UiT
THE ARCTIC
UNIVERSITY
OF NORWAY

Faculty of Science and Technology
Department of Physics and Technology

Forest Health Monitoring by Detection of Biotic Agents in Satellite Images

—

Stine Hansen

EOM-3901 Master's Thesis in Energy, Climate and Environment, 30 SP

June 2018

Acknowledgements

First and foremost, I would like to thank my supervisor, Associate professor Stian Normann Anfinsen for the opportunity to write this master's thesis and for his invaluable assistance and guidance throughout this entire project.

I want to thank Science [&] Technology AS for assistance with the data, and Dr. Vasco Mantas at the University of Coimbra for supplying ground truth and good advice.

Further, I would also like to thank all the great people I have come to know during my studies here in Tromsø. Thank you for making this experience unforgettable!

To my family, my parents and my sister, to whom I can (and do) call with good news, bad news and no news at all: Thank you for your endless love and support.

Last but not least, I want to thank Olaf for all help and support, and for bearing with me these past months.

Stine Hansen
Tromsø, May 2018

Abstract

Since 1999, the *Pine Wood Nematode* (PWN) has spread, infected and damaged growing areas of pine forest in Portugal. The pest is subject to strict quarantine measures, that require forest owners to register, fell and dispose of infected trees. As remote sensing from satellites provides repetitive and consistent data sets with high temporal resolution and large spatial coverage, a surveillance based on satellite images would be advantageous. Moreover, since multispectral images from the European Space Agency's (ESA's) Sentinel-2 mission are freely available, a monitoring based on these data is preferrable. The most commonly used tool for monitoring vegetation by remote sensing is the *Normalized Difference Vegetation Index* (NDVI) [Franke et al., 2006]. However, the infection by PWN may appear more prominent in other combinations of channels, and by introducing machine learning techniques, allowing for the exploration of changes in *all* spectral bands, one may end up with a better discrimination. This thesis is intended as a feasibility study that aims to explore the viability of an operational PWN detection system based on Sentinel-2 data. To this purpose, traditional feature extraction algorithms, including *Fisher's linear discriminant analysis* (FLDA), and *sparse linear discriminant analysis* (SLDA) are examined, as well as some spectral unmixing algorithms, including *iterated constrained energy minimization* (ICEM), *mixture-tuned matched filtering* (MTMF) and the *2D-Corr-NLS* algorithm. The methods are examined with respect to the impact of spatial and spectral resolution, and the performances are measured and compared using a *constant miss rate* (CMR) detector. Multisource data with different spatial and spectral resolution are used in experiments to investigate how these resolutions constrain successful PWN detection. The results show that the spatial resolution of the Sentinel-2 data is too low for the selected methods to be useful for PWN detection, whilst it remains an open question whether the spectral resolution of Sentinel-2 is a limitation.

Contents

List of Figures

List of Tables

List of Notation

Throughout this thesis, the following notations have been used:

- *Vectors* are written in lower case **bold**.
- *Scalars* are written in regular font.
- *Matrices* are written in upper case **bold**.

Further, $\mathbf{1}$ should be understood as the all-1 column vector of appropriate size, and $\boldsymbol{I}$ as the identity matrix of appropriate size.

Also, the following index notation is used for 1D, 2D and 3D-matrices:

$$\boldsymbol{A} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix},$$

$$\boldsymbol{B} = \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{12} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix},$$

$$\boldsymbol{C} = \begin{bmatrix} \begin{bmatrix} c_{111} & c_{121} & c_{131} \\ c_{211} & c_{221} & c_{231} \\ c_{311} & c_{321} & c_{331} \end{bmatrix} \\ \begin{bmatrix} c_{112} & c_{122} & c_{132} \\ c_{212} & c_{222} & c_{232} \\ c_{312} & c_{322} & c_{332} \end{bmatrix} \end{bmatrix}.$$

1 Introduction

This introductory chapter starts by motivating the work of the thesis. It then provides a brief overview of the state of the art before it goes on to the contributions of the project and presents the topic of research. Finally, a short summary of the preceding chapters is given.

1.1 Motivation

The establishment of an invasive species can potentially cause serious damage to an environment. The consequences of an intrusion may be difficult to predict and can include both threats to the health of people, animals and plants, as well as threats to the economy [NEA, 2017].

The pine wood nematode (PWN, *Bursaphelenchus xylophilus*) entered Europe through Portugal in 1999, and by 2008 it had made its way to Spain. This intrusive species infects pine trees, and brings pine wilt and death to trees which are significantly stressed[1]. The PWN is for this reason considered to be one of the top five most damaging biotic agents in European forests [BIO IS, 2011, p. 112, 69, 43], and all European Union (EU) member states are by Commission Decision 2006/133/EC required to do annual surveys to ensure that the PWN has not entered their forests [BIO IS, 2011, p. 118].

1. Stress already inflicted by other factors, such as drought, wind, temperature, etc.

Remote sensing from space provides the opportunity to collect repetitive and consistent multispectral data sets with high temporal resolution and large spatial coverage. Detection of PWN from satellite data could thus serve as a fast and cost effective alternative to manual ground surveys and aerial photography missions.

Since multispectral images from the European Space Agency's (ESA's) Sentinel-2 mission[2] are freely available[3] at a 10 m spatial resolution and five days temporal resolution[4] [ESA, n.d.-b], a surveillance based on detection from these data would be advantageous. Moreover, an operational forest monitoring *depends* on both economic and temporal availability of data. However, at such a coarse resolution, one pixel will usually contain multiple trees (some of which may not be infected) and/or other objects (buildings, bare soil, grass, etc.), and the signal originating from the potentially infected tree may be weak or even undetectable.

1.2 State of the Art

The only way to establish for certain whether a tree is infected by PWN or not, is to physically examine it. However, a direct consequence of the pine wilt disease is dehydration [Kuroda, 2008], which again leads to changes in other physical properties like color, texture and internal structure of the needles. These are the changes that are being exploited in PWN detection by remote sensing.

There has been some research on the detection of PWN by remote sensing, including a recent, high budget pilot study conducted by the Joint Research Centre on behalf of the European Commission, which resulted in the comprehensive report "The feasibility of detecting trees affected by the Pine Wood Nematode using remote sensing" [Beck, 2015]. This study included spectral analysis of aerial images (hyperspectral, near infrared and thermal infrared) as well as high resolution satellite data (WorldView2/3 and Pleiades), where the intensities of the reflection in different spectral channels were compared. A variety of spectral vegetation indices were calculated and used to examine the health of a trees.

Other previous works worth noting are some studies on the detectability of

2. The mission comprises a constellation of the two polar-orbiting satellites Sentinel-2A and Sentinel-2B [ESA, n.d.-b].
3. Images can be downloaded from `https://scihub.copernicus.eu`.
4. At equator, assuming cloud-free conditions.

early signs of stress in coniferous species, where the red-edge information proved to be helpful:

- "Broadband, red-edge information from satellites improves early stress detection in a New Mexico conifer woodland", [Eitel et al., 2011].
- "Early Detection of Bark Beetle Green Attack Using TerraSAR-X and RapidEye Data", [Ortiz et al., 2013].
- "Evaluating the impact of red-edge band from Rapideye image for classifying insect defoliation levels", [Adelabu et al., 2014].

One essential feature that is common to the aforementioned studies is that they regard *mid-* to *high* spatial resolution (< 6.5 m) satellite data and aerial images. This type of data is not free (or inexpensive) and it is questionable whether it is possible to run an operational monitoring service based on commercial data.

The Silvisense project (2016-2017), an ESA funded collaborative project between UiT The Arctic University of Norway, the University of Coimbra (Portugal) and Science [&] Technology AS (Norway), investigated the detectibility of Pine Wilt disease (ESA contract no. 4000117722/16/INB).

Based on NDVI composites of Landsat images and MODIS images from 2001 to 2012, time series of homogeneous pixels (at the MODIS resolution of 250 m $\times$ 250 m) of pine forest (>90 % pine) were extracted. These time series were analysed to extract a seasonal component, a trend component and a random component. The algorithm that was developed then compares satellite measurements at two dates: one pre- and one post-decline. Based on the seasonal trend, the pixel's pre-decline value is used to predict its expected post-decline value, if not infected. The predicted value is then compared to the measured post-decline value and if the difference exceeds a threshold, a change is detected.

The project resulted in a pilot forest monitoring service to help forest owners in Portugal monitor their forests (`https://www.silvisense.com`). However, some issues remain unsolved by the Silvisense project: The number of false alarms is high and the detection capability (especially for small objects and weak/mixed signals) is limited.

1.3 Contributions

This master's project is an extension of the Silvisense project, conducted in collaboration with project partners at the University of Coimbra and Science [&] Technology AS, and is intended as a feasibility study that aims to explore the viability of PWN detection by Sentinel-2 data. To this purpose, two different main approaches are pursued. The first one is detection based on classical *feature extraction*, and seeks alternative test statistics to the NDVI, that could potentially discriminate better between vegetation of slightly different spectral signature. Here, two different feature extraction algorithms, including Fisher linear discriminant analysis (FLDA) and sparse linear discriminant analysis (SLDA), are explored. The second approach involves *spectral unmixing*, where the goal is to unmix the signal from each pixel to retrieve the proportion of target material in each pixel. To this end, three different unmixing algorithms, including the *2D-Corr-NLS* algorithm, the *iterated constrained energy minimization* (ICEM) algorithm and the *mixture-tuned matched filtering* (MTMF) algorithm, are examined.

In order to compare the proposed methods and explore the effects of spectral and spatial resolution in a quantitative manner, the methods are carried out on a fully ground truthed data set as well as the PWN data set, and a constant miss rate detector is used to evaluate the performances of the methods.

Further, this thesis will also contribute through comments, analyzes and interpretations of the methods, as well as highlight key points in the original articles.

An operational PWN detection system is assumed to be a twofold system where the first step is a pre-processing step, applied to narrow the search down to pixels that have *previously* shown "healthy pine"-like signatures. In this way, the detection problem becomes a two-class problem where the background class is *healthy forest* and the target class is *infected pines*. This step should be assumed before any further processing. However, to limit the scope of this thesis, this step is replaced with a simple NDVI threshold operation.

1.4 Research Questions

The aim of this study is to examine the feasibility of PWN detection from Sentinel-2 images based on signal processing and machine learning techniques such as spectral unmixing and classical feature extraction methods. In particular, this thesis will address two main research questions:

1. *Is the spatial resolution of the Sentinel-2 data a limitation to successful use of the selected methods for PWN detection?*

2. *Is the spectral resolution of the Sentinel-2 data a limitation to successful use of the selected methods for PWN detection, and do the red-edge bands help the discrimination?*

Furthermore, this thesis will also examine and compare the methods' performances on data with varying spatial and spectral resolution in order to decide what type of data they require to provide good results.

1.5 Structure of Thesis

Chapter 2 reviews the essential principles of passive remote sensing and describes passive remote sensing of PWN, the use of vegetation indices, their limitations and the approach of this thesis to obtain a better discrimination between infected and healthy trees.

Chapter 3 introduces the application and theory of classical supervised feature extraction and presents the specific algorithms to be tested in this thesis.

Chapter 4 presents the application and theory of spectral unmixing. The first part describes the mixture models, with focus on the linear mixing model, while the second part discusses the different unmixing algorithms to be tested in this thesis.

Chapter 5 presents the basic theory on hypothesis testing, introduces the constant miss rate detector and discusses the detector output distributions for the methods described in the previous chapters. Finally, a short introduction to the receiver operating characteristics curve is given.

Chapter 6 addresses the methodology. It presents the data sets and describes the design and setup of the experiment.

Chapter 7 provides and discusses the results of the detection based on the different algorithms for both data sets.

