tag{5.9}$$

Then, for $k = 3, \ldots, K$, we calculate $J(k,p)$ as follows:

$$J(k,p) = \min_{r=k-1}^{p-1} \left(J(k-1,r) + D_{[r+1,p]}\right). \tag{5.10}$$

At the end of the iterative process, the resulting $J(K,M)$ gives the optimal value of the objective function $J(\mathbf{b}_K)$ defined in (5.5). The optimal partition in terms of the breakspoints $\{b_1, \ldots, b_K\}$ is obtained by recursively backtracking the minimizers r^* for which the optimal sequence $\{J(K,M), J(K-1,r_K^*), \ldots, J(2,r_3^*)\}$ was obtained.

5.2 Experiments with Real Data

In this section we evaluate the performance of the RX algorithm, which, as mentioned, is a benchmark anomaly detector for Hyperspectral Imagery [32]. We applied it to Hyperspectral Data before and after the dimensionality reduction by PCA, FFR and the proposed MXMN algorithm. To demonstrate the results, the RX algorithm was applied to 6 real hyperspectral image cubes, collected by an AISA airborne sensor configured to 65 spectral bands, uniformly covering VNIR range of $400nm$ - $1000nm$ wavelengths. At 4 km altitude, a pixel resolution corresponds to $(0.8m)^2$. The obtained image cubes are $b \times r \times c = 65 \times 300 \times 479$

hyperspectral images, where b,r and c denote the number of hyperspectral bands, the number of rows and the number of columns in the image, respectively.

In Fig. 5.1, we show the 30th band of a typical hyperspectral image cube. The image contains ground-truth anomalies (vehicles and small agriculture facilities, which occupy a few pixel segments marked in white and encircled by red ellipses), which were manually identified using side information collected from high resolution RGB images of the corresponding scenes. All 6 images are not shown here just because of space limitations.

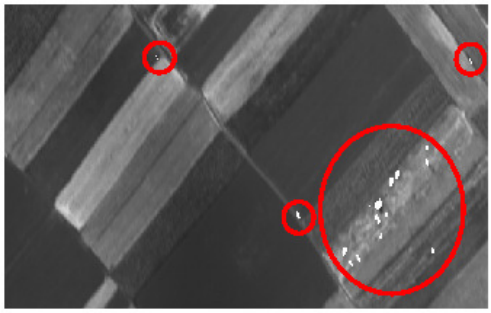


Figure 5.1: 30th band of a hyperspectral image cube with anomalies marked in white and encircled by red ellipses.

We applied FFR and the proposed MXMN algorithms to an image cube that *does not contain anomalies* to reduce the hyperspectral dimensionality from 65 to 10 by the corresponding piece-wise constant spectral segments. We also applied PCA to obtain an ℓ_2 optimal 10-dimensional basis. In Fig. 5.2, one can see the obtained piece-wise constant approximations by FFR (cyan (bright) thick line) and MXMN (blue (dark) thick line) for 3 selected hyperspectral pixels (blue thin lines). The leftmost pixel is an anomaly, whereas the other two pixels were selected from different background regions. As can be seen from the figure, the partition obtained by MXMN has a denser granularity in bands $[1-35]$, in which the anomaly is expressed. This is on the expense of other bands, which, in spite of being energetically prominent, are less important for anomaly detection. On the contrary, FFR adapts better to the energetical bands and, as a result, assigns less channels to bands $[1-35]$ which makes it prone to loosing anomalies.

In Fig. 5.3, we compare FFR, MXMN and PCA in terms of Receiver Operation Characteristic (ROC) curves obtained by applying the RX algorithm on

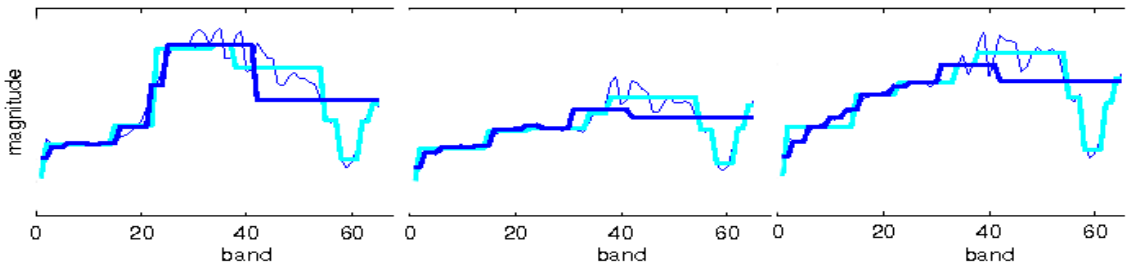


Figure 5.2: **Piecewise constant approximation.** The leftmost graph is anomaly pixel, whereas two right graphs are background pixels. Original spectrum is in blue (dark) thin line, MXMN approximation is in blue (dark) thick line, FFR approximation is in cyan (bright) thick line.

hyperspectral data after the dimensionality reduction. For the purpose of ROC curves generation, all 6 hyperspectral images were used, in which the total number of anomaly segments count is 25. It is cleraly seen from the figure that the MXMN algorithm corresponds to a better ROC curve (blue solid line) compared to other dimensionality reduction techniques such as FFR (cyan dashed line) or PCA (red solid line with solid circles) for all tested parameters. It is important to note that the performance of the RX algorithm applied to the data obtained by the MXMN is even better than applying RX to the full-dimensional (original) images (green dot-dashed line), for the range of low false-alarm rates. This can be explained by the fact that MXMN performs averaging of hyperspectral bands corresponding to the background clutter, which alleviates the effect of masking out anomaly contributions by background clutter. As a matter of fact, since $J(\mathbf{b}_K)$ of (5.5) is designed to favor partitions that produce errors which are "more Gaussian", its minimization using typical data *without anomalies*, results in a coarse partition in the spectral bands containining background clutter. These bands are noisy, they have less discriminative power and they may mask out subtle anomaly-related contributions that may appear in other bands. The fine partition is obtained in the spectral bands that are "less Gaussian", they have more discriminative power and, therefore, may potentially contain anomaly-related information. This may explain why the proposed algorithm corresponds to a better ROC curve compared to the other algorithms, although the optimal partition was obtained using *an image that does not contain anomalies.*

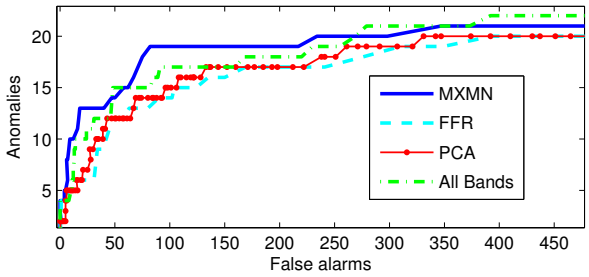


Figure 5.3: **ROC curves.**

5.3 Summary

In this chapter we proposed a novel approach for channel reduction in hyperspectral images that allows designing multispectral filters that facilitate a good performance of local anomaly detection algorithms. The channel reduction is performed by replacing subsets of adjacent hyperspectral bands by their means, producing a piecewise constant pixel approximation. An optimal partition of hyperspectral bands is obtained by *Minimizing the Maximum of Mahalanobis Norms* of errors, obtained due to missrepresentation of these bands by constants. Hence, the proposed algorithm is denoted as MXMN. The minimization is performed by a dynamic programming technique, as used by the Fast Hyperspectral Feature Reduction (FFR) algorithm proposed in [77]. We compared MXMN with FFR and SVD by examining the results of the RX algorithm [32] applied after the dimensionality reduction. It was demonstrated that the proposed MXMN algorithm results in a better ROC curve in the whole range of false alarm values, and even better than applying RX on the original data without the dimensionality reduction in the important range of low false-alarm rates.

Chapter 6

Conclusion

6.1 Summary

In this research we have studied how to perform redundancy reduction of high-dimensional noisy signals for applications where a good representation of *both* the abundant and the anomaly vectors is essential. The combined subspace of anomaly and abundant vectors is obtained by using the proposed $\ell_{2,\infty}$-norm that penalizes individual data-vector miss-representations.

In the first part of the research, a sub-optimal greedy algorithm is developed that is designed to optimize the $\ell_{2,\infty}$-based criterion. It uses a combination of SVD and direct selection of vectors from the data to form the signal-subspace basis. The rank is determined by applying Extreme Value Theory results to model the distribution of the maximal noise-residual ℓ_2-norms. In simulations, conducted for various rare-vectors signal-to-noise conditions, the proposed approach is shown to yield good results for practically-significant RSNR values (RSNR essentially measures the SNR of rare-vectors with respect to noise), for which the classical methods of SVD and MDL fail to determine correctly the signal-subspace and rank, respectively, of high dimensional signals composed of abundant and rare vectors.