Chapter 8 concludes the work of this thesis and suggests future work based on obstacles and findings.

Part I

Theory

2

Passive Remote Sensing

This chapter starts by presenting the underlying theory of passive remote sensing, connecting the physical processes to the detection problem. The second section moves on to describe in greater detail passive remote sensing of PWN, the use of vegetation indices, their limitations, and the approach of this thesis to obtain a better discrimination between healthy and infected trees.

2.1 Principle of Passive Remote Sensing

When an object is illuminated by an external source of electromagnetic (EM) radiation, the type of interaction between the incoming waves and the object is governed by *i*) the physical properties of the waves (mainly frequency), and *ii*) the energy levels of the material [Elachi, 2006]. If the energy of a photon,

$$E = h\nu, \tag{2.1}$$

where h is Planck's constant and ν is the frequency of the wave, matches an energy step in the material (energy difference between two states), the energy will get absorbed. Waves with energies too low to correspond to any energy step, on the other hand, can interact by *heating* the material or (depending on the geometry and dielectric properties of the matter) *reflecting/scattering* off the surface [Elachi, 2006].

In a passive remote sensing system, the sun is the source of radiation, and the final product[1] derived from the signal received at the sensor is the spectral surface *reflectance*. Since the sun is considered a blackbody in thermal equilibrium, its radiation is described by Planck's radiation law [Elachi, 2006]

$$S(\lambda, T) = \frac{2\pi hc^2}{\lambda^5} \frac{1}{e^{hc/\lambda kT} - 1}, \tag{2.2}$$

where S is the spectral radiance, c is the speed of light, λ denotes wavelength, T is temperature and k is Boltzmann's constant. Any object is thus illuminated by waves throughout the entire spectrum (except the bands absorbed in the atmosphere), and absorbs and reflects the different wavelengths in accordance with its energy levels. In this way, the spectral reflectance (entity measured by the sensor) reveals the characteristic signature of the ground material (Figure 2.1).

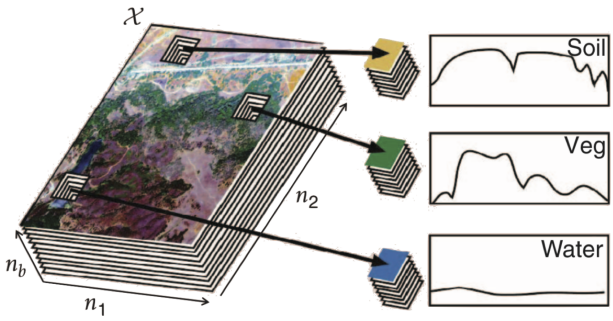

Figure 2.1: The reflectance at different wavelengths (bands) reveals the characteristic spectral signature of each material. From [Bioucas-Dias et al., 2013].

2.2 Passive Remote Sensing of the PWN

The idea of detecting PWN by remote sensing relies on the fact that the spectral signature of a tree *changes* after the infection. When a tree gets infected, it starts to dehydrate, and different physical properties change.

1. The signal received at the sensor consists of multiple *unwanted* contributions from atmospheric scattering and redirected scattering from the surroundings. The atmospheric correction (ideally) removes all unwanted paths and converts the surface reflected radiation from the pixel of interest to surface reflectance [Manolakis et al., 2014].

One visible change is the discoloration of the needles caused by the degradation of chlorophyll. Since it is the chlorophyll that accounts for most of the absorption in the visible part of the spectrum [Campbell, 2011], a difference in absorption (between healthy and infected trees) is expected. The dehydration also changes the internal structure of the needles. In a healthy needle, the structure of the mesophyll tissue accounts for the high reflectivity of near infrared radiation, while the reflectivity in a infected tree is somewhat lower. A difference is also expected in the far infrared region due to strong water absorption bands. However, it is worth noting that these differences may also occur due to atmospheric water absorption [Elachi, 2006]. Figure 2.2 shows a sketch of a typical spectral response for green vegetation with its characteristic absorption bands.

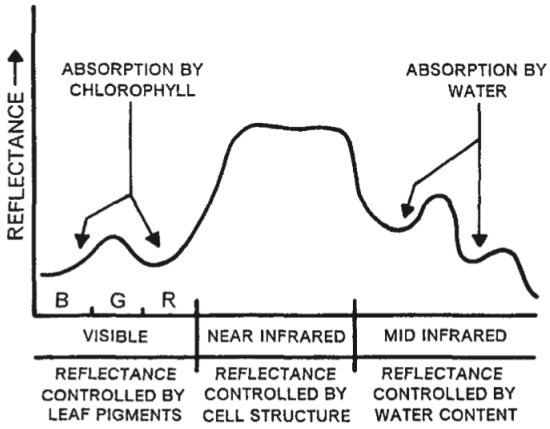

Figure 2.2: Changes in chlorophyll content may cause differences in the visible part of the spectrum, and changes in water content in the far infrared part. From [Campbell, 2011].

2.2.1 Vegetation Indices

Vegetation indices (VIs) are computed by combining intensities from different spectral bands that are known to be able to tell something about the biomass and/or the health of the vegetation. One group of VIs of particular importance for living biomass is the band ratios where the red and the infrared channels are included [Campbell, 2011].

NDVI

The *Normalized Difference Vegetation Index* (NDVI) is the most commonly used tool in vegetation monitoring from space [Franke et al., 2006]. It utilizes the well known inverse relationship between reflection in the red (RED) and near infrared (NIR) part of the spectrum for healthy vegetation [Campbell, 2011].

The test statistic is given by

$$\text{NDVI} = \frac{\text{NIR} - \text{RED}}{\text{NIR} + \text{RED}}, \tag{2.3}$$

and yields high values for healthy vegetation and lower values for stressed vegetation or non-vegetated areas.

Furthermore, an important strength of the quotients is that effects influencing *all* bands (such as angle of incidence) get canceled [Chuvieco, 2009].

Limitations

The detection of PWN by investigating differences in NDVI exploits the fact that healthy trees absorb more of the visible light and reflect more of the near infrared light (compared to that of an infected tree, as discussed in the previous section), resulting in a higher NDVI value [NASA, n.d.]. However, the spectral signatures of healthy pine trees and pine trees infected by PWN may also differ in other regions of the spectra. In particular, the red-edge bands have shown potential in the detection of early signs of stress in coniferous species [Eitel et al., 2011, Adelabu et al., 2014].

By introducing machine learning techniques, allowing for the exploitation of these changes as well, one may end up with a better discrimination between healthy and infected trees.

In the two following chapters, two different approaches to (possibly) obtain a better discrimination between healthy and infected trees are explored. Chapter 3 presents two different dimension reducing transformations where the algorithms decide which features to combine (and how) in order to reduce the pixel vectors to scalars that discriminate between infected trees and the background. Chapter 4 concerns spectral unmixing of the pixel signals to obtain the *fractional abundance map* of the target material. That is, a one-band image is produced, showing the proportion of target material (infected trees) in each pixel.

As will be seen (section 4.3.2), these approaches may under certain conditions lead to the same solution.

3 Feature Extraction

One of the thesis' main approaches to the detection problem is through classical feature extraction methods. This chapter starts by introducing the application of the dimensionality reducing methods to be discussed. Then comes a short introduction to supervised feature extraction and a presentation of the different algorithms.

3.1 Application

In this thesis, the goal is to reduce the dimension of the data to one dimension, i.e. one number per pixel (like the NDVI), in such a way that the statistical distance between pixels containing healthy and infected trees is maximized[1]. The following section gives a short introduction to supervised feature extraction and presents two supervised dimensionality reducing transformations.

1. This number is the test statistic used in the hypothesis test of the detector.

3.2 Supervised Feature Extraction Algorithms

In a multispectral image, the spectral information of the area depicted by a single pixel is typically stored as discrete values[2] in a feature vector

$$\boldsymbol{x} = [x_1, ..., x_p]^T, \tag{3.1}$$

where $[\cdot]^T$ denotes transpose and p is the number of spectral bands sampled by the sensor. Reducing the dimension refers to reducing the number of features in these vectors, and it is desirable for different reasons[3]. That is, supervised feature extraction algorithms seek the transformation

$$\boldsymbol{y} = \boldsymbol{w}^T \boldsymbol{x}, \tag{3.2}$$

from the input data, $\boldsymbol{x}$, to a lower dimension, $\boldsymbol{y} \in \mathbb{R}^m$, $m < p$, using a transformation matrix, $\boldsymbol{w} \in \mathbb{R}^{p \times m}$, learned from a set of training pairs (observations with known labels, i.e. ground truth) [Cunningham et al., 2008].

However, the dimension reduction should be done in a way that preserves as much *information* as possible. Different dimensionality reduction algorithms have different measures of information and different approaches to the selection of features. The following sections present two different supervised feature extraction algorithms, seeking a direction $\boldsymbol{w}$ to project the data onto, such that as much as possible of the discriminant information about the classes[4] is preserved.

3.2.1 Fisher Linear Discriminant Analysis

The Fisher linear discriminant analysis (FLDA) is a linear, supervised[5] transform

$$\begin{aligned} \boldsymbol{y} = \boldsymbol{w}^T \boldsymbol{x}, \quad & \boldsymbol{y} \in \mathbb{R}^m, \\ & \boldsymbol{w} \in \mathbb{R}^{p \times m}, \end{aligned} \tag{3.3}$$

which seeks to project the data, $\boldsymbol{x} \in \mathbb{R}^p$, onto a possibly[6] lower dimension, $\mathbb{R}^m$, where the within-class variance is minimized and the between-class variance is

2. The mean intensity within each spectral band that the sensor samples, discretized according to the radiometric resolution.
3. High dimensional data is demanding in the sense of storage and computational operations, but equally important is the *curse of dimensionality*, which is the general term for the problems arising due to the fact that the necessary number of data points to obtain statistical significance and reliable estimates of the model parameters increases exponentially with the dimensionality of the data [Cooper, 1995, Theodoridis, 2009].
4. The detection problem can be considered a two-class classification problem where each sample is classified as target (detected) or background.
5. Requires labeled data to compute $\boldsymbol{w}$.
6. If $m < p$, the dimension gets reduced. This is however *not* necessary.

maximized, yielding a maximized *class separation* [Theodoridis, 2009, Bishop, 2006].

A cost function that achieves these properties when maximized is the *Fisher discriminant ratio* (FDR). For a two-class problem (that is, each observation belongs to either class ω_1 or class ω_2) where the data is projected onto the real line ($\boldsymbol{y} = y$), the ratio can be expressed as [Theodoridis, 2009]

$$FDR = \frac{(\mu_1 - \mu_2)^2}{(\sigma_1^2 + \sigma_2^2)}, \tag{3.4}$$

where σ_i^2 and μ_i are the within-class variance and mean of the ith class in the transformed domain. In this way of writing, the ratio is thus expressed in the *transformed domain*. However, to be able to optimize the ratio, it is necessary to express it in the domain of the original (input) data as a function of the transformation vector $\boldsymbol{w}$. A closer look at the within- and between-class scattering matrices in the input domain will reveal this expression.

In the input space, the within-class scattering matrix can be expressed as [Theodoridis, 2009]

$$\mathbf{S}_w = \sum_{i=1}^{2} P_i \boldsymbol{\Sigma}_i, \tag{3.5}$$

where $\boldsymbol{\Sigma}_i$ is the covariance matrix of class i and P_i is the a priori probability of class i. The sum of within-class variances in the output space can be written as [Theodoridis, 2009]

$$\sigma_1^2 + \sigma_2^2 = E_{y \in \omega_1} \left\{ (y - \mu_1)^2 \right\} + E_{y \in \omega_2} \left\{ (y - \mu_2)^2 \right\}, \tag{3.6}$$

where $E_{y \in \omega_i}$ denotes the mathematical expectation with respect to class ω_i. In terms of input data, this is equal to

$$\begin{aligned} E_{\boldsymbol{x} \in \omega_1} \left\{ \boldsymbol{w}^T (\boldsymbol{x} - \boldsymbol{\mu}_1)(\boldsymbol{x} - \boldsymbol{\mu}_1)^T \boldsymbol{w} \right\} + E_{\boldsymbol{x} \in \omega_2} \left\{ \boldsymbol{w}^T (\boldsymbol{x} - \boldsymbol{\mu}_2)(\boldsymbol{x} - \boldsymbol{\mu}_2)^T \boldsymbol{w} \right\} \\ = \boldsymbol{w}^T \boldsymbol{\Sigma}_1 \boldsymbol{w} + \boldsymbol{w}^T \boldsymbol{\Sigma}_2 \boldsymbol{w} \\ \propto \boldsymbol{w}^T \mathbf{S}_w \boldsymbol{w}. \end{aligned} \tag{3.7}$$

In the input domain, the between-class scattering matrix can be written as [Theodoridis, 2009]

$$\mathbf{S}_b = \sum_{i=1}^{2} P_i (\boldsymbol{\mu}_i - \boldsymbol{\mu}_g)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_g)^T, \tag{3.8}$$

where $\boldsymbol{\mu}_i$ is the mean of class i and $\boldsymbol{\mu}_g$ is the global mean. The between-class variance in the output space can be written as [Theodoridis, 2009]

$$(\mu_1 - \mu_2)^2, \tag{3.9}$$

which in terms of input data is equivalent to

$$(\boldsymbol{w}^T\boldsymbol{\mu}_1 - \boldsymbol{w}^T\boldsymbol{\mu}_2)(\boldsymbol{w}^T\boldsymbol{\mu}_1 - \boldsymbol{w}^T\boldsymbol{\mu}_2)^T = \boldsymbol{w}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{w}. \tag{3.10}$$

Since the global mean is a mean of the two class means, this can be written as (substituting the global mean and doing some algebra)

$$\boldsymbol{w}^T\left\{\frac{1}{P_1^2P_2 + P_1P_2^2}\Big(P_1(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_g)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_g)^T + P_2(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_g)(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_g)^T\Big)\right\}\boldsymbol{w} \propto \boldsymbol{w}^T\mathbf{S}_b\boldsymbol{w}. \tag{3.11}$$

This means that the FDR in terms of input data can be expressed as (the generalized Rayleigh quotient)

$$FDR = \frac{\boldsymbol{w}^T\mathbf{S}_b\boldsymbol{w}}{\boldsymbol{w}^T\mathbf{S}_w\boldsymbol{w}}, \tag{3.12}$$

which is the object of maximization.

Since the *direction* of $\boldsymbol{w}$ is the only thing of importance, it is valid to constrain the denominator to equal to one and solve the constrained optimization problem given by [Hubert et al., 2005]

$$\begin{aligned} \max_{\boldsymbol{w}} \quad & \boldsymbol{w}^T\mathbf{S}_b\boldsymbol{w} \\ \text{subject to} \quad & \boldsymbol{w}^T\mathbf{S}_w\boldsymbol{w} = 1 \end{aligned} \tag{3.13}$$

The Karush–Kuhn–Tucker (KKT) conditions necessary at the optimal solution are as follows [Theodoridis, 2009]:

i) $\frac{\partial L(\boldsymbol{w},\boldsymbol{\lambda})}{\partial \boldsymbol{w}} = 0$

ii) $\lambda_i = 0, \quad \forall i$

iii) $\lambda_i\left(\boldsymbol{w}^T\mathbf{S}_w\boldsymbol{w} - 1\right) = 0,$

where λ_i denotes a Lagrangian multiplier and

$$L(\boldsymbol{w}, \lambda) = \boldsymbol{w}^T\mathbf{S}_b\boldsymbol{w} - \lambda\left(\boldsymbol{w}^T\mathbf{S}_w\boldsymbol{w} - 1\right). \tag{3.14}$$

The solution of the constrained optimization problem is thus solved by taking the derivative of $L(\boldsymbol{w}, \lambda)$ with respect to $\boldsymbol{w}$ and setting it equal to zero. This results in the eigenvalue-eigenvector problem given by

$$\mathbf{S}_b\boldsymbol{w} = \lambda\mathbf{S}_w\boldsymbol{w} \leftrightarrow \lambda\boldsymbol{w} = \mathbf{S}_w^{-1}\mathbf{S}_b\boldsymbol{w}, \tag{3.15}$$

which by substituting into (3.12) gives

$$FDR = \lambda. \tag{3.16}$$

Thus, the cost function (FDR) is maximized by choosing $\boldsymbol{w}$ to be the eigenvector corresponding to the *largest* eigenvalue of $\mathbf{S}_w^{-1}\mathbf{S}_b$. However, for the simple case of two classes, the expression of this solution can be found without any eigendecomposition [Theodoridis, 2009]. Since $\mathbf{S}_b \propto (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T$, (3.15) can be written as [Theodoridis, 2009]

$$\lambda \mathbf{S}_w \boldsymbol{w} \propto (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{w} = \alpha(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \tag{3.17}$$

where α is a constant. The direction of the projection vector is thus given by

$$\hat{\boldsymbol{w}} = \mathbf{S}_w^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2). \tag{3.18}$$

The projection is illustrated in Figure 3.1, and a more detailed interpretation in a target detection setting is discussed in Section 4.3.2.

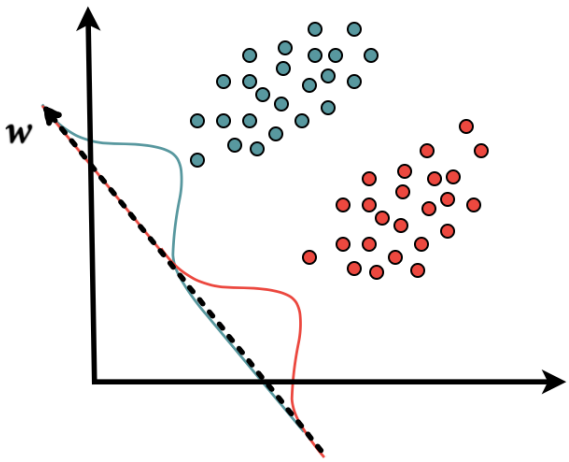

Figure 3.1: The green and red markers correspond to two different classes. In FLDA, the data gets projected onto the direction, $\boldsymbol{w}$, of maximum class separation.

Remarks

- Since the solution for $\boldsymbol{w}$ requires inversion of $\mathbf{S}_w$, it must necessarily be invertible, i.e. non-singular. This is a problem when the dimensionality is greater than the number of samples. One solution is to reduce the dimension by principal component analysis (PCA), such that matrix inversion becomes numerically stable, and then do FLDA (PCA + FLDA) [Theodoridis, 2009].

- FLDA does not assume Gaussian distributed classes, but if the projected data are unimodal and bell-shaped, the difference between the projected class means is a good measure of class separability [Manolakis and Shaw, 2002b].

- Since the pooled covariance of the background and the target is used, the FLDA implicitly assumes that the covariance matrices are equal [Johnson and Wichern, 2007].

3.2.2 Sparse Linear Discriminant Analysis

Sparse Linear Discriminant Analysis (SLDA) comprises a variety of proposed algorithms to obtain a sparse solution[7] to the linear discriminant analysis (LDA) problem. There are two main reasons for the various algorithms: 1) The LDA classification rule can be considered to originate from three different starting points, including Bayes' rule, Fisher's discriminant problem and the optimal scoring problem. 2) The various types of regularization techniques lead to different algorithms [Hastie et al., 2015].

The SLDA method to be investigated in this thesis is based on the optimal scoring problem. Before proceeding any further, the essential principles and theory of optimal scoring and regularization are reviewed.

Optimal Scoring

The idea of optimal scoring is to transform the categorical class lables, $y_i = 1/0$, into quantitative ones (i.e. continuous variables), and in this way turn the LDA into a *regression problem*. For a data set $\{\mathbf{x}_i, y_i\}_{i=1}^N$, a *linear regression model* assumes that the dependent variables, y_i, are a linear function of the predictor variables, $\boldsymbol{x}_i = [x_{i1}, ..., x_{ip}]^T$, and take the form

$$y_i = w_0 + \sum_{j=1}^{p} x_{ij} w_j, \tag{3.19}$$

where $w_0, ..., w_p$ are the parameter coefficients [Hastie, 2009, Hastie et al., 2015]. In matrix form, this is equivalent to

$$\boldsymbol{y} = w_0 \mathbf{1} + \boldsymbol{X}\boldsymbol{w}, \tag{3.20}$$

where $\boldsymbol{y} = [y_1, ..., y_N]^T$, $\boldsymbol{w} = [w_1, ..., w_p]^T$, $\boldsymbol{X}$ is the matrix whose rows are the predictor variables and $w_0\mathbf{1}$ is the $N \times 1$ vector whose elements are all

7. There are various definitions of sparse matrices/vectors, but common to all is that they describe matrices/vectors where several elements are zero [Pissanetzky, 1984].

equal to w_0. The regression problem is then all about estimating the parameter coefficients by optimizing some cost function. However, in LDA, it is only the *direction* of $\boldsymbol{w}$ that is of importance and the threshold w_0 is omitted from now.

To turn a K class LDA problem with N samples into a regression problem, an $N \times K$ indicator matrix, $\boldsymbol{Y}$ with elements

$$Y_{ij} = \begin{cases} 1, & \text{if sample } i \text{ belongs to class } j \\ 0, & \text{else,} \end{cases} \tag{3.21}$$

and a $K \times 1$ score vector $\boldsymbol{\theta}$, containing the *scores*[8] of the classes are formed. In this way, the $N \times 1$ vector $\boldsymbol{Y\theta}$ is the vector containing the scored/scaled training data that is to be regressed onto the predictor matrix, $\boldsymbol{X}$ [Hastie et al., 1995].

The optimal scoring algorithm then generates a sequence of $(\boldsymbol{w}_k, \boldsymbol{\theta}_k)$-pairs[9] by solving the optimal scoring criterion defined as [Clemmensen et al., 2011, Atkins et al., 2017]

$$\begin{aligned} \min_{\boldsymbol{w}_k, \boldsymbol{\theta}_k} \quad & \left\{ |\boldsymbol{Y\theta}_k - \boldsymbol{X w}_k|_2^2 \right\} \\ \text{subject to} \quad & \frac{1}{N} \boldsymbol{\theta}_k^T \boldsymbol{Y}^T \boldsymbol{Y \theta}_k = 1 \\ & \boldsymbol{\theta}_k^T \boldsymbol{Y}^T \boldsymbol{Y \theta}_l = 0 \quad \forall \quad l < k, \end{aligned} \tag{3.22}$$

where $|\cdot|_2$ denotes the L^2-vector norm, $\boldsymbol{w}_k$ is a $p \times 1$ vector containing the regression coefficients, and the constraints assure normalized and mutually orthogonal scores [Hastie et al., 1994].

In this way, the data points within one class will be projected onto $\boldsymbol{w}_k$ at a position dependent on the score of the class. The solution to the optimization problem will thus separate the classes in one dimension [Kay, 1999].

It can be shown [Hastie et al., 1994] that the optimal solution $\boldsymbol{w}_k$ is proportional to the solution of (3.13) (the optimal solution of Fisher's discriminant problem).

8. Each class is assigned a score, i.e. a real number, in such a way that they (the scores) are optimally predicted in the linear regression on $\boldsymbol{X}$ [Hastie et al., 1994].
9. There are at most $K - 1$ non-trivial solutions to (3.22) [Clemmensen et al., 2011]. In the $(K = 2)$-class detection problem, this gives only *one* solution/direction.