The proposed approach was also applied for the signal-subspace and rank determination of a hyperspectral image with and without anomaly pixels. The results of MOCA were found to be equal to those of MDL (or, when necessary, robust MDL) for the pure-background subimage, whereas in the presence of

anomalies, MOCA has detected a higher rank than MDL, while MDL produced the same rank as in the pure-background case. This indicates that MDL failed to determine correctly the signal-subspace rank of a hyperspectral image composed of both abundant and rare vectors, whereas MOCA succeeded in representing it well.

In the next part of the research, we have proposed an algorithm for anomaly detection, discrimination and population estimation of anomalies of the same type, called AXDA. The algorithm is based on a signal-subspace and rank estimation provided by MOCA. By its construction, the signal basis consists of two groups of basis vectors. One group spans the subspace of anomalies. The second group is designed to represent background pixel residuals belonging to the subspace that is complementary to the subspace of the anomalies. The proposed AXDA extracts anomaly pixels by removing an anomaly basis vector from the anomaly vectors group and compensating for its removal by augmenting the background vectors related subspace. This operation causes a violation of the noise hypothesis condition in vectors that are highly correlated with the removed anomaly basis vector. Such vectors are detected, associated with the removed basis vector, and depleted from the data. This way we obtain groups of data vectors associated with each one of the anomaly basis vectors.

In experiments with real hyperspectral image cubes AXDA was shown to have a better performance than GMRX and MSD, in most of the range of the tested parameters. Since the anomaly and background subspaces are unknown in advance, the MSD algorithm was provided the anomaly-free estimation of the background basis $\mathbf{\Psi}_s$ obtained from AXDA and the anomaly subspace obtained from MOCA. This provides MSD subspace-related information that is (at least) as good as AXDA has for the detection of anomalies. It is also important to note, that in contrast to MSD and GMRX, AXDA is equipped with an unsupervised determination of the nominal operating point. AXDA also has a capability to discriminate between different types of anomalies, though the accuracy of this discrimination, as well as the accuracy of population estimation of anomalies of the same type, are topics for future research. Moreover, AXDA allows also an anomaly-free (robust) estimation of the background-subspace and rank.

It turns out now that MOCA in combination with AXDA provide means to meet a wide range of signal-subspace estimation scenarios:

1. *Estimation of a signal-subspace that includes anomaly-vectors.*
2. *Detection of anomaly-vectors and determination of their subspace.*
3. *Providing a natural (nominal) operating point for anomaly detection.*
4. *Estimation of a pure (free of outliers) background-subspace.*

Next, we propose an algorithm that is denoted as (MOOSE), which is the optimal version of the suboptimal greedy algorithm for anomaly preserving signal subspace estimation provided by MOCA. MOOSE is optimal in the sense that the estimated subspace (locally) minimizes the proposed $\ell_{2,\infty}$-norm of misrepresentaion residuals. The optimization is performed via a natural conjugate gradient learning approach carried out on the set of n dimensional subspaces in $\mathbb{R}^m$, $m > n$, known as the Grassmann manifold. As any local minimization of a non-convex objective function, MOOSE is prone to getting trapped in a local minimum. Therefore, a proper initialization is crucial and is obtained by employing MOCA that uses global principles to find a suboptimal solution that is close to the global minimum. The results of MOOSE were compared to the results of SVD and MOCA by applying them both on simulated data and on real hyperspectral images. It was demonstrated that the results of MOOSE and MOCA are much better than those of SVD in terms of max-norm residual error, obtained in both simulated and real data, and in terms of the subspace estimation error obtained for simulated data. Although MOCA exhibits results inferior to those of MOOSE, the results of MOCA are quite comparable to those of MOOSE meaning that practically, the greedy signal subspace estimation algorithm of MOCA is a good choice, since it is more computationally efficient.

Finally, we have proposed a novel unsupervised technique for Designing Multispectral Filters that facilitates a better performance of local anomaly detection algorithms. We have shown that the problem of designing Multispectral Filters can be considered as a special case of the problem of Channel Reduction in Hyperspectral Images. Here, the channel reduction is performed by replacing subsets

of adjacent hyperspectral bands by their means, producing a piecewise constant pixel approximation. An optimal partition of hyperspectral bands is obtained by *Minimizing the Maximum of Mahalanobis Norms* of errors, obtained due to missrepresentation of these bands by constants. Hence, the proposed algorithm is denoted as MXMN. The minimization is performed by a dynamic programming technique, as used by the Fast Hyperspectral Feature Reduction (FFR), a similar channel reduction algorithm based on minimizing the ℓ_2-norm of errors. We compared MXMN with FFR and SVD by examining the results of the RX algorithm applied after the dimensionality reduction. It was demonstrated that the proposed MXMN algorithm results in a better ROC curve in the whole range of false alarm values, and even better than applying RX on the original data without the dimensionality reduction in the important range of low false-alarm rates.

6.2 Future Directions

The proposed anomaly detection algorithm AXDA exploits the greedy structure of MX-SVD - a suboptimal anomaly preserving signal-subspace estimation algorithm. Although the optimal signal-subspace obtained by MOOSE improves the performance of MX-SVD in terms of signal-subspace estimation error, it does not produce a signal-subspace basis with a special structure like MX-SVD. This makes the direct use of MOOSE in AXDA impossible. Therefore, in a future research, one may need to develop an anomaly detection algorithm that directly uses the optimal signal-subspace produced by MOOSE.

As discussed in chapter 3, AXDA also has a capability to discriminate between different types of anomalies, and to estimate populations of anomalies of the same type. However, the accuracy of this discrimination, as well as the accuracy of population estimation of anomalies of the same type, are not evaluated. As a matter of fact, we have observed that the results of the discrimination and population estimation depend on the order of removing columns from the matrix $\mathbf{\Omega}$ while applying AXDA. The selection of a proper column removal order and a more accurate classification of anomalies to their corresponding class may be a subject for future research.

In chapter 5, we have proposed the MXMN technique for channel reduction in hyperspectral images, which allows designing multispectral filters that are tuned for local anomaly detection algorithms. In simulations with real data, 65 hyperspectral channels were reduced by MXMN to 10 multispectral channels. On one hand, the channel reduction alleviates the distracting influence of noise or the background clutter. On the other hand, the channel reduction reduces the amount of information needed for the anomaly detection. In that simulation we have demonstrated that MXMN results in a better ROC curve in the whole range of false alarm values as compared to FFR and SVD, and even better than applying RX on the original data without the dimensionality reduction in the important range of low false-alarm rates. A reasonable question is whether there is another dimensionality, other than 10, which could have produced even better results. Therefore, a future research subject can be the development of an unsupervised algorithm for the selection of multispectral dimensionality, that is optimal in terms of the performance of an anomaly detection algorithm.

Appendix A

Distribution of maximum-norm noise realizations

In this appendix we characterize the pdf $f_{\nu_k}(\cdot)$ of section 2.4.1. We assume that the noise is a zero-mean white Gaussian process, with known standard deviation σ. Then, its residual squared norms

$$\zeta_{k,i} \triangleq \|\mathcal{P}_{\hat{\mathcal{S}}_k^{\perp}} \mathbf{z}_i\|^2, \tag{A.1}$$

$i = 1, \ldots, N$, have a Chi-squared distribution of order $l \triangleq \text{rank}\, \hat{\mathcal{S}}_k^{\perp} = p - k$, denoted by $\chi^2(l, \sigma^2)$ with the following pdf [**?**]:

$$f(u) = \frac{1}{2^{l/2}\Gamma(l/2)\sigma^2} \left(\frac{u}{\sigma^2}\right)^{(l/2)-1} e^{-u/2\sigma^2}. \tag{A.2}$$

For large l, the Central Limit Theorem can be used to obtain the following approximation:

$$\zeta_{k,i} \sim \chi^2(l, \sigma^2) \approx \mathcal{N}\left(l\sigma^2, 2l\sigma^4\right). \tag{A.3}$$

Now, the limiting distribution of ν_k, which satisfies

$$\nu_k = \max_{i=1,\ldots,N} \zeta_{k,i}, \tag{A.4}$$

can be obtained using the following Extreme Value Theory result:

Theorem 1 *[54]*
If $\{\zeta_i\}_{i=1}^N$ is i.i.d., with absolutely continuous distribution $F(x)$ and density $f(x)$, and letting

(i) $h(x) = f(x)/(1 - F(x))$

(ii) $b_N = F^{-1}(1 - \frac{1}{N})$

(iii) $a_N = h(b_N)$

(vi) $\omega = lim_{x \to x^*} \frac{dh(x)}{dx}$,
where x^* *is the upper end-point of* F,

then, for $M_N = \max\{\zeta_1 \dots \zeta_N\}$,

$$P(a_N(M_N - b_N) \leq u) \underset{N \to \infty}{\longrightarrow} \begin{cases} \exp(-e^{-u}), & if\ \omega = \infty \\ \exp\{-[1 + \frac{u}{\omega}]^{\omega}\}, & if\ \omega < \infty \end{cases}, \tag{A.5}$$

The proof is found in [54].

In words: Theorem 1 says that the maximum of N i.i.d random variables has a limiting distribution that depends on ω - a parameter derived from their individual distributions. For the purposes of the present work, we consider normal and chi-squared distributions, which lead to $\omega = \infty$.