Regularization

Optimizing the least squares cost function, as in the optimal scoring criterion in 3.22, is one of the most common ways to estimate the parameter variables in a regression problem. Ordinary LS estimates $\boldsymbol{w}$ as [Theodoridis, 2009]

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} |\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}|_2^2 = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}, \tag{3.23}$$

which gives the unbiased estimate with the lowest variance of all unbiased methods [Kuhn and Johnson, 2013]. That is, the LS method produces the minimum-variance unbiased estimator (MVUE). However, the expected squared prediction error of a new sample, $\boldsymbol{x}_0$, can be decomposed into variance and bias squared[10]:

$$E\left\{\left(y - \boldsymbol{x}_0^T\hat{\boldsymbol{w}}\right)^2\right\} = \text{Var}\{\boldsymbol{x}_0^T\hat{\boldsymbol{w}}\} + \text{Bias}\{\boldsymbol{x}_0^T\hat{\boldsymbol{w}}\}^2. \tag{3.24}$$

This is known as the *bias-variance trade-off* [Kuhn and Johnson, 2013, Hastie, 2009]. In some cases, a small increase in bias may lead to a significant decrease in the variance, in that way reducing the squared error. One such scenario is when (some of) the predictor variables are highly linearly dependent [Kuhn and Johnson, 2013]. This is because the variance of the parameter coefficients is large when the predictor variables are highly linearly dependent ($\text{Var}\{\hat{\boldsymbol{w}}\} \propto (\boldsymbol{X}^T\boldsymbol{X})^{-1}$) [Seber and Lee, 2012]. However, by adding a *penalty* to the LS problem, the regularized optimal solution becomes $\hat{\boldsymbol{w}}_{\ell_2} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ [Bishop, 2006][11], which has a lower variance. Specifically, the order of magnitude of $(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}$ will be much smaller than that of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ if $\boldsymbol{X}$ is ill-posed.

Adding penalties to the regression parameters can in this way help reducing the variance of the model and hence also reduce the possibility of overfitting the model to the training data. This yields better regularization and prediction power [Kuhn and Johnson, 2013].

The different types of regularizers have different properties. Common to all is that they shrink the magnitude of the coefficients. For a fixed, real number $q \geq 1$, a general expression for penalized LS is given by [Bishop, 2006]

$$\hat{\boldsymbol{w}}_{\ell_q} = \arg\min_{\boldsymbol{w}} |\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}|_2^2 + \lambda|\boldsymbol{w}|_q^q, \tag{3.25}$$

where $|\cdot|_q$ denotes the L^q-vector norm and λ is the regularization weight, controlling how strict the penalty on non-zero coefficients should be.

10. Sometimes an irreducible error is also included as the contribution from the noise variance.
11. Here, a quadratic regularizer $\lambda||\boldsymbol{w}||_2^2$ is added to the LS objective function.

For $q = 1$ (3.25) yields *least absolute shrinkage selection operator* (LASSO) regression (where the penalty is referred to as the ℓ_1-penalty), and for $q = 2$, *ridge* regression (where the penalty is referred to as the ℓ_2-penalty) [Hastie et al., 2015]. *Elastic net* regression solves [Zou and Hastie, 2005a]

$$\min_{\boldsymbol{w}} \ |\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}|_2^2 + \lambda|\boldsymbol{w}|_1 + \gamma|\boldsymbol{w}|_2^2, \tag{3.26}$$

which is a combination of LASSO and ridge regression.

In a two-dimensional case, if the variables X_1 and X_2 are highly linearly dependent, and/or λ is sufficiently large, LASSO regression will suppress the coefficient of X_1 or X_2 ($\hat{w}_1 = 0$ or $\hat{w}_2 = 0$), in this way giving a sparse model. Ridge regression, on the other hand, tends to shrink both of them to the same value [Clemmensen et al., 2011]. The reason for this is that, in the case of correlated variables, numerous linear combinations of X_1 and X_2 can explain the Y-value ($0.1X_1 + 0.9X_2$, $0.2X_1 + 0.8X_2$ etc.). The ridge penalty will simply choose the combination where $\hat{w}_1^2 + \hat{w}_2^2$ is minimized, and that is the case when $\hat{w}_1$ and $\hat{w}_2$ are similar. The elastic net regression is a compromise between ridge and LASSO regression.

An alternative way to write the penalized LS problem, that allows for a geometrical interpretation (Figure 3.2), is [Hastie, 2009]

$$\min_{\boldsymbol{w}} \ |\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}|_2^2 \tag{3.27}$$

$$\text{subject to} \ \ |\boldsymbol{w}|_q^q \leq t, \tag{3.28}$$

where t is a one-to-one mapping of λ.

Sparse Linear Discriminant Analysis Algorithm

[Witten and Tibshirani, 2011] proposes an SLDA algorithm by adding an ℓ_1-penalty to Fisher's discriminant problem as reviewed in the previous section. However, solving this problem is challenging. The SLDA as proposed in [Clemmensen et al., 2011] is more easily solved and will thus be the SLDA algorithm considered in this thesis.

[Clemmensen et al., 2011] proposes a modified version of the optimal scoring approach to LDA, where an elastic net penalty is added to the coefficient vector to (possibly[12]) favor sparse solutions [Hastie et al., 2015]. The optimization

12. If the regularization weight on the ℓ_1-penalty is sufficiently large [Hastie et al., 2015].

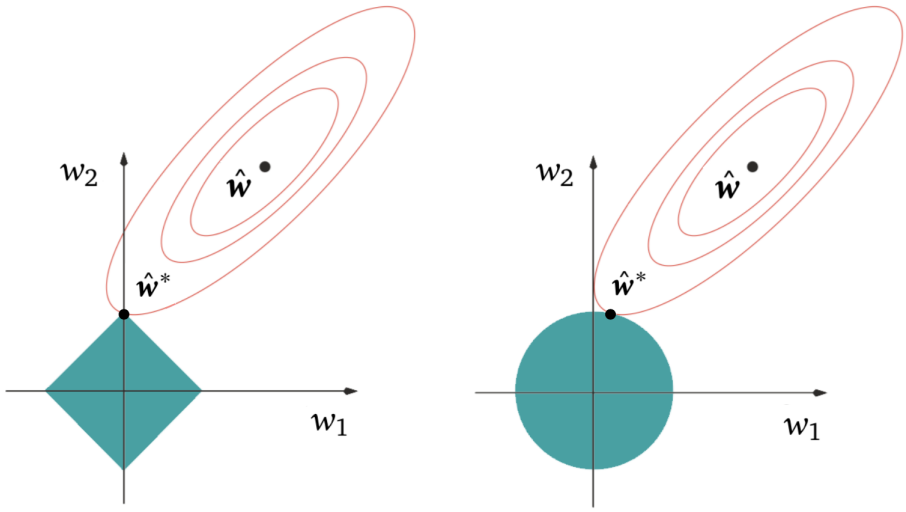

Figure 3.2: The red elliptic contours represent the solution space of the objective function in (3.28). Left: The green diamond represents the regularization function of $q = 1$ (LASSO): $|w_1| + |w_2| \leq t$. Right: The green circle represents the regularization function of $q = 2$ (ridge): $w_1^2 + w_2^2 \leq t$. The LASSO tends to suppress some parameters to give a sparse solution, while the ridge tends to shrink to small values instead. Modified from [Hastie, 2009].

problem is given by [Clemmensen et al., 2011]

$$\begin{aligned} \min_{\boldsymbol{w}_k, \boldsymbol{\theta}_k} \quad & \left\{ |\boldsymbol{Y}\boldsymbol{\theta}_k - \boldsymbol{X}\boldsymbol{w}_k|_2^2 + \gamma \boldsymbol{w}_k^T \boldsymbol{\Omega} \boldsymbol{w}_k + \lambda |\boldsymbol{w}_k|_1 \right\} \\ \text{subject to} \quad & \frac{1}{N} \boldsymbol{\theta}_k^T \boldsymbol{Y}^T \boldsymbol{Y} \boldsymbol{\theta}_k = 1 \\ & \boldsymbol{\theta}_k^T \boldsymbol{Y}^T \boldsymbol{Y} \boldsymbol{\theta}_l = 0 \quad \forall \ l < k, \end{aligned} \tag{3.29}$$

where $\boldsymbol{X}$ is centered and $\boldsymbol{\Omega}$ is a positive semi-definite matrix. When $\boldsymbol{\Omega} = \boldsymbol{I}$, the second term in the objective function becomes the ℓ_2-penalty, in total resulting in the elastic net penalty.

The optimization in (3.29) is a non-convex problem, and the solution is approximated iteratively by alternately holding $\boldsymbol{w}_k$ and $\boldsymbol{\theta}_k$ constant [Clemmensen et al., 2011]. That is, for fixed $\boldsymbol{\theta}_k$, the optimization problem is given by

$$\min_{\boldsymbol{w}_k} \quad \left\{ |\boldsymbol{Y}\boldsymbol{\theta}_k - \boldsymbol{X}\boldsymbol{w}_k|_2^2 + \gamma \boldsymbol{w}_k^T \boldsymbol{\Omega} \boldsymbol{w}_k + \lambda |\boldsymbol{w}_k|_1 \right\}, \tag{3.30}$$

which is (for $\boldsymbol{\Omega} = \boldsymbol{I}$) an elastic net optimization that is solved by the LARS-EN [Zou and Hastie, 2005b] algorithm. Then, for fixed $\boldsymbol{w}_k$, the optimization

problem is defined by

$$\begin{aligned} \min_{\boldsymbol{\theta}_k} \quad & \left\{ |\boldsymbol{Y}\boldsymbol{\theta}_k - \boldsymbol{X}\boldsymbol{w}_k|_2^2 \right\} \\ \text{subject to} \quad & \frac{1}{N}\boldsymbol{\theta}_k^T \boldsymbol{Y}^T \boldsymbol{Y}\boldsymbol{\theta}_k = 1 \\ & \boldsymbol{\theta}_k^T \boldsymbol{Y}^T \boldsymbol{Y}\boldsymbol{\theta}_l = 0 \quad \forall \ l < k, \end{aligned} \tag{3.31}$$

which results in

$$\boldsymbol{\theta}_k = s(\boldsymbol{I} - \boldsymbol{Q}_k \boldsymbol{Q}_k^T \boldsymbol{Y}^T \boldsymbol{Y})(\boldsymbol{Y}^T \boldsymbol{Y})^{-1} \boldsymbol{Y}^T \boldsymbol{X}\boldsymbol{w}_k, \tag{3.32}$$

where s is a constant to assure that $\frac{1}{N}\boldsymbol{\theta}_k^T \boldsymbol{Y}^T \boldsymbol{Y}\boldsymbol{\theta}_k = 1$ and $\boldsymbol{Q}_k$ is the matrix whose columns are the previous solutions to $\boldsymbol{\theta}_k$ along with the trivial $\boldsymbol{\theta}_k = \mathbf{1}$ solution [Clemmensen et al., 2011].

These steps are repeated for each $(\boldsymbol{w}_k, \boldsymbol{\theta}_k)$-pair (each direction) until convergence or until the maximum number of iterations is reached [Clemmensen et al., 2011].

Remarks

- The SLDA algorithm is designed for high-dimensional data, but is included in this thesis because of its ability to improve the generalization properties (in the case of correlated variables, for instance).

4

Spectral Unmixing

The thesis' second main approach to the detection problem is through the inverse problem of spectral unmixing. This chapter begins with introducing the application of spectral unmixing and continues with discussing different mathematical mixture models. Finally, the different algorithms to be tested in this thesis are presented.

4.1 Application

When an image is acquired, the ground scene gets subdivided into discrete pixels whose sizes depend on the spatial resolution of the sensor. Since the sensor integrates the brightness within the entire ground pixel, all information about spatial structure on subpixel level is lost, and the resulting spectrum is a combination of the spectra related to the materials in the pixel. If the material within the ground pixel is non-uniform, the pixel is referred to as *mixed* and the response is a *composite signature* that does not resemble any of the pure signals of the materials in the scene [Campbell, 2011, Manolakis, 2016]. In such situations, to be able to detect a specific material, spectral unmixing is commonly applied to the pixels to decompose the composite signatures into a set of constituent spectra and their associated contribution fractions to the total signal [Keshava and Mustard, 2002].

A further discussion on spectral unmixing requires some assumptions about

the mixing process. In the next section, two different mixing models and their assumptions are presented.

4.2 Mixture Models

An essential presumption of mixture modelling is that the surface is comprised of a small number of disparate materials, called *endmembers*, with relatively constant spectral signatures [Keshava and Mustard, 2002]. With this as a basis, and the fact that mixed pixels occur for different reasons, two classes of mixing models are defined: linear and nonlinear mixing models.

A linear mixing model (LMM) assumes that the mixing scale is macroscopic such that each photon interacts with *one* surface material only. In this way, the actual mixing happens *inside* the sensor due to insufficient spatial resolution [Bioucas-Dias et al., 2013]. The resulting pixel spectra thus becomes a linear combination of the endmember spectra, where the weights correspond to the relative pixel area that each endmember occupies (Figure 4.1.a).

In other scenarios, where the mixing scale is microscopic[1] and/or there exist multiple scattering such that each photon may interact with multiple materials, the mixing is nonlinear in nature and follows a nonlinear mixing model (Figure 4.1.b-c).

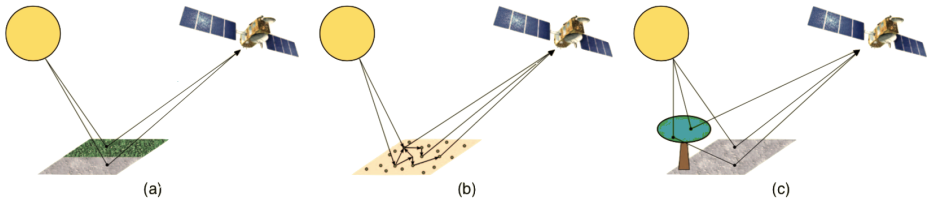

Figure 4.1: (a) Linear mixing due to insufficient spatial resolution. **(b)** Nonlinear mixing due to microscopic mixing scale. **(c)** Nonlinear mixing due to multiple scattering. From [Dobigeon et al., 2014].

Although the assumptions of the LMM is not always satisfied, it is widely acknowledged as an acceptable model for many real world applications [Ma et al., 2014]. In the problem of this thesis it is assumed that the macroscopic mixing is the most prominent effect, and the LMM will for this reason suffice as mixture

1. Endmembers are microscopically mixed if the mixing scale is smaller than the photon path length (traveling length) in the mixture [Boas et al., 2016]. In this way, photons emitted by molecules in one material may get absorbed and re-emitted by molecules in another material [Bioucas-Dias et al., 2012].

model. The following subsection presents a mathematical representations of the LMM.

4.2.1 The Linear Mixing Model

There are two main approaches to linear mixture modelling: statistical and geometrical. The statistical approach models the endmembers using probability distributions while the geometrical approach utilizes the fact that linear combinations of vectors lie in a simplex set[2] [Manolakis et al., 2014, Bioucas-Dias et al., 2012]. In this thesis, the geometrical approach is considered.

In a scene with L endmembers, the p-dimensional signal from each ground pixel can be expressed as

$$\boldsymbol{x} = \boldsymbol{M}\boldsymbol{\alpha} + \boldsymbol{n}, \tag{4.1}$$

where $\boldsymbol{M} = [\boldsymbol{m}_1, ..., \boldsymbol{m}_L]$ is the $p \times L$ matrix whose columns are the spectal signatures[3] of the L endmembers, $\boldsymbol{\alpha} = [\alpha_1, ..., \alpha_L]^T$ is the L-dimensional fractional abundance vector[4] and $\boldsymbol{n}$ represents additive noise [Manolakis et al., 2014]. For N pixels, this can be written as

$$\boldsymbol{X} = \boldsymbol{M}\boldsymbol{A} + \boldsymbol{N}, \tag{4.2}$$

where $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_N]$, $\boldsymbol{A} = [\boldsymbol{\alpha}_1, ..., \boldsymbol{\alpha}_N]$ and $\boldsymbol{N} = [\boldsymbol{n}_1, ..., \boldsymbol{n}_N]$. Since $\boldsymbol{\alpha}$ represents fractions of the pixel area, its components are subject to two constraints to assure that they make physical sense. The first one is the *abundance non-negativity constraint*, $\alpha_i \geq 0 \; \forall i$, and the second one is the *abundance sum constraint*, $\sum_{i=1}^{L} \alpha_i = 1$ [Bioucas-Dias et al., 2013].

Under the LMM, all pixels are thus by definition convex combinations of the endmembers. The *convex hull*[5] of the L columns of $\boldsymbol{M}$, conv($\boldsymbol{M}$), is thus a $(L-1)$-simplex, assuming the columns of $\boldsymbol{M}$ are *affinely independent*[6] [Preparata and Shamos, 2012, Bioucas-Dias et al., 2012]. This means that all pixel vectors lie within the $(L-1)$-simplex whose vertices are the endmembers. For $L = 2$ endmembers, this yields a *line* between the two endmembers, on which all mixtures will fall. $L = 3$ endmembers yields a *triangle* (Figure 4.2), and $L = N$ endmembers an $N-1$-simplex in $N-1$-dimensional space [Boardman and Kruse, 2011, Weisstein, n.d.-b].

2. A simplex is the generalization of a tetrahedral space in n dimensions [Weisstein, n.d.-b].
3. Spectral signatures sampled according to the spectral resolution of the sensor.
4. The fractional abundance vector indicates the fractional abundaces of the L endmembers in the pixel.
5. The smallest subset of points in $\mathbb{R}^p$ containing the L columns of $\boldsymbol{M}$ [Preparata and Shamos, 2012].
6. If a linear combination of an n-family $(\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$, $\lambda_1\boldsymbol{x}_1 + ... + \lambda_n\boldsymbol{x}_1$ with $\sum_{i=1}^{n} \lambda_i = 0$ is zero iff $\lambda_1 = ... = \lambda_n = 0$, the family is said to be *affinely independent* [Brondsted, 2012].

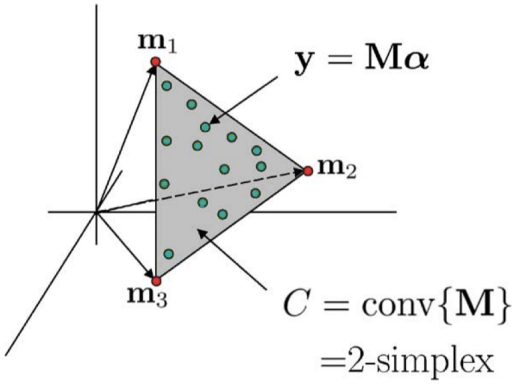

Figure 4.2: C is the convex hull of the columns of $\boldsymbol{M}$. All mixtures of the three endmembers fall within the triangle. Modified from [Bioucas-Dias et al., 2012].

Remarks

- The spectral signature of a material may vary for several reasons; variations in illumination (due to topology, seasonal/diurnal variations in solar illumination, etc.), atmospheric conditions (level of gasses and aerosols) and inherent variations in the material. However, since the LMM represents each endmember by a single vector, it does not account for any spectral variability, which may lead to inaccuracies in the abundance estimation [Zare and Ho, 2014].

4.3 Linear Unmixing Algorithms

Linear unmixing may be divided into two subgroups: full unmixing and partial unmixing. In full unmixing algorithms, the spectra of *all* endmembers in the scene are estimated and their associated abundances are extracted in each pixel. However, in a target detection problem, it is only the abundance of the target material that is of any interest, and much of the trouble in the full unmixing algorithms is avoidable. Partial unmixing algorithms are supervised[7] algorithms that searches only the abundance of the *desired* endmember(s), and are thus convenient in detection problems.

Most unmixing algorithms are designed for *hyperspectral* images [Karoui et al.,

7. Supervised in the way that the user must provide the spectrum of the desired endmember(s).

2012]. As opposed to multispectral remote sensing, which samples in a relative small (but wisely chosen) number of *broad* channels, the hyperspectral sensors may sample in over 200 *narrow* spectral bands [Campbell, 2011]. This gives an almost continuous spectral response of each pixel, which significantly helps to discriminate between surface materials. However, considering an operational system, it is desirable to use free multispectral data instead of costly hyperspectral data. While the literature is rich with research on the unmixing of hyperspectral images, there are few methods designed specifically for multispectral unmixing. This thesis will thus also explore the application of some hyperspectral unmixing algorithms on multispectral data.

The reminder of this chapter presents three different algorithms for linear unmixing, including one unsupervised full unmixing algorithm and two supervised partial unmixing algorithms.

4.3.1 2D-Corr-NLS

The 2D-Corr-NLS algorithm, proposed by [Karoui et al., 2012], is an *unsupervised* algorithm designed for full spectral unmixing of multispectral images. Since this is an unsupervised algorithm, the problem consists of recovering the unobserved sources mixed in unknown proportions in the pixel signals (both $\boldsymbol{M}$ and $\boldsymbol{A}$ in (4.2) are unknown). This is indeed a *blind source separation* (BSS) problem, and the 2D-Corr-NLS algorithm combines the first stage of a correlation-based BSS algorithm with clustering to detect *single-source-zones*[8] (SSZs) and identify the endmember spectra ($\boldsymbol{M}$). Then, the *nonnegative least squares* (NLS) method is used to unmix the signals and obtain the abundances ($\boldsymbol{A}$) [Karoui et al., 2012].

The algorithm is roughly divided into *four* stages: *i)* Detection, *ii)* Estimation, *iii)* Clustering and *iv)* Extraction. Before explaining these stages in more detail, a quick look at the BSS model is needed.

The standard BSS model consists of two elements, namely the mixing matrix, $\boldsymbol{A}_{BSS}$ and the source signals, $\boldsymbol{M}_{BSS}$, explaining the observations through [Karoui et al., 2012]

$$\boldsymbol{X}_{BSS} = \boldsymbol{A}_{BSS}\boldsymbol{M}_{BSS}, \tag{4.3}$$

where $\boldsymbol{X}_{BSS} = \boldsymbol{X}^T$, $\boldsymbol{A}_{BSS} = \boldsymbol{A}^T$ and $\boldsymbol{M}_{BSS} = \boldsymbol{M}^T$.[9] However, in this way of writing, BSS algorithms are *not* applicable to multispectral images. This is because the algorithms require a high number of "samples" which translates into spectral bands. The proposed solution in [Karoui et al., 2012] is to do

8. Regions in the image corresponding to an endmember.
9. $\boldsymbol{X}, \boldsymbol{A}$ and $\boldsymbol{M}$ as defined in (4.2).

an alternative interpretation[10] of the sources and mixing matrix in the BSS problem, which essentially results in transposing (4.3) to obtain

$$\boldsymbol{X}_{BSS}^T = (\boldsymbol{A}_{BSS}\boldsymbol{M}_{BSS})^T = \boldsymbol{M}_{BSS}^T\boldsymbol{A}_{BSS}^T, \tag{4.4}$$

which is equal to

$$\boldsymbol{X} = \boldsymbol{M}\boldsymbol{A}.^{11} \tag{4.5}$$

i) Detection

The first stage in the 2D-Corr-NLS algorithm is to automatically detect the SSZs by applying the first part of a spatial correlation based *sparse component analysis* (SCA) algorithm (as proposed in [Deville and Puigt, 2007]) [Karoui et al., 2012]. This includes moving a two-dimensional *analysis window* over the image and, based on the neighborhood of pixels inside the window, calculating the cross-correlation coefficients between *all* bands[12]. Mathematically, for a given neighborhood Ω, the cross-correlation coefficient between the two bands p and q is expressed as

$$\rho_{\boldsymbol{x}_p,\boldsymbol{x}_q}(\Omega) = \frac{\left\langle \boldsymbol{x}_p(\Omega), \boldsymbol{x}_q(\Omega) \right\rangle}{|\boldsymbol{x}_p(\Omega)|_2 \; |\boldsymbol{x}_q(\Omega)|_2}, \tag{4.6}$$

where $\boldsymbol{x}_i(\Omega)$ is the vector of all neighborhood pixels in band i, $\langle \cdot, \cdot \rangle$ denotes the inner product between two vectors and $|\cdot|_2$ denotes the L^2-vector norm.