Therefore, from (A.5), the limiting distribution of interest is

$$\mathcal{G}(u) \triangleq \exp(-e^{-u}), \tag{A.6}$$

also known as the Gumbel distribution [1]. The mean and std of a variable distributed as (A.6) are $\eta = 0.5772$ and $\gamma = 1.6450$, respectively. The normalizing coefficients a_N and b_N are also functions of the ζ_i distribution. Theorem 1 also describes how to calculate the normalizing coefficients given the distribution function of ζ_i.

Unfortunately, there are no known analytical expressions for the normalizing coefficients a_N and b_N corresponding to $\{\zeta_{k,i}\}$ (defined in (A.1)) that are chi-square distributed. In our evaluations of the asymptotic pdf of ν_k in Fig. 2.4(b)

[1] Extreme Value Distributions are the limiting distributions of the minimum or the maximum of a very large collection of random observations from the same arbitrary distribution. Gumbel (1958), [55] showed that for any well-behaved initial distribution (i.e., F(x) is continuous and has an inverse), only a few models of limiting distributions are needed, depending on whether one is interested in the maximum or the minimum, and also if the observations are bounded from above or below (see [52]).

above and in the sequel, we used the results of Theorem 1 to calculate a_N and b_N numerically. Note, that a_N and b_N are also functions of l and σ, since they depend on the $\chi^2(l, \sigma^2)$ distribution, which is a function of l and σ.

However, in order to explain why the pdf of $\|\mathcal{P}_{\hat{\mathbf{S}}^\perp}\mathbf{Z}\|^2_{2,\infty}$, shown in Fig. 2.4(a) and Fig. 2.5(b), is so narrow; one can use the approximation in (A.3) to obtain the following asymptotic analysis, which can be conducted analytically. It can be shown [52] that for $\{\zeta_i\}$ of Theorem 1, which are Gaussian, M_N is distributed as follows:

$$P(M_N \leq u) \underset{N\to\infty}{\longrightarrow} \mathcal{G}(a_N(u - b_N)), \tag{A.7}$$

with

$$\begin{aligned} a_N &= (2\ln N)^{1/2} \\ b_N &= (2\ln N)^{1/2} - \\ &\quad \frac{1}{2}(2\ln N)^{-1/2}(\ln\ln N + \ln 4\pi). \end{aligned}$$

Therefore,

$$P(\nu_k \leq x) \approx \mathcal{G}\left(a_N\left[\frac{x - \sigma^2 l}{\sigma^2\sqrt{2l}} - b_N\right]\right) \tag{A.8}$$

with mean and std:

$$\mu_N = \sigma^2\left(\frac{\sqrt{2l}}{a_N}\eta + b_N\sqrt{2l} + l\right) \tag{A.9}$$

$$\sigma_N = \frac{\sigma^2\sqrt{2l}}{a_N}\gamma \tag{A.10}$$

While this approximation doesn't provide us with an accurate mean and std of ν_k, it is instructive to look at the following ratio that defines a relative width of a pdf for $N \gg 1$, $l \gg 1$:

$$\frac{\mu_N}{\sigma_N} \propto 2\ln N + \sqrt{l\ln N}. \tag{A.11}$$

It is observed that this ratio doesn't depend on σ^2, and it is log-dependent on N. Thus, the ratio μ_N/σ_N tends to infinity as $N \to \infty$ or $l \to \infty$. For example, for $l = 100, N = 10^5$ and white noise, $\mu_N/\sigma_N \approx 23$ corresponding to quite a small relative width. The dominant factor in obtaining such a high ratio is the high

dimensionality of $l = 100$.

Appendix B

Derivation of posterior hypothesis probabilities

In the following, we derive the conditional probabilities $p(H_0|\eta_k)$ and $p(H_1|\eta_k)$ in (2.25) and (2.26), based on pdfs f_{ν_k} and f_{ξ_k}:

$$\begin{aligned}
f(H_0,\eta_k) &= f_{\nu_k}(y)p(\xi_k<\eta_k) = \\
&\quad f_{\nu_k}(\eta_k)F_{\xi_k}(\eta_k) = f_{\nu_k}(\eta_k)\frac{\eta_k}{\eta_{k-1}}, \\
f(H_1,\eta_k) &= f_{\xi_k}(\eta_k)p(\nu_k<\eta_k) = \\
&\quad f_{\xi_k}(\eta_k)F_{\nu_k}(\eta_k) = F_{\nu_k}(\eta_k)\frac{1}{\eta_{k-1}}, \\
f_{\eta_k}(\eta_k) &= f(H_0,\eta_k)+f(H_1,\eta_k) = \\
&\quad \frac{1}{\eta_{k-1}}\left[\eta_k f_{\nu_k}(\eta_k)+F_{\nu_k}(\eta_k)\right], \\
p(H_0|\eta_k) &= \frac{f_{\nu_k}(\eta_k)F_{\xi_k}(\eta_k)}{f_{\eta_k}(\eta_k)} = \frac{\eta_k f_{\nu_k}(\eta_k)}{\eta_k f_{\nu_k}(\eta_k)+F_{\nu_k}(\eta_k)}, \\
p(H_1|\eta_k) &= \frac{f_{\xi_k}(\eta_k)F_{\nu_k}(\eta_k)}{f_{\eta_k}(\eta_k)} = \frac{F_{\nu_k}(\eta_k)}{\eta_k f_{\nu_k}(\eta_k)+F_{\nu_k}(\eta_k)},
\end{aligned}$$

which are the expressions shown in (2.25) and (2.26).

Appendix C

Assessment of MOCA reliability in terms of RSNR

In the following we assess the dependence of MOCA rank estimation error (see chapter 2) on the value of RSNR.

Let's recall that the RSNR notion was introduced in the context of SVD performance assessment in the presence of rare-vectors. It measures the ratio between the contribution of rare-vectors and the contribution of noise to the signal covariance matrix. Thus, being ℓ_2-based, RSNR is an ambiguous measure for MOCA performance assessment, which is affected by individual data-vector contributions. For example, two identical rare-vectors of the same ℓ_2-norm value l, have the same RSNR as that of a single rare-vector of an ℓ_2-norm value of $l\sqrt{2}$, and thus have the same SVD performance. However, MOCA may behave differently in each of the two cases in this example. Thus, in some applications, there is typically only one rare-vector out of 10^5 data-vectors, whereas in other applications, even 10 collinear vectors out of 10^5 are considered to be rare. Different rare-vector multiplicities cause MOCA to depend differently on the RSNR. In order to eliminate this ambiguity, we constrain the rare-vectors in the following analysis to be linearly independent. Otherwise, the RSNR value should be corrected by an appropriate rare-vectors multiplicity factor in order to obtain an equivalent MOCA performance.

If the SNR of abundant vectors is high enough, then we can assume that for $k \geq r_a$, where r_a is the abundant vectors subspace rank, the SVD-part of MOCA estimates well the abundant vectors subspace, and that MOCA iterations don't

terminate before $k = r_a$. Thus, in the complementary subspace for $r_a \leq k < r$, one would find only residuals of abundant vectors, composed of noise only, and residuals of rare-vectors. Let's denote

$$\tilde{\mathbf{Y}}_{rare} \triangleq \mathcal{P}_{\mathbf{Y}_{abund}^{\perp}} \mathbf{Y}_{rare}, \tag{C.1}$$

i.e., the projection of the rare-vectors sub-matrix onto the abundant-vectors null-space. Our purpose here is to characterize RSNR values for k values satisfying $r_a \leq k \leq r$, for which there is a high probability that rare-vectors will be selected among $\mathbf{\Omega}_{k-r_a}$ columns (see (2.12)).

Let's assume that for some iteration k, $r_a < k \leq r$, the matrix $\mathbf{\Omega}_{k-r_a}$ is composed of rare vectors. We are looking for conditions on RSNR that guarantee selecting the next rare-vector at iteration k as in (2.12), with probability close to 1. This RSNR value would also justify the assumption on the matrix $\mathbf{\Omega}_{k-r_a}$ above, since (as we'll see later) it would guarantee the rare-vectors selection for all $r_a < k \leq r$, with probability close to 1. If one neglects the effect of noise on the rare vectors selected in $\mathbf{\Omega}_{k-r_a}$, then the ℓ_2-norms of the remaining $r-k$ rare vectors can equivalently be obtained as the last $r-k$ diagonal entries of the upper triangular matrix $\mathbf{R}$ obtained via the following QR decomposition:

$$\mathbf{QR} = \tilde{\mathbf{Y}}_{rare}\mathbf{\Pi}, \tag{C.2}$$

where $\mathbf{\Pi}$ is a permutation matrix that moves $\tilde{\mathbf{Y}}_{rare}$ columns of rare-vectors selected in $\mathbf{\Omega}_{k-r_a}$ to the leading positions. Now, we use the following lemma in order to obtain a relation between the RSNR of $\mathbf{Y}$ and the diagonal entries of $\mathbf{R}$.