Each analysis zone will result in a $\frac{p(p-1)}{2}$-dimensional vector of cross-correlation coefficients, where *high* (close to one) correlation coefficients indicate *uniform* neighborhoods, meaning neighborhoods with pixels of uniform signature. This can occur when the pixels contain the same pure material *or* when the pixels contain a mixture of equal materials in similar proportions[13] (Figure 4.3). Then, the neighborhood vectors between the bands will be highly linearly dependent (yielding a high correlation coefficient).

For each neighborhood cross-correlation vector, the *smallest* element is then selected, and if that value exceeds a threshold t, the neighborhood is said to be a SSZ [Karoui et al., 2012].

10. The traditional approach is to view the endmember spectra as the sources (spectral sources) and each pixel as observations. Alternatively, one may consider the abundance fractions associated with one material as the sources (spatial sources) and the spectral bands as observations [Karoui et al., 2012].
11. With respect to BSS terminology, this yields *mixing matrix* $\times$ *source matrix*.
12. For p spectral bands, this yields $\frac{p(p-1)}{2}$ cross-correlation coefficients.
13. The proportion of material n is similar in all pixels in the neighborhood.

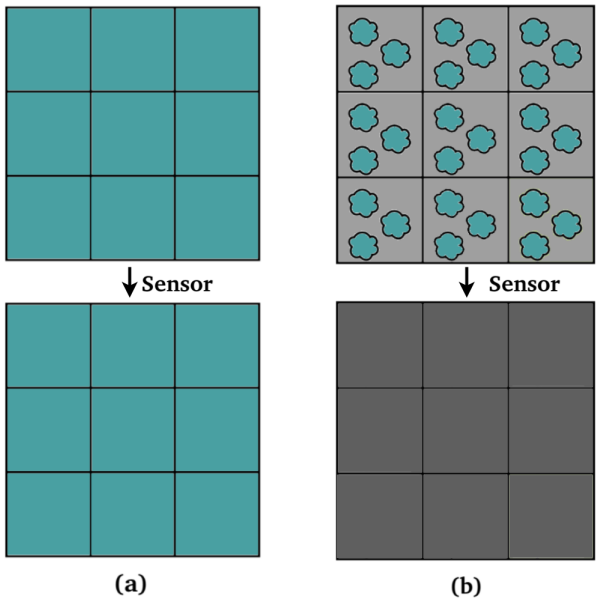

Figure 4.3: Two neighborhoods of 3×3 pixels detected as SSZs. **(a)** All pixels contain the same pure material. **(b)** All pixels contain a mixture of the same materials in equal proportions.

ii) Estimation

In the second stage, *tentative* estimates of the columns of the mixing matrix[14], $\boldsymbol{M}$, are computed as the *medians* of the pixels in each band in the detected SSZs:

$$\hat{\boldsymbol{m}}_j = \begin{bmatrix} \text{median}(\boldsymbol{x}_1(\Omega_j)) \\ \vdots \\ \text{median}(\boldsymbol{x}_p(\Omega_j)) \end{bmatrix}. \tag{4.7}$$

iii) Clustering

In the third stage, L columns are selected as the final estimate of $\boldsymbol{M}$. Since some of the tentative estimates obtained in the previous section may belong to the same endmember (there might have been detected multiple SSZs representing the same endmember), a clustering algorithm is used to cluster all the tentative estimates into $L = k^*$ clusters[15]. More specifically, the *fuzzy c-means* algorithm [Bezdek et al., 1984] is applied. In [Karoui et al., 2012], they set the center of

14. In terms of the BSS model.
15. The choice of number of clusters is important for the result. In [Karoui et al., 2012], the optimal number of clusters, k^*, is chosen by considering the *cluster validity index*, as described in [Kim et al., 2001], for different values of k. In this thesis, however, the long established Calinski-Harabasz criterion, as proposed in [Caliński and Harabasz, 1974], is applied.

each cluster to the element with the *highest confidence*, given by the smallest element in the neighborhood cross-correlation vector. However, because of the coarse spatial resolution of the Sentinel-2 images, the endmember of highest interest will *not* be "infected trees" but rather "pixels containing infected trees" (Figure 4.3.b), and the element with the highest confidence may not be the best representative. In this thesis, the centers are thus chosen as the mean value of the cluster members. This may allow for some more variation in proportions and/or spectral variability between pixels, which may enhance the probability of detecting "pixels containing infected trees" as an endmember.

When the clustering algorithm has converged, the centers of the clusters will form the estimate of $\boldsymbol{M}$, $\hat{\boldsymbol{M}} = [\hat{\boldsymbol{m}}_1, ..., \hat{\boldsymbol{m}}_L]$.

iv) Extraction
In the fourth and final step, the goal is to extract the abundance maps (rows of $\boldsymbol{A}$). This is done one pixel (column of $\boldsymbol{X}$) at a time by the nonnegative least squares method, which is a constrained version of the ordinary least squares method. In this way, the columns of $\boldsymbol{A}$ are estimated one by one in the order according to $\hat{\boldsymbol{M}}$ and $\boldsymbol{X}$, hereby avoiding the permutation indeterminacy. Each operation is expressed as [Karoui et al., 2012]

$$\begin{aligned} \min_{\boldsymbol{\alpha}_k} \quad & |\hat{\boldsymbol{M}}\boldsymbol{\alpha}_k - \boldsymbol{x}_k|_2^2 \\ \text{subject to} \quad & \alpha_{ki} \geq 0 \quad i = 1, ..., L, \end{aligned} \tag{4.8}$$

where $\boldsymbol{x}_k$ is the kth column of $\boldsymbol{X}$. Solving this problem for $k = 1, ..., N$ gives the estimate of $\boldsymbol{A}$, $\hat{\boldsymbol{A}} = [\hat{\boldsymbol{\alpha}}_1, ..., \hat{\boldsymbol{\alpha}}_N]$.

The algorithm as presented above will solve the BSS problem subject to the abundance non-negativity constraint. The abundance sum constraint, however, has not been mentioned so far. To assure this constraint as well, some simple modifications of $\boldsymbol{X}$ and $\hat{\boldsymbol{M}}$ can be made before extracting $\hat{\boldsymbol{A}}$. The technique consists in adding a row of a strictly positive constant to both matrices [Karoui et al., 2012, Heinz et al., 2001].

$$\hat{\boldsymbol{M}} \rightarrow \begin{bmatrix} \delta\hat{\boldsymbol{M}} \\ \mathbf{1}^T \end{bmatrix}, \tag{4.9}$$

$$\boldsymbol{x} \rightarrow \begin{bmatrix} \delta\boldsymbol{x} \\ 1 \end{bmatrix}, \tag{4.10}$$

where δ is a constant controlling the effect of the abundance sum constraint. In this way, both constraints are handled.

Remarks

- Because the 2D-Corr-NLS algorithm uses spatial information to detect the single source zones, it requires an image (not a bunch of vectors) as input.

- Because of the way the 2D-Corr-NLS algorithm estimates endmembers and the coarse resolution of the Sentinel-2 images, one can at most expect to find "pixels containing infected trees" as an endmember. Spectral variability in infected trees and background, and/or variations in proportions between infected trees and background may lead to detection of multiple endmembers, all representing "pixels containing infected trees". For these reasons, the 2D-Corr-NLS algorithm may seem like a unlikely winner. However, since this is one of very few unmixing methods designed for multispectral imagery, it is included in this thesis. Also, since it is possible to force the method to detect any number (smaller or equal to the total number of detected single-source-zones in the scene) of endmembers by restricting the number of clusters to evaluate by the cluster index, it is possible to distinguish between spectrally similar endmembers.

4.3.2 Iterated Constrained Energy Minimimization

The *constrained energy minimization* (CEM) algorithm is a *partial* unmixing algorithm where only the desired endmember spectra are estimated (as opposed to full unmixing where the spectra of all endmembers present are estimated) [Nielsen, 2001]. The idea is to project the data onto a direction $\boldsymbol{w}$ where the total energy of the image is minimized while the response of the desired endmember(s) is constrained to a constant value [Resmini et al., 1997].

Partial unmixing, and thus CEM, is based on a modified version of the LMM where the $\boldsymbol{M\alpha}$ term is split into two. One part corresponds to the desired endmember and the other part represents the undesired enmembers [Nielsen, 2001]:

$$\begin{aligned} \boldsymbol{x} &= \boldsymbol{M\alpha} + \boldsymbol{n} \\ &= \boldsymbol{d}\alpha_p + \boldsymbol{U\gamma} + \boldsymbol{n}, \end{aligned} \tag{4.11}$$

where $\boldsymbol{d}$ is the $p \times 1$ vector representing the desired endmember spectrum with associated abundance α_p and $\boldsymbol{U}$ is the $p \times (L-1)$ matrix containing the endmember spectra of the unwanted endmembers with corresponding $(L-1) \times 1$ abundance vector $\boldsymbol{\gamma}$.

The CEM algorithm searches the direction $\boldsymbol{w}$ to project $\boldsymbol{x}$ onto such that [Nielsen, 2001]:

i) The projected value is one for the desired spectrum, that is $\boldsymbol{w}^T\boldsymbol{d} = 1$.

ii) The expected value of the projection is zero: $E\{\boldsymbol{w}^T\boldsymbol{x}\} = 0$.

iii) The expectation of the squared error, $E\{(\boldsymbol{w}^T\boldsymbol{x} - E\{\boldsymbol{w}^T\boldsymbol{x}\})^2\} = E\{(\boldsymbol{w}^T\boldsymbol{x})^2\}$ (which is the expectation of the *energy*), is minimized.

Since $E\{(\boldsymbol{w}^T\boldsymbol{x})^2\} = \text{Var}\{\boldsymbol{w}^T\boldsymbol{x}\} \overset{16}{=} \boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w}$, where $\boldsymbol{\Sigma}$ is the covariance matrix of the data, the requests above yield the constrained optimization problem given by [Nielsen, 2001]:

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & \boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w} \\ \text{subject to} \quad & \boldsymbol{w}^T\boldsymbol{d} = 1, \end{aligned} \tag{4.12}$$

with solution

$$\hat{\boldsymbol{w}} = \frac{\boldsymbol{\Sigma}^{-1}\boldsymbol{d}}{\boldsymbol{d}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{d}}. \tag{4.13}$$

By projecting the data onto this direction, each pixel vector is reduced to a single *score* (the projection value), and the result is a one-band output image (abundance image) with zero mean and minimum variance. Because of the scaling by $\frac{1}{\boldsymbol{d}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{d}}$, the score (abundance value) of a pure target pixel ($\boldsymbol{x} \approx \boldsymbol{d}$)[17] is close to one [Boardman and Kruse, 2011]. The projection vector is illustrated in Figure 4.4, and a more detailed geometric interpretation is given at the end of this section.

In [Nielsen, 2001], an alternative eigenvalue formulation of the CEM is proposed by considering the variance of the projection of $\boldsymbol{x}$ onto $\boldsymbol{w}$:

$$\begin{aligned} \text{Var}\{\boldsymbol{w}^T\boldsymbol{x}\} &= \text{Var}\{\boldsymbol{w}^T\boldsymbol{d}\alpha_p + \boldsymbol{w}^T\boldsymbol{U}\boldsymbol{\gamma} + \boldsymbol{w}^T\boldsymbol{n}\} \\ &\overset{18}{=} \text{Var}\{\boldsymbol{w}^T\boldsymbol{d}\alpha_p\} + \text{Var}\{\boldsymbol{w}^T\boldsymbol{U}\boldsymbol{\gamma}\} + \text{Var}\{\boldsymbol{w}^T\boldsymbol{n}\} \\ &\quad + 2\text{Cov}\{\boldsymbol{w}^T\boldsymbol{d}\alpha_p, \boldsymbol{w}^T\boldsymbol{U}\boldsymbol{\gamma}\} \\ &= \text{Var}\{\alpha_p\}\boldsymbol{w}^T\boldsymbol{d}\boldsymbol{d}^T\boldsymbol{w} + \boldsymbol{w}^T\boldsymbol{U}\text{Var}\{\boldsymbol{\gamma}\}\boldsymbol{U}^T\boldsymbol{w} \\ &\quad + \boldsymbol{w}^T\boldsymbol{\Sigma}_n\boldsymbol{w} + 2\boldsymbol{w}^T\boldsymbol{d}\text{Cov}\{\alpha_p, \boldsymbol{\gamma}\}\boldsymbol{U}^T\boldsymbol{w}, \end{aligned} \tag{4.14}$$

where $\boldsymbol{\Sigma}_n$ is the noise covariance matrix. By capturing all undesired effects in one term, $\boldsymbol{E}$, this can be written as the sum of the variance of the desired and

16. Assuming $E\{\boldsymbol{x}\} = \boldsymbol{0}$. That is, the data needs to be centered (mean corrected). In this way, the CEM method is equivalent to the *matched filtering* method [Manolakis et al., 2003].
17. Since the target class *will* inhabit some spectral variability (as discussed in Section 4.2.1), the desired signature $\boldsymbol{d}$ is typically represented by the mean target signature, and most target pixels will, to some extent (depending on the spectral variability of the target), deviate from this representative.
18. Assuming $\text{Cov}\{\boldsymbol{w}^T\boldsymbol{n}, \boldsymbol{w}^T\boldsymbol{d}\alpha_p\} = \text{Cov}\{\boldsymbol{w}^T\boldsymbol{n}, \boldsymbol{w}^T\boldsymbol{U}\boldsymbol{\gamma}\} = 0$, i.e. the noise is uncorrelated with any endmember spectra.

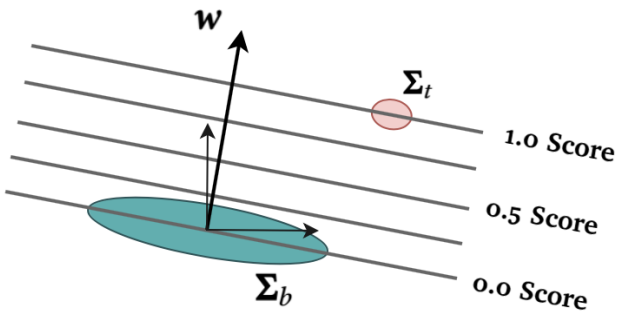

Figure 4.4: CEM (Assuming that the target is sparsely distributed in the scene, such that the covariance of the data is approximately equal to the covariance of the background): The projection onto $\boldsymbol{w}$ highlights the parts of the target signal that does *not* resemble the background and suppress the parts that do. Pixels dominated by background will get *scores* around zero and target pixels get scores around one. A pixel where half the area is covered with target material will get a score of 0.5. Modified from [Boardman and Kruse, 2011].

undesired effects (in the projected space):

$$\text{Var}\{\boldsymbol{w}^T\boldsymbol{x}\} = \text{Var}\{\alpha_p\}\boldsymbol{w}^T\boldsymbol{d}\boldsymbol{d}^T\boldsymbol{w} + \boldsymbol{w}^T\boldsymbol{E}\boldsymbol{w}, \tag{4.15}$$

which, by dividing by $\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w}$, gives the constant sum

$$1 = \text{Var}\{\alpha_p\}\frac{\boldsymbol{w}^T\boldsymbol{d}\boldsymbol{d}^T\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w}} + \frac{\boldsymbol{w}^T\boldsymbol{E}\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w}}. \tag{4.16}$$

From this way of writing, it is clear that it is possible to minimize the second term by maximizing the first term. Minimizing the variance of the undesired effects is thus equivalent to

$$\max_{\boldsymbol{w}} \quad \frac{\boldsymbol{w}^T\boldsymbol{d}\boldsymbol{d}^T\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w}}, \tag{4.17}$$

which is the same problem as in (3.12) (the FDR in Fisher LDA) and translates (in the same way) into the eigenvalue-eigenvector problem given by

$$\boldsymbol{d}\boldsymbol{d}^T\boldsymbol{w} = \lambda\boldsymbol{\Sigma}\boldsymbol{w}. \tag{4.18}$$

Requiring $\boldsymbol{w}^T\boldsymbol{d} = 1$ yields

$$\hat{\boldsymbol{w}} = \frac{1}{\lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{d}, \tag{4.19}$$

which is equal to (4.13) for $\lambda = \boldsymbol{d}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{d}$.

In this way, the solution of the eigenvalue formulation of the CEM is equivalent to the original CEM formulation.

It may seem like nothing was gained by introducing this formulation, but the *iterated constrained energy minimization* (ICEM) scheme was inspired by the expression in (4.16) [Nielsen, 2001]. [Nielsen, 2001] notes that for the desired spectrum to dominate (first term on the right hand side in (4.16)), the covariance matrix could be estimated exclusively from target pixels. Further, [Nielsen, 2001] proposes to estimate the covariance matrix iteratively using pixels weighted by their projected value (onto $\boldsymbol{w}$). That is; in the first iteration, $\hat{\boldsymbol{\Sigma}}^{(1)}$ is estimated using the known target pixels. In iteration number t, $\hat{\boldsymbol{\Sigma}}^{(t)}$ is estimated from all pixels weighted by $\hat{\boldsymbol{w}}^{(t-1)T}\boldsymbol{x}$. When the algorithm converges, the final estimate of $\boldsymbol{w}$ is returned.

Geometrical Interpretation

To obtain a geometrical interpretation of the CEM method, it is convenient to consider the projection of a pixel vector $\boldsymbol{x}$ onto the CEM vector $\boldsymbol{w}$ and rewrite the inverse covariance matrix using the square-root decomposition [Manolakis and Shaw, 2002a]:

$$\begin{aligned}\boldsymbol{x}^T\boldsymbol{w} &\overset{19}{\propto} \boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{d} \\ &= \underbrace{\left(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}\right)}_{\text{I}}\underbrace{\left(\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{d}\right)}_{\text{II}},\end{aligned} \tag{4.20}$$

where $\boldsymbol{x}$ is a mean corrected pixel vector and $\boldsymbol{d}$ is the mean corrected target signature. From this way of writing, it is clear that the CEM method can be understood as two whitening transformations. The first term (I), can be thought of as the *zero-phase component analysis* (ZCA) *whitening* (See Appendix A.2) of $\boldsymbol{x}$ [Manolakis and Shaw, 2002a]. The ZCA whitening is simply a *scaling* that (without rotation) scales all pixels such that the covariance matrix of the pixels used to estimate $\boldsymbol{\Sigma}$ in (4.20) becomes the identity matrix. That is, pixel structures resembling the pixels "driving" the estimated covariance matrix are whitened while pixel structures *not* resembling the "driving"-pixels are stretched/squeezed into some unknown shape. The second term (II) is the same whitening transform applied to the target signature vector $\boldsymbol{d}$, and gives the "new" (scaled) direction[20] onto which the first term is projected (dashed line in the rightmost plot in Figure 4.5) [Manolakis and Shaw, 2002a]. Figure 4.5 illustrates this geometrical interpretation when the covariance matrix is

19. $\frac{1}{\boldsymbol{d}^T\boldsymbol{\Sigma}^{-1},\boldsymbol{d}}$ is a constant.
20. The direction in the transformed space.

estimated from all data points in the case where the target probability is low (sparsely distributed target).

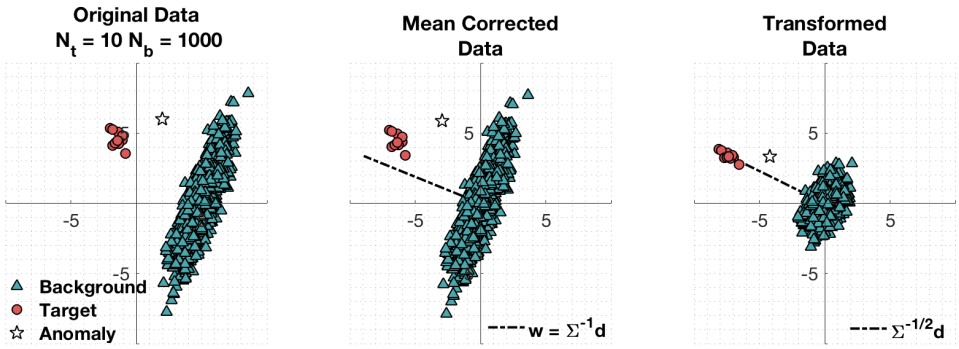

Figure 4.5: Original CEM: $\mathbf{\Sigma}$ is estimated based on all data points. However, since $N_t \ll N_b$, $\mathbf{\Sigma}_g \approx \mathbf{\Sigma}_b$ and $\boldsymbol{\mu}_g \approx \boldsymbol{\mu}_b$ (subscript b and g refer to *background* and *global*, respectively). Left: Original data. Middle: Mean corrected data (Equivalent to Figure 4.4). Since the global mean is approximately equal to the background mean this centers the background approximately around the origin. The dashed black line represent $\boldsymbol{w}$. Right: Mean corrected and transformed (by multiplication of term I in (4.20))) data with projection direction (term II in (4.20)) indicated by the dashed black line.

By estimating $\mathbf{\Sigma}$ in (4.20) based on target pixels only, as proposed in the ICEM method, it is the target class that is being whitened. According to [Manolakis and Shaw, 2002b], the optimum version of the statistic is estimated based on "target free" data. However, the statistic based on "background free" data may be just as good, or better. The results in either case will depend on the covariance structure (spread of the data) of the two classes, as illustrated in Figure 4.6. From this figure, it is clear that the projection of an anomaly (star) onto $\mathbf{\Sigma}^{-\frac{1}{2}}\boldsymbol{d}$ (dashed line) may indicate target for one choice of covariance estimate and background for the other.

Based on this analysis, this thesis will also (in addition to the ICEM method proposed in [Nielsen, 2001], weighting the target pixels), investigate an alternative ICEM method weighting the background pixels. From now on, the original ICEM algorithm, as proposed in [Nielsen, 2001], will be refered to as the ICEM1 method while the alternative ICEM algorithm will be refered to as the ICEM2 method.

Remarks

- The original CEM algorithm uses the data covariance to model the spectral variability of the background [Boardman and Kruse, 2011] and requires thus *low-probability* targets ($P_t \ll P_b$), such that $\mathbf{\Sigma} \approx \mathbf{\Sigma}_b$. For the ICEM algorithm, on the other hand, this assumption is not that important. By adjusting the weights, the contribution from background and target to the covariance estimate can be controlled such that $\mathbf{\Sigma} \approx \mathbf{\Sigma}_b$ or $\mathbf{\Sigma} \approx \mathbf{\Sigma}_t$.