Lemma 1 *The minimal singular value s_{min} of a full-rank $m \times n$ matrix $\mathbf{M}$ with $m > n$, satisfies $s_{min} \leq \rho_j$, $j = 1, \ldots n$, where ρ_j are the diagonal entries of a triangular matrix in the QR decomposition of $\mathbf{M\Pi}$, with $\mathbf{\Pi}$ - any permutation matrix.*

Proof

Let ρ_j be a diagonal entry for some $j = 1, \ldots, n$, and let $\hat{\mathbf{\Pi}}$ be another permutation matrix that moves column j of $\mathbf{M\Pi}$ to the last. Then, the corresponding $\hat{\rho}_n$ of $\mathbf{M\Pi\hat{\Pi}}$ satisfies: $\hat{\rho}_n \leq \rho_j$, since it is a norm of a projection onto a smaller (contained) subspace. Now, according to [21], the following holds: $\hat{s}_{min} \leq \rho_n$. Therefore, $s_{min} \leq \rho_j$.

Using the lemma above and the definition of RSNR (2.4), one obtains:

$$\varsigma_k \geq \text{RSNR}\sigma^2(p-r), \tag{C.3}$$

where $\varsigma_k \triangleq \|\tilde{\mathbf{y}}_{max}\|^2$, and $\tilde{\mathbf{y}}_{max}$ is the maximum-norm rare-vector residual in $\hat{\mathbb{S}}_k^{\perp}$. Since the termination condition of MOCA is based on testing the maximum squared norm of residuals $\eta_k = \|\mathcal{P}_{\hat{\mathbb{S}}_k^{\perp}}\mathbf{X}\|_{2,\infty}^2$, it is important to calculate the pdf of η_k, which satisfies:

$$\eta_k = \max(\xi_k, \nu_k), \tag{C.4}$$

where,

$$\nu_k = \|\mathcal{P}_{\hat{\mathbb{S}}_k^{\perp}}\mathbf{X}_{abund}\|_{2,\infty}^2 \tag{C.5}$$

$$\xi_k = \|\tilde{\mathbf{y}}_{max} + \mathbf{n}\|^2, \tag{C.6}$$

we also assume here that the RSNR value is large enough, so that:

$$\underset{\tilde{\mathbf{y}}_i \in \text{columns } \tilde{\mathbf{Y}}}{\text{argmax}} \|\tilde{\mathbf{y}}_i + \mathbf{n}\| = \underset{\tilde{\mathbf{y}}_i \in \text{columns } \tilde{\mathbf{Y}}}{\text{argmax}} \|\tilde{\mathbf{y}}_i\|, \tag{C.7}$$

with probability close to 1.

Now, the distribution function of η_k for $r_a \leq k < r$ is given by:

$$F_{\eta_k}(\cdot) = \mathcal{G}_{p-k}(\cdot) NC\chi^2_{p-k,\delta}(\cdot), \tag{C.8}$$

where $\mathcal{G}_{p-k}(\cdot)$ is the Gumbel distribution of the noise max-norm with $p-k$ degrees of freedom, as described in Appendix A, and $NC\chi^2_{p-k,\delta}(\cdot)$ is the noncentral chi-square distribution [22], with $p-k$ degrees of freedom and δ is its non-centrality parameter. The results of [22] and relation (C.3) can be used to obtain:

$$\delta = \frac{\varsigma_k}{\sigma^2} \geq \text{RSNR}\ (p-r). \tag{C.9}$$

The pdf of η_{r-1}, $f_{\eta_{r-1}}$, corresponding to a situation where $\varsigma_{r-1} = \text{RSNR}\sigma^2(p-r)$ (selecting the worst case in (C.3)), RSNR $= 2$, $p = 100$, $r = 10$, $r_a = 5$, $\sigma = 1$, $N = 10^4$ is shown in Fig. C.1, solid line. The choice of $k = r-1$ is arbitrary for numerical demonstration purpose only. Now, the distribution of η_r, $F_{\eta_r}(\cdot)$,

equals to the distribution of maximum-norm noise residual $\mathcal{G}_{p-r}(\cdot)$, since $\hat{\mathbb{S}}_r^{\perp}$ is supposed to include only noise. The pdf of η_r, f_{η_r}, is plotted in dashed line. The rank-determination threshold τ_{r-1} at iteration $r-1$ (marked by a vertical line) equals to η_{r-1}, satisfying:

$$p(H_0|\eta_{r-1}) = p(H_1|\eta_{r-1}), \tag{C.10}$$

where H_0, H_1 are defined in subsection 2.4.1.

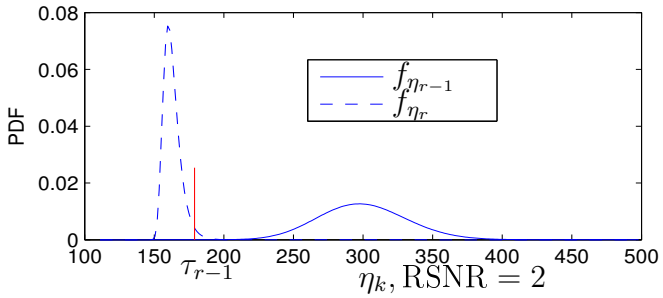


Figure C.1: Pdf of the maximum residual norm η_{r-1} and η_r for $k = r-1$, $\varsigma_{r-1} = \text{RSNR}\sigma^2(p-r)$, RSNR $= 2$, $p = 100$, $r = 10$, $r_a = 5$, $\sigma = 1$, $N = 10^4$ at iteration $r-1$ (solid line) and iteration r (dashed line), respectively. The rank-determination threshold τ_r at iteration r is marked by a vertical line.

Now, the probability p_u of rank underestimation, given that iteration $r-1$ is reached, is given by $p_u = F_{\eta_{r-1}}(\tau_{r-1})$, which for the parameters above is of the order of 10^{-6}! It turns out that for the parameters above, the order of the rank underestimation error is approximately the same for all k values $r_a \leq k < r$, which is small enough to be neglected.

It is important to note that typically, $f_{\eta_{r-1}}$ would lie farther from the threshold τ_{r-1}, since selecting equality in (C.3), in this example, corresponds to the worst case. This decreases the probability of the rank underestimation even further. Due to properties of $\mathcal{G}_{p-k}(\cdot)$, the distribution of η_k has a weak $\log N$ dependence on the data sample size N (see (A.8)). Whereas $NC\chi^2_{p-k,\delta}(\cdot)$ doesn't depend on N at all. Therefore, the rank underestimation error is also negligible for $N = 10^3$ as well as for $N = 10^5$.

The probability of rank overestimation p_o at iteration $k = r$, is given by $p_u = 1 - F_{\eta_r}(\tau_r) = 1 - \mathcal{G}_{p-r}(\tau_r)$, which for the parameter values above gives

$p_o \approx 0.027$. This value is nearly constant for all RSNR values above 2, which, as we have seen earlier, guarantee a negligible p_u. It can be decreased by modifying the hypotheses equality test of (C.10) to the following likelihood ratio test:

$$p(H_0|\eta_r) \leq \gamma p(H_1|\eta_r), \qquad \gamma < 1. \tag{C.11}$$

This should produce a lower error-rate at the expense of a higher τ_{r-1}. Fortunately, as it is clearly seen in Fig. C.1, the pdf $f_{\eta_{r-1}}$ lies far from τ_{r-1}, which means that a lower p_o can be obtained by choosing an appropriate $\gamma < 1$ leaving p_u still negligible.

Appendix D

Robust MDL with a modification that accounts for noise dependence between bands

In section 4.3 we apply the RMDL approach [27] as an ℓ_2-based alternative to the classical MDL approach for signal-subspace rank determination. The assumption of RSNR that the noise covariance matrix is diagonal, but with different diagonal entries $\sigma_1^2, \ldots, \sigma_p^2$, makes the algorithm robust to deviations of noise variances from being equal in all spectral bands. In order to model also the observed small dependence of noise components between adjacent bands, we assume that the secondary-diagonal noise covariance matrix entries are all-equal to a parameter β_k. As in [27], let's define $\sigma^2 \triangleq \frac{1}{p}\sum_{i=1}^{p}\sigma_i^2$ and $w_i \triangleq \sigma_i^2 - \sigma^2$. Now the model parameters vector of (2.27) can be expressed via:

$$\boldsymbol{\Theta}(k) = (\lambda_1, \ldots, \lambda_k, \mathbf{V}_1, \ldots, \mathbf{V}_k, \sigma, w_1, \ldots, w_p, \beta_k). \tag{D.1}$$

This modification requires changing steps 3 and 4 in [27] (p. 3547) as follows:
In Step 3: Adding the computation of β_k as follows:

$$\beta_k = \text{mean}\left(\text{offdiag}\left(\hat{\mathbf{R}} - \mathbf{A}_k\mathbf{R}_{s,k}(\mathbf{A}_k)^H - (\sigma_{n,k})^2\mathbf{I}\right)\right), \tag{D.2}$$

where offdiag($\mathbf{R}$) returns a second diagonal of the matrix $\mathbf{R}$.

In step 4: Changing the computation of $\mathbf{E} = \hat{\mathbf{R}} - \mathbf{w}_k$ to $\mathbf{E} = \hat{\mathbf{R}} - \mathbf{w}_k - \beta_k\mathbf{I}_{\text{off}}$, where $\mathbf{I}_{\text{off}}$ denotes a $p \times p$ matrix with ones on its second diagonals and zeros

everywhere else.

Appendix E

Noise variance estimation procedure

AXDA and MOCA strongly rely on the assumption of additive white Gaussian noise of known variance. The correct specification of the noise variance is of paramount importance since it determines the signal subspace rank (see (2.23)) and, as a result, affects the detection/false-alarm rates. In this appendix we describe a technique used for the estimation of noise variance in each hyperspectral channel. The estimated noise variance is then used for a band-wise normalization of the noise variance to 1.

It was observed in experiments with real hyperspectral data that *overestimation* of the noise variance by about a half an order of magnitude has little influence on AXDA performance. Although using an overestimated value of the noise variance (causing a poorer representation of the background) would result in the underestimation of signal subspace rank, the false alarm rate remains mostly unchanged. This happens since background misrepresentations are tested by rule (2.23) (used in steps 9 and 13 of AXDA), which depends on the noise variance as well. Thus, the overestimated noise variance raises the "effective threshold value", which naturally leaves the background misrepresentations undetected. Thus, an overestimated noise level may just slightly impair detection rate of anomalies that aren't prominent enough.

Using an *underestimated* value of the noise variance is less favorable because of special statistical properties of the maximal norm of noise. As shown in chapter 2, the maximal norm of noise has a narrow distribution, explained by Extreme Value Theory results. Therefore, there is a high likelihood that the maximal norm

of noise would obtain an almost deterministic value (see Figs. 4 and 5 in chapter 2). Thus, if the underestimated noise variance makes the "effective threshold value" implied by (2.23) lower than the almost deterministic maximal norm of noise, MOCA would never terminate its iterations (or will terminate too late). This would result in a significant signal-subspace rank overestimation, which may cause the background subspace of increased-rank to include the anomalies and to significantly impair the anomaly detection rate. Therefore, the noise variance estimation technique proposed below prefers *noise variance overestimation*.

In CCD-based hyperspectral systems, the noise is a combination of dark current noise, photon (shot) noise and fixed pattern noise (FPN) [49]. The FPN is due to different sensor responsivities, which is estimated and compensated out by calibrating the sensor. It turns out that even at mild light intensities, the *photon noise* may be dominant [49]. The photon noise problem arises from the statistical nature of photon production. The probability distribution for n photons in an observation window is known to be Poisson:

$$p(n) = \frac{M^n e^{-M}}{n!}, \tag{E.1}$$

where M is the average number of photons within the given observation window. For the linear part of the CCD response function, the image intensity I is linearly proportional to n, i.e.,

$$I = gn, \tag{E.2}$$

for some proportionality coefficient g. Since the Poisson distribution approaches a normal distribution for large M, the photon noise in I can be modelled as having a zero-mean normal distribution with std σ_e satisfying:

$$\sigma_e(I) = \sqrt{gH}, \tag{E.3}$$

where H being the mean of I satisfying $H = gM$, is considered to be the clean signal.

Thus, the photon noise variance is not constant and, therefore, doesn't meet the noise stationarity property assumed in MOCA. Nevertheless, in our real-data

simulations, we have empirically found that using

$$\sigma_{0.98} = \sqrt{gH_{0.98}} \tag{E.4}$$

in AXDA, as an estimate of the noise std in each band (with $H_{0.98}$ denoting the 0.98 quantile of image intensities in the band), decreases false alarm rate caused by high image intensity pixels, while allowing a reasonably high anomaly detection rate. The 0.98 quantile corresponds to almost maximum image intensity, ignoring 2% of the most intense image values that may stem from anomalies.

The only thing left is to estimate g. According to (E.3), g satisfies:

$$g = var\left(\frac{e}{H}\right), \tag{E.5}$$

where e denotes pixel noise and *var* denotes variance. Note, that random variables $\{e_i/H_i\}$, where i denotes pixel index, are identically distributed. If one assumes that they are independent, then g can be estimated by:

$$\hat{g} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\hat{e}_i}{\hat{H}_i}\right)^2, \tag{E.6}$$

where $\hat{e}_i$ is a noise estimation, and $\hat{H}_i$ is a clean image intensity estimation.

The estimation of $\{H_i\}$ can be obtained via a 2D linear prediction as follows:

$$\hat{H}_i = \sum_{j\in\mathcal{K}} a_j I_{i,j}, \tag{E.7}$$

where $\mathcal{K}$ denotes the set of a 2D neighborhood indices, $I_{i,j}$ denotes the image intensity at position j in the neighborhood of a pixel i, and $\{a_j\}$ denote the linear prediction coefficients, obtained via least squares over the *whole image*. In our simulations we used a 5×5 neighborhood.

The estimation of $\{e_i\}$ is then given by

$$\hat{e}_i = I_i - \hat{H}_i. \tag{E.8}$$

Unfortunately, the estimation of $\{H_i\}$ given in (E.7) is inaccurate in non-smooth image regions such as edges and/or anomaly pixels. Therefore, the estimates $\hat{e}_i$ and $\hat{H}_i$ from these regions should not be accounted in (E.6) for the estimation of

g. In order to filter out the undesired contributions of $\hat{e}_i$ and $\hat{H}_i$, we estimate $\sqrt{g}$ using median absolute deviation (MAD), proposed in [57], for a robust estimation of the standard deviation of e/H as follows:

$$\sqrt{\hat{g}} = MAD\left(\hat{e}/\hat{H}\right) = \underset{i=1,\ldots,N}{median}\left|\hat{e}_i/\hat{H}_i\right|, \tag{E.9}$$

where N is the total number of hyperspectral pixels.

Using the estimated values of g and H_i and substituting to (E.4), we obtain an estimate of the effective noise std $\sigma_{0.98}$ in each band and normalize the noise to unity variance in each band of the hyperspectral cube for further processing.

References

[1] O. Kuybeda, D. Malah and M. Barzohar, *Rank Estimation and Redundancy Reduction of High-Dimensional Noisy Signals with Preservation of Rare Vectors,* IEEE Trans. Signal Proc., vol. 55, no. 12, pp. 5579-5592, Dec. 2007.

[2] O. Kuybeda, D. Malah and M. Barzohar, *Anomaly Preserving $\ell_{2,\infty}$-Optimal Dimensionality Reduction over a Grassmann Manifold,* submitted to IEEE Trans. Signal Proc. on Aug. 2008.

[3] O. Kuybeda, D. Malah and M. Barzohar, *Global Unsupervised Anomaly Extraction and Discrimination in Hyperspectral Images via Maximum- Orthogonal Complement Analysis,* EUSIPCO - European Signal Processing Conference, Aug. 2008, Lausanne. Switzerland.

[4] O. Kuybeda, D. Malah and M. Barzohar, *Hyperspectral Channel Reduction for Local Anomaly Detection,* submitted to EUSIPCO 2009.

[5] Chein-I Chang, *Hyperspectral Imaging Techniques For Spectral Detection and Classification,* Springer, 2003.

[6] A. Hyvarinen, J. Karhunen, and E. Oja. *Independent Component Analysis,* Series on Adaptive and Learning Systems for Signal Processing, Communications, and Control. Wiley, 2001.

[7] M. Belkin, *Problems of learning on manifolds,* Ph. D. thesis, University of Chicago, 2003.

[8] B. SchÃűlkopf, A. Smola, K.R. Muller, *Kernel Principal Component Analysis,* Bernhard SchÃűlkopf, Christopher J. C. Burges, Alexander J. Smola

(Eds.), Advances in Kernel Methods-Support Vector Learning, 1999, MIT Press Cambridge, MA, USA, pp. 327 – 352.

[9] A. Gorban, B. Kegl, D. Wunsch, A. Zinovyev, *Principal Manifolds for Data Visualisation and Dimension Reduction,* LNCSE 58, Springer, Berlin - Heidelberg - New York, 2007.

[10] S. T. Roweis and L. K. Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding," *Science, vol. 290, December 2000, pp. 2323 – 2326.*

[11] M. Belkin and P.Niyogi, "Laplacian Eigenmaps for dimensionality reduction and data representation," *Neural Computation, 2003.*

[12] D. L. Donoho, C. Grimes, "Hessian Eigenmaps: new locally linear embedding techniques for high-dimensional data," *Proc. of the National Academy of Sciences, 2003.*

[13] J. Ham, D. D. Lee, S. Mika, B. SchÃűlkopf "A kernel view of the dimensionality reduction of manifolds," *Proc. of* 21^{st} *Intern. Conf. on Machine Learning, 2004.*

[14] E. Moulines, P. Duhamel, J. F. Cardoso, and S. Mayrargue, "Subspace methods for the blind identification of multichannel FIR filters," *IEEE Trans. Signal Processing, vol. 43, pp. 516 525, Feb. 1995.*

[15] G. Xu, H. Liu, L. Tong, and T. Kailath, "A least-squares approach to blind channel identification" *IEEE Trans. Signal Processing, vol. 43, pp. 2982 2993, Dec. 1995.*

[16] D. Slock, "Blind fractionally-spaced equalization, perfect reconstruction filterbanks, and multilinear prediction" *in Proc. ICASSP, Adelaide, Australia, Apr. 1994.*

[17] P. V. Overshee and B. D. Moor, "Subspace algorithms for the stochastic identification problem" *Automatica, vol. 29, pp. 649 - 660, 1993.*

[18] E. Moulines, P. Duhamel, J. Cardoso, and S. Mayrargue "Subspace methods for the blind identification of multichannel FIR filters," *IEEE Trans. Signal Processing, vol. 43, pp. 516 - 526, Feb. 1995.*

[19] M. Viberg, "Subspace-based methods for the identification of linear time-invariant systems," *Automatica, vol. 31, no. 12, pp. 1835 - 1853, 1995.*

[20] Y. Wu and K. W. Tam "On Determination of the Number of Signals in Spatially Correlated Noise," *IEEE Trans. Signal Processing, vol. 46, no. 11, pp. 3023 - 3029 November 1998.*

[21] Y. P. Hong, C.T. Pan, "Rank-Revealing QR Factorizations and the Singular Value Decomposition", *Mathematics of Computation, vol. 58, No. 197 (Jan. 1992), pp. 213-232.*

[22] N. Johson, S. Kotz "Distributions in Statstics: Continuous Univariate Distributions-2" *John Wiley and Sons, 1970, pp. 130-148.*

[23] K. M. Wong, Q. T. Zhang, J. P. Reilly, and P. Yip, , "A new criterion for the determination of the number of signals in high-resolution array processing", *Advanced algorithms and architectures for signal processing III; Proceedings of the Meeting, San Diego, CA, Aug. 15-17, 1988, pp. 352 - 357*

[24] A. P. Liavas, P. A. Regalia and J. P. Delmas, "Blind Channel Approximation: Effective Channel Order Determination," *IEEE Trans. Signal Processing, vol. 47, no. 12, December 1999.*

[25] J. Rissanen "Modeling by shortest data description," *Automatica, vol. 14, no. 5, pp. 465471, 1978.*

[26] J. Rissanen "Estimation of structure by minimum description length," *Circuits, Syst. Signal Process., vol. 1, no. 4, pp. 395406, 1982.*

[27] E. Fishler and H. V. Poor "Estimation of the Number of Sources in Unbalanced Arrays via Information Theoretic Criteria" *IEEE Trans. on signal processing, vol. 53, no. 9, Sept. 2005, pp. 3543-3553.*

[28] S. A. Kassam, H. V. Poor "Robust Techniques for Signal Processing: A Survey," *Proceedings of the IEEE Vol.73, Issue 3, March 1985 Page(s):433 - 481.*

[29] N. A. Campbell, "Robust procedures in multivariate analysis I: Robust covariance estimation," *Applied Statistics*, 29(3):231 2137, January 1980.

[30] L. L. Scharf and B. Friedlander "Matched Subspace Detectors," *IEEE Trans. Signal Proc., vol. 42, no. 8, pp. 2446 – 2157, August 1994.*

[31] Specim, Spectral Imaging LTD., *www.specim.fi*

[32] I. S. Reed and X. Yu "Adaptive multiple-band CFAR detection of an optical pattern with unknown spectral distribution," *IEEE Trans. Acoust., Speech, Signal Process., vol. 38, no. 1, pp. 1760 – 1770, Oct. 1990.*

[33] H. Kwon and N. M. Nasrabadi "Kernel RX-Algorithm: A Nonlinear Anomaly Detector for Hyperspectral Imagery," *IEEE Trans on Geoscience and Remote Sensing, vol. 43, no. 2, pp. 388 – 397, Feb. 2005.*

[34] M. Zontak and I. Cohen, "Defect detection in patterned wafers using anisotropic kernels," *Machine Vision and Applications, Springer-Verlag 2008..*

[35] H. Ren and C.I. Chang "Automatic spectral target recognition in hyperspectral imagery," *IEEE Trans. Aerosp. Electron. Syst., vol. 39, pp. 1232 – 1249, Oct. 2003.*

[36] A. Banerjee, P. Burlina, and C. Diehl "A Support Vector Method for Anomaly Detection in Hyperspectral Imagery" *IEEE Trans on Geoscience and Remote Sensing, vol. 44, no. 8, pp. 2282 – 2291, Aug 2006*

[37] S. G. Beaven, D. Stein, and L. E. Hoff "Comparison of Gaussian mixture and linear mixture models for classification of hyperspectral data," *in Proc. IGARSS, Honolulu, HI, pp. 1597 – 1599, Jul. 2000*

[38] D. Stein, S. Beaven, L. E. Hoff, E. Winter, A. Shaum, and A. D. Stocker "Anomaly detection from hyperspectral imagery," *IEEE Signal Process. Mag., vol. 19, no. 1, pp. 58 - 69, Jan. 2002.*

[39] P. Stoica and Y. Selen. "Model-order selection: a review of information criterion rules", *Signal Processing Magazine, IEEE, vol. 21, Issue 4, July 2004, pp. 36-47.*

[40] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-33, pp. 387-392, Apr. 1985.*

[41] M. Tipping and C. Bishop "Probabilistic principal component analysis", *Journal of the Royal Statistical Society: Series B (Statistical Methodology), Vol. 61, No. 3, pp. 611 - 622, 1999*

[42] C-I Chang and Q. Du "Estimation of Number of Spectrally Distinct Signal Sources in Hyperspectral Imagery", *IEEE Trans. on Geoscience and Remote Sensing, Vol. 42, No. 3, pp. 608 - 619, March 2004*

[43] D. Manolakis and G. Shaw "Detection algorithms for hyperspectral imaging applications," *IEEE Signal Proc. Mag., pp. 29-43, Jan. 2002.*

[44] B. Thai and G. Healey "Invariant subpixel material detection in hyperspectral imagery", *IEEE Trans. on Geoscience and Remote Sens., vol. 40, no. 3, pp. 599-608, March 2002*

[45] Q. Du and and C.-I Chang "A signal-decomposed and interference-annihilated approach to hyperspectral target detection,", *IEEE Trans. on Geoscience and Remote Sens. vol. 42, no. 4, pp. 892-906, April 2004.*

[46] N.I. Ramey and M. Scoumekh "Hyperspectral anomaly detection within the signal subspace,", *IEEE Trans. Geoscience and Remote Sens. Let., vol. 3, no. 3, pp. 312-316, July 2006.*

[47] M. J. Carlotto "A Cluster-Based Approach for Detecting Man-Made Objects and Changes in Imagery" *IEEE Trans. on Geoscience and Remote Sensing, vol. 43, no. 2, pp. 374-386 Feb. 2005.*

[48] L.L. Scharf *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis,* Addison-Welsey Publishing Company, 1993.

[49] H. Faraji and W. J. MacLean "Noise Removal in Digital Images" *IEEE Trans. Image Proc., vol. 15, no. 9, pp. 2676-2685 Sept. 2006.*

[50] P.J. Huber, *Robust Statistics,* Wiley: New York 1981.

[51] M. E. Winter and E. M. Winter, "Comparison of approaches for determinig end-members in hyperspectral data," *Aerospace Conference Proceedings, IEEE*, vol. 3, pp. 305 – 313, March 2000.

[52] M. Leadbetter, *Extremes and related properties of random sequences and processes*, Springer Series in Statistics, 1982.

[53] S.I. Resnick, *Extreme Values, Regular Variation, and Point Process*, Springer Verlag, New York., 1987.

[54] Stuart Coles, *An Introduction to Statistical Modeling of Extreme Values*, Springer Series in Statistics., 2001.

[55] Gumbel, E.J., *Statistics of Extremes.*, Columbia University Press., 1958.

[56] C.I. Chang and Q. Du "Interference and noise-adjusted principal components analysis." *IEEE Trans. on Geoscience and Remote Sensing, Vol 37, No. 5, pp. 2387 – 2396, 1999*

[57] D.L. Donoho, "De-noising by soft-thresholding", *IEEE Trans. Inform. Theory, vol. 41, no. 3, pp. 613-627, May 1995.*

[58] D. Manolakis, C. Siracusa, and G. Shaw "Hyperspectral Subpixel Target Detection Using the Linear Mixing Model", *IEEE Trans. Geoscience and Remote Sensing, vol. 39, no. 7, pp. 1232-1249, July 2001.*

[59] J. Dias and J. Nascimento "Estimation of Signal Subspace on Hyperspectral Data", *in Proc. of SPIE, vol. 5982, pp. 191-198, Bruges, Belgium, September 2005.*

[60] W. X. Zheng "A Least-Squares Based Method for Autoregressive Signals in the Presence of Noise" *IEEE Trans. on circ. and syst.II: analog and digital signal processing, vol. 46, no. 1, pp. 81-84 Jan. 1999.*

[61] Jos M.P. Nascimento, Jos M.B. Dias, "Signal Subspace Identification in Hyperspectral Linear Mixtures", *Lecture Notes in Computer Science, Vol.3523, Jan 2005, pp. 207 - 214*

[62] A. Edelman, T. A. Arias and S. T. Smith "The Geometry of Algorithms with Orthogonality Constraints," *Siam J. Matrix Anal. Appl., vol. 20, no. 2, pp. 303-353, 1998.*

[63] A. Edelman, T. A. Arias and S. T. Smith "A Comprehensive Introduction to Differential Geometry," *vols. 1âĂŞ3, 2nd ed., Publish or Perish, Houston, TX, 1979.*

[64] C. T. Kelley, Iterative Methods for Optimization, Siam, 1999.

[65] J. A. Snyman "Practical Mathematical Optimization: An Introduction to Basic Optimization Theory and Classical and New Gradient-Based Algorithms," *Springer Publishing (2005).*

[66] S. Boyd, L. Vandenberghe Convex Optimization Cambridge University Press, March 2004.

[67] G. H. Golub and D. P. OâĂŹLeary "Some History of the Conjugate Gradient and Lanczos Algo- rithms," *1948âĂŞ1976, SIAM Review 31 (1989), no. 1, pp. 50âĂŞ102.*

[68] S. T. Smith "Optimization techniques on Riemannian manifolds," *Fields Institute Communications, vol. 3, AMS, Providence, RI, 1994, pp. 113âĂŞ146*

[69] A. Björck and G. H. Golub "Numerical Methods for Computing Angles Between Linear Subspaces," *Mathematics of Computataion, vol. 27, no. 123, July 1973*

[70] G.W. Stewart *Matrix Algorithms Volume II: Eigensystems,* SIAM, Philadelphia, PA, 2001.

[71] G. W. Stewart and J. G. Sun, "Matrix Perturbation Theory", *Academic Press, Boston, MA, 1990.*

[72] D. Stein, S. Beaven, L. E. Hoff, E. Winter, A. Shaum, and A. D. Stocker "Anomaly detection from hyperspectral imagery," *IEEE Signal Process. Mag., vol. 19, no. 1, pp. 58 – 69, Jan. 2002.*

[73] G.F. Hughes "On The Mean Accuracy Of Statistical Pattem Recognizers," *IEEE Trans. Infor. Theory, vol. IT-14, NO. I , pp 55 – 63, 1968.*

[74] S. B. Serpico, M. DâĂŹInca, and G. Moser, "Design of spectral channels for hyperspectral image classification," *Proc. IGARSS, vol. 2, pp. 956 – 959, 2004.*

[75] Fodor I. K. "A survey of dimension reduction techniques" *LLNL technical report, June 2002.*

[76] S. Kumar, J. Ghosh, and M. Crawford, "Best-bases feature extraction algorithms for classification of hyperspectral data," *IEEE Trans. Geoscience and Remote Sensing, vol. 39, no. 7, pp. 1368 – 1379, Jul. 2001.*

[77] A. C. Jensen and A. S. Solberg, "Fast Hyperspectral Feature Reduction Using Piecewise Constant Function Approximations," *IEEE Geoscience and Remote Sensing Letters, vol. 4, no. 4, Oct. 2007.*

[78] D. Ververidis and C. Kotropoulos "Gaussian Mixture Modeling by Exploiting the Mahalanobis Distance," *IEEE Trans. Sig. Proc., vol. 56, no. 7., July 2008.*

[79] G. Winkler and V. Liebscher "Smoothers for discontinuous signals," *J. Nonparametric Statist., vol. 14, no. 1/2, pp. 203 – 222, 2002.*

This page is left blank on purpose.

הפחתת יתירות באותות רב מימדיים תוך שימור אנומליות

חיבור על מחקר

לשם מילוי חלקי של הדרישות לקבלת התואר
דוקטור לפילוסופיה

אולג קוייבדה

הוגש לסנט הטכניון - מכון טכנולוגי לישראל

אייר תשס״ט חיפה מאי 2009

הכרת תודה

ברצוני להודות מקרב לב למנחה העבודה, פרופ׳ דוד מלאך, על הנחייתו האישית והמסורה, הבנתו, תרומתו הגדולה למחקר ותמיכתו במשך כל שנות המחקר. הייתה לי זכות גדולה לעבוד לצידו של חוקר ברמתו ושל אדם חם ותומך כמוהו.

תודות רבות נתונות גם למנחה הנוסף, דר׳ מאיר בר-זוהר, על הערותיו הקונסטרוקטיביות והדייקניות ועל ניסיונו המעשי שתרם רבות לניסוח ופיתרון בעיות ברמה הנדסית גבוהה. תמיכתו הייתה חשובה לי מאוד לאורך כל המחקר.

במשך המחקר עבדתי לצידם של עמיתים רבים והנני רוצה להביע הערכה רבה לתמיכתם. תודות חמות לנמרוד, זיוה, אבי ויאיר אשר עזרו לי בעבודתי במעבדה לעיבוד אותות ותמונות (SIPL). הנני רוצה לציין במיוחד את רמתו המקצועית של נמרוד, המהנדס הראשי של המעבדה, שלדעתי מצליח להדביק את כולם בשאיפתו למצויינות ועמידה בסטנדרטים גבוהים.

תודותיי באהבה לאשתי, אלונה, על הרבה אהבה, תמיכה וסובלנות שליוו אותי בתקופה זאת.

אני מודה לטכניון על התמיכה הכספית הנדיבה בהשתלמותי.

תקציר

הפחתת יתירות הוא נושא מרכזי בקהילות המקצועיות שמטפלות באותות רב-מימדיים. ישנם לא מעט יישומים המשתמשים במערכי חיישנים, שבהם המימדיות האמיתית של המידע היא נמוכה בהרבה ממימדיות וקטורי האות הנקלט. תת-המרחב של האות האמיתי נושא מידע חשוב על המבנה הפיסיקלי של מקורות האות ולכן, בהרבה יישומים, ישנה חשיבות רבה לשערך את תת המרחב של האות האמיתי. מכיוון שהאות הנקלט כולל בדרך כלל גם רעש, הטלתו על תת המרחב המשוערך מאפשרת גם הורדת תרומת הרעש במימדים שלא נושאים מידע חשוב.

גישות קלאסיות לשערוך תת המרחב הפיסיקלי של האות מבוססות על מזעור אנרגית שגיאת הייצוג, או, במילים אחרות, מזעור נורמת ℓ_2 של שגיאת הייצוג. גישה זאת מאפשרת למצוא תת-מרחב אשר יכלול בתוכו את רוב אנרגיית האות. למרבה הצער, הגישה של מזעור של נורמת ℓ_2 אינו יכולה להביא למציאת ייצוג טוב לתופעות אנומליות באות הפיסיקלי, מכיוון שהתרומה האנרגטית של הוקטורים האנומליים, במובן של נורמת ℓ_2, היא בדרך כלל נמוכה בהשוואה לתרומה האנרגטית של הרעש. תרומה זו יכולה לגרום להסטת תת-המרחב המשוערך, כך שיחמיץ את האנומליות וישאירן מחוצה לו . לכן, הגישות הקלאסיות שמבוססות על נורמת ℓ_2 מוגבלות ביכולתן לייצג וקטורים אנומליים. כך, למשל, בבעיית גילוי אנומליות בתמונות היפרספרטרליות מעוניינים, מצד אחד, לצמצם את המימדים בכדי לנקות את האות מרעש, ומצד שני רוצים לשמר את האנומליות בתוך המרחב המשוערך כדי לגלותן בהמשך.

במחקר זה אנו מתמודדים עם בעיית הפחתת היתירות באותות רב-ממדיים שיכולים להכיל וקטורים אנומליים. הנושא המרכזי במחקר הוא, לפיכך, שימור האנומליות לאחר הורדת היתירות. בחלקו הראשון של המחקר, אנו מציעים גישה חדשנית שנקראת Maximum Orthogonal Complements Analysis (MOCA), אשר משלבת נורמות ℓ_2 ו-ℓ_∞ לשם שערוך תת-המרחב והמימד של האות, כולל האנומליות. הגישה מורכבת משתי פעולות עיקריות: הפעולה הראשונה מטפלת בשערוך של תת-מרחב האות בהנחת מימד נתון, אשר ממזער את המכסימום של נורמות ℓ_2 של שגיאת הייצוג של וקטורי האות הנקלט, לאחר הטלתם

לתת-המרחב המשוערך. מכיוון שהמדד החדש משלב פעולות מכסימום, קרי - נורמת ℓ_∞, עם נורמת ℓ_2, הוא מכונה נורמת $\ell_{2,\infty}$. $\ell_{2,\infty}$ הוא אכן נורמה מכיוון שמכסימום של נורמות גם היא נורמה. הפעולה השנייה מטפלת בקביעת נכונות היפותזת המימד, שעבורו נתקבל תת-המרחב. המימד נקבע על ידי הפעלת תוצאות של תורת הערכים הקיצוניים (Extreme Value Theory) על מודל פילוג של נורמת $\ell_{2,\infty}$ של הרעש. שתי פעולות אלה מבוצעות לסירוגין תוך כדי שימוש באלגוריתם חמדני תת-אופטימלי למציאת תת-מרחב אשר הופך את הגישה לבת מימוש בבעיות מעשיות. בסוף התהליך מתקבלים תת-המרחב והמימד המשוערכים של האות הפיסיקלי, שכולל אנומליות, וכן תת-מרחב ומימד של האנומליות בנפרד. בסימולציות שערכנו עבור יחסי אות לרעש שונים (בתחום ערכים החשובים מבחינה מעשית) של וקטורים אנומליים, הגישה המוצעת הביאה לתוצאות טובות, בעוד הגישות הקלסיות, המבוססות על SVD לשערוך תת-המרחב ועל MDL לשערוך המימד, נכשלו במציאת תת-המרחב והמימד של האות, שכולל אנומליות ותרומת תהליך הרקע.

היתרונות של שילוב נורמות ℓ_2 ו-ℓ_∞ אינן מסתכמות רק באלגוריתם להורדת יתירות תוך שימור אנומליות. אנו מראים גם שניתן להתאים שילוב זה ואת המבנה של אלגוריתם MOCA למציאת תת-מרחב האות (כולל אנומליות) לשם פיתוח אלגוריתם לגילוי, הפרדה ושערוך אוכלוסיות אנומליות בתמונות היפרספקטרליות. אלגוריתם זה שייך למשפחה של אלגוריתמי גילוי אנומליות גלובליים מכיוון שהוא משתמש בכל התמונה ההיפרספקטרלית כדי לשערך מודל תהליך הרקע ומאפשר כך לגלות אנומליות כוקטורים אשר אינם מסכימים עם מודל זה. תהליך זה קרוי בעבודה "אלגוריתם למיצוי והפרדה של אנומליות", Anomaly Detection and Discrimination Algorithm (AXDA). הרעיון המרכזי של AXDA מבוסס על הקטנה איטרטיבית של מימד מרחב האנומליות שנמצא על ידי MOCA. פעולה זאת גורמת לשגיאה בייצוג של הוקטורים האנומליים הקשורים למימד שהופחת ומאפשרת גילוי האנומליות בעזרת אנליזה של נורמת $\ell_{2,\infty}$ של שגיאות הייצוג. תוצאה נלווית של AXDA היא שבתום התהליך מתקבל גם שערוך של תת-מרחב הרקע הנקי מאנומליות. בסימולציות עם תמונות היפרספקטראליות אמיתיות, ראינו כי AXDA נותן תוצאות טובות יותר מאלה של GMRX ו-MSD – האלגוריתמים לגילוי אנומליות גלובליים הקלאסיים שקיימים בספרות.

כדי להשלים את הגישה לשערוך תת-מרחב שמשמר אנומליות, הרחבנו את MOCA על ידי פיתוח אלגוריתם אופטימלי למזעור נורמת $\ell_{2,\infty}$ של שגיאות הייצוג. אלגוריתם זה מכונה "שערוך תת-מרחב אופטימאלי במובן של מזעור המשלים האורתוגונאלי המכסימלי" Maximum of Orthogonal Complements Optimal Subspace Estimation (MOOSE). האופטימיזציה מבוצעת על ידי גישת natural conjugate gradient learning שפועלת על קבוצת תת-מרחבים n-ממדיים ששייכים ל $\mathbb{R}^m, m > n$. קבוצה זו מהווה יריעה של גרסמן (Grassmann manifold). בהיותו תהליך אופטימיזציה מקומי, MOOSE רגיש לבעיית התכנסות למינימום מקומי. לכן אתחול טוב הוא חיוני ביותר. אנו משתמשים בתוצאות של MOCA כתנאי התחלה של MOOSE מכיוון שהאלגוריתם החמדני של MOCA משתמש בשיקולים גלובליים כדי למצוא תת-מרחב תת-אופטימלי שקרוב לתת-המרחב האופטימלי. אנחנו מראים כי MOCA ו-MOOSE מניבים תוצאות טובות משמעותית מאלה של SVD, במובן של מזעור נורמת $\ell_{2,\infty}$ של שגיאת הייצוג באותות שהוכנו בסימולציות וגם בתמונות היפרספקטרליות אמיתיות. למרות ש-MOCA מניב תוצאות פחות טובות מ-MOOSE, הן קרובות לתוצאות של MOOSE. זאת אומרת, שבאופן מעשי, האלגוריתם החמדני לשערוך תת-מרחב של MOCA, מהווה בחירה טובה מכיוון שהוא הרבה יותר פשוט מבחינה חישובית.

העושר של מידע ספקטראלי המצוי בתמונות היפרספקטראליות מאפשר לקבל ביצועים טובים בהפעלת אלגוריתמים לגילוי אנומליות כגון AXDA. אולם, מערכות היפרספקטראליות עדיין אינן נוחות לשימוש ביישומים שדורשים ניידות מכיוון שהן יקרות, כבדות, וצורכות הרבה הספק. לכן, ישנה דרישה למערכות מולטיספקטראליות שהן יותר קומפקטיות במידה רבה, למרות שהן מספקות מספר מוגבל של ערוצים ספקטראליים. תכנון נכון של המסננים המולטיספקטראליים יכול להיות חיוני לפיתוח אלגוריתמים לגילוי אנומליות בתמונות מולטיספקראליות. המחקר מסתכם בהצעת גישה חדשנית לתכנון מסננים מולטיספקטראליים שמטרתם לשפר ביצועים של אלגוריתמים לגילוי אנומליות. גישה זאת מבוססת על עיבוד תמונה היפרספקטראלית אופיינית שמשמשת כקלט לאלגוריתמים לגילוי אנומליות. למעשה, בעיית תכנון מסננים מולטיספקטראליים ניתנת לניסוח כבעיית הפחתת ערוצים בתמונות היפרספקטראליות. הורדת מספר הערוצים מבוצעת על ידי החלפת קבוצות של ערוצים היפרספקטראליים סמוכים על ידי הממוצע שלהם.

החלוקה האופטימלית של ערוצים היפרספקטראליים מתקבלת על ידי מזעור מרחק מהאלנוביס (Mahalanobis) מכסימלי (שמסומן על ידינו ב-MXMN - minimizing Maximum of Mahalanobis Norms) של השגיאות המתקבלות מייצוג ערוצים היפרספקטראליים על ידי ערכים קבועים. האופטימיזציה מבוצעת על ידי אלגוריתם תכנות דינאמי בדומה לאלגוריתם שהוצע בגישה הקרוייה Fast Hyperspectral Feature Reduction (FFR). גם לפי גישת FFR הערוצים ההיפרספקטראליים מצומצמים על ידי החלפת קבוצות של הערוצים הסמוכים בממוצע שלהם. אולם הקריטריון לחלוקה אופטימלית של הערוצים הוא אנרגיית שגיאת הייצוג. כמו שהזכרנו לעיל, קריטריון זה אינו רגיש לאנומליות ולכן עלול לפגוע בייצוגן על ידי הערוצים המצומצמים. מזעור מטיפוס MXMN של שגיאות הייצוג גורם להקטנת תרומת האנומליות לשגיאות אלה. הדבר מאפשר שימור מאפיינים אנומליים בערוצים המצומצמים, כשישנן אנומליות בתמונה שעליה מבצעים את התהליך. במקרה שתהליך צמצום הערוצים מופעל על תמונה שאינה כוללת אנומליות, מזעור מטיפוס MXMN של שגיאות הייצוג עדיין מביאה להחלקת הערוצים הספקטראליים שכוללים clutter השייך לתהליך הרקע. הערוצים האלה אינם רצויים למטרת גילויי אנומליות מכיוון שהם עלולים למסך תרומות עדינות של אנומליות בערוצים אחרים שבהם ה- clutter של תהליך הרקע מתבטא פחות. תוצאות הרצת MXMN, FFR ו-SVD למטרת צמצום הערוצים והפעלת אלגוריתם RX – האלגוריתם הקלאסי לגילוי אנומליות בתמונות מולטיספקטראליות, על התמונות לאחר צמצום הערוצים, מראות ש-MXMN מניב תוצאות יותר טובות מ-FFR ו-SVD ואף מהפעלת RX על האות המקורי ללא צמצום הערוצים בתחום החשוב של ערכי גילוי שווא נמוכים. הביצועים נמדדו על ידי בחינת עקומות Receiver Operating Characteristic (ROC) של RX שהופעל על האות לאחר צמצום הערוצים.