- In the CEM algorithm, *all* anomalies (pixels not resembling the background) will get a score not equal to zero, even if their signatures differ from that of the target signature. This may lead to false positives and is referred to as the *selectivity problem* of CEM [Boardman and Kruse, 2011].

- As mentioned earlier, some of the algorithms presented in this thesis will under some conditions lead to the same solution. From the expressions of the FLDA and the CEM projection vectors,

$$\boldsymbol{w}_{\text{FLDA}} = (P_t\mathbf{\Sigma}_t + P_b\mathbf{\Sigma}_b)^{-1}(\boldsymbol{\mu}_t - \boldsymbol{\mu}_b), \tag{4.21}$$

$$\boldsymbol{w}_{\text{CEM}} \propto \mathbf{\Sigma}_g^{-1}(\boldsymbol{\mu}_t - \boldsymbol{\mu}_g), \tag{4.22}$$

  where $\mathbf{\Sigma}_t$, $\mathbf{\Sigma}_b$ and $\mathbf{\Sigma}_g$ are the target, background and global covariances, repectively, it is clear that $\boldsymbol{w}_{\text{FLDA}}$ reduces to $\boldsymbol{w}_{\text{CEM}}$ for low probability targets [Manolakis and Shaw, 2002b]. That is, when $P_t \longrightarrow 0$, the direction of the FLDA solution approaches that of the CEM solution since

$$P_t\mathbf{\Sigma}_t + P_b\mathbf{\Sigma}_b \longrightarrow \mathbf{\Sigma}_b, \tag{4.23}$$

$$\mathbf{\Sigma}_g \longrightarrow \mathbf{\Sigma}_b, \tag{4.24}$$

$$\boldsymbol{\mu}_g \longrightarrow \boldsymbol{\mu}_b. \tag{4.25}$$

  This means that the geometrical interpretation above is valid also for the FLDA when the target probability is low.

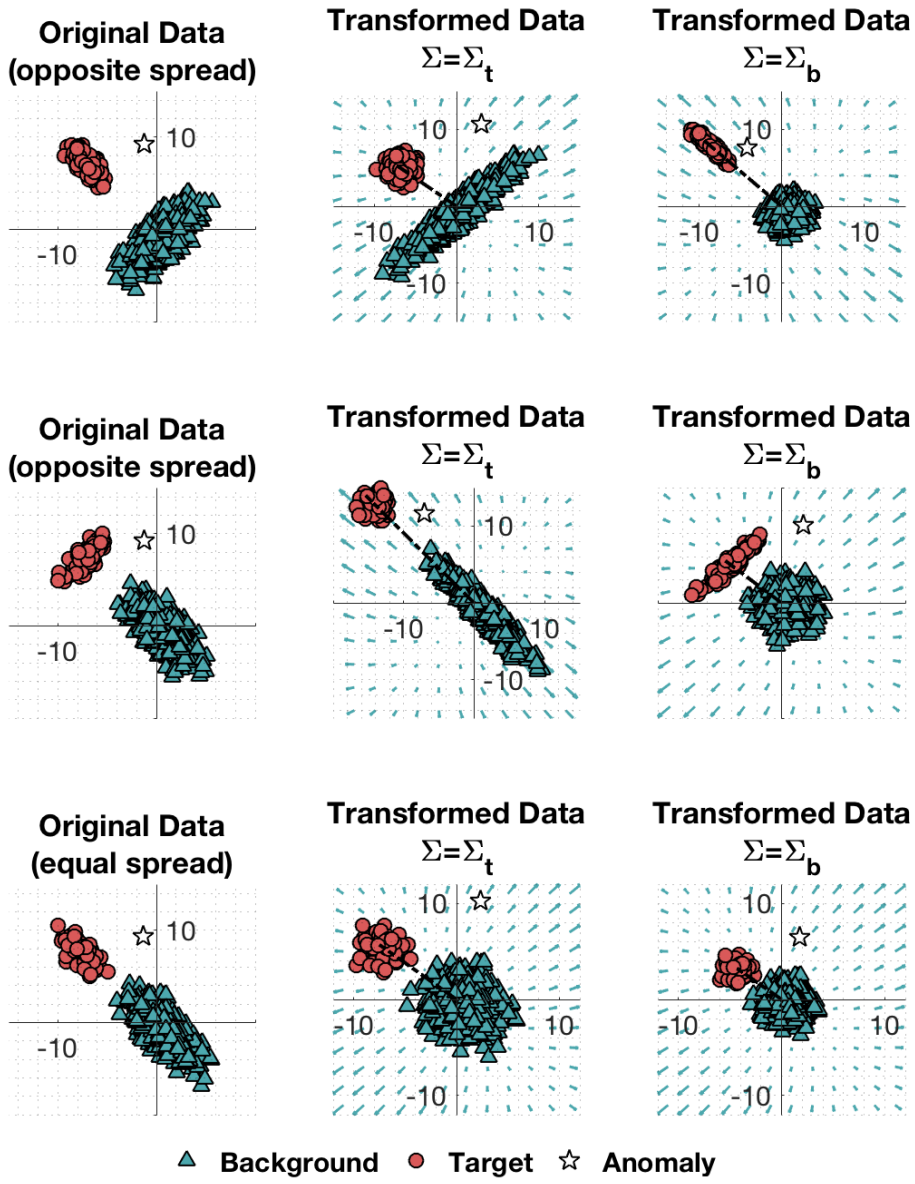

Figure 4.6: This figure illustrates the differences between estimating the covariance matrix based on target points and background points. Top: The target and the background have *opposite* covariance structures (northwest-southeast and southwest-northeast spread). Middle: The target and the background have *opposite* covariance structures (southwest-northeast and northwest-southeast spread). Bottom: The target and the background have *equal* covariance structures (northwest-southeast and northwest-southeast spread). The arrows indicate the scaling of the transformation. When the covariance structure of the classes are equal, the methods give similar results. When the covariance structures are different, on the other hand, the methods differ (the projection of the anomaly (star) onto $\mathbf{\Sigma}^{-\frac{1}{2}}\boldsymbol{d}$ (dashed line) may indicate target for one choice of transformation and background for the other)

.

4.3.3 Mixture-Tuned Matched Filtering

Mixture-Tuned Matched Filtering (MTMF) as proposed in [Boardman, 1998] is a spectral unmixing algorithm designed for hyperspectral data. The matched filtering (MF) part of MTMF is equivalent to the CEM method described in the previous section[21], and consists in suppressing the background and estimating the abundance of the known target. The mixture-tuning (MT) part of the MTMF algorithm is added to solve the *selectivity* problem of the classical MF methods, and rejects false positives by considering convex geometry theory [Boardman and Kruse, 2011].

The MTMF algorithm is divided into three main steps [Boardman and Kruse, 2011]: *i)* Pre-processing and Data Conditioning, *ii)* Matched Filtering and *iii)* Mixture Tuning. The rest of this section presents these steps in more detail.

i) Pre-processing and Data Conditioning
The first step in the MTMF algorithm is to apply the *Minimum Noise Fraction* (MNF) transformation. This is a twofold process and consists in PCA *whitening* the noise and decorrelating the data [Boardman and Kruse, 2011].

To be able to whiten the noise, the noise covariance matrix must be estimated, and in [Boardman and Kruse, 2011] this is done by the *shift difference* method. This method utilizes the fact that neighboring pixels usually are strongly spatially correlated, while the noise show a much lower correlation [Sun et al., 2014]. In this way, the difference between neighboring pixels will mainly depict the noise[22].

The input image is represented by a size $r \times c \times p$ matrix, $\boldsymbol{X}$, where r, c and p are the number of rows, columns and bands, respectively. For each pixel in each band, x_{ijk}, the *differences* between the pixel value and the pixel values in the previous row and column are computed. The average value of these *shift differences* is stored in a size $r-1 \times c-1 \times p$ *noise* image, $\boldsymbol{N}$. That is, the elements of $\boldsymbol{N}$ is given by:

$$n_{ijk} = \frac{1}{2}\left(x_{ijk} - x_{(i-1)jk}\right) + \frac{1}{2}\left(x_{ijk} - x_{i(j-1)k}\right), \tag{4.26}$$

where i, j, k indicates row, column and band number, respectively [Boardman and Kruse, 2011]. Then, $\boldsymbol{N}$ is treated as a 2D-matrix of size $((r-1) * (c -$

21. The MF projection vector is derived from the log-likelihood ratio test statistic when assuming Gaussian distributions with equal covariance matrices for the target and the background. The CEM reduces to MF when the mean is removed from all data points and the desired target signature [Manolakis et al., 2003].
22. Sharp edges in the scene will also (incorrectly) contribute to the noise estimate

1)) $\times\, p$, and the noise covariance estimate, $\hat{\boldsymbol{\Sigma}}_n$, is calculated by the sample covariance[23].

After that, the image is mean corrected (MC) ($\boldsymbol{X} \rightarrow \boldsymbol{X}_{MC}$) and treated as an $N \times p$ matrix, where the $N = r * c$ rows represent the p-dimensional pixel vectors. The noise whitened (NW) data, $\boldsymbol{X}_{NW}$, is then obtained by applying the PCA whitening transform (see Appendix A.1) based on the estimated noise covariance [Boardman and Kruse, 2011]:

$$\boldsymbol{X}_{NW} = \boldsymbol{W}^T \boldsymbol{X}_{MC} = \boldsymbol{\Lambda}_n^{-\frac{1}{2}} \boldsymbol{E}_n^T \boldsymbol{X}_{MC}, \tag{4.27}$$

where $\boldsymbol{W} = \boldsymbol{E}_n \boldsymbol{\Lambda}_n^{-\frac{1}{2}}$ is the PCA whitening matrix and $\boldsymbol{\Lambda}_n$ and $\boldsymbol{E}_n$ are the diagonal eigenvalue matrix and eigenvector matrix of $\hat{\boldsymbol{\Sigma}}_n$, respectively.

The first step of the MTMF method is then completed by applying the PCA algorithm (see Appendix B) to $\boldsymbol{X}_{NW}$ to obtain the MNF transformed data:

$$\boldsymbol{X}_{MNF} = \boldsymbol{E}_{x_{NW}}^T \boldsymbol{X}_{NW}, \tag{4.28}$$

where $\boldsymbol{E}_{x_{NW}}$ is the matrix whose columns are the eigenvectors of the covariance matrix estimated from $\boldsymbol{X}_{NW}$, $\hat{\boldsymbol{\Sigma}}_{x_{NW}}$ [Boardman and Kruse, 2011].

In this way, the data in the MNF space has uncorrelated noise with unit variance, and all variables are uncorrelated. As will be seen later, this step is essential for the MT part of the algorithm, where the validity of the detections are quantified to remove the false positives.

ii) Matched Filtering

The second step in the MTMF algorithm is the matched filtering. The form of the MF projection is equivalent to that of the CEM projection in 4.13:

$$\boldsymbol{w} = \frac{\hat{\boldsymbol{\Sigma}}_{x_{MNF}}^{-1} \boldsymbol{d}_{MNF}}{\boldsymbol{d}_{MNF}^T \hat{\boldsymbol{\Sigma}}_{x_{MNF}}^{-1} \boldsymbol{d}_{MNF}}, \tag{4.29}$$

where $\boldsymbol{d}_{MNF}$ is the $p \times 1$ vector describing the desired target signature in the MNF space and $\hat{\boldsymbol{\Sigma}}_{x_{MNF}}$ is the estimated covariance matrix of the MNF transformed data (which is approximately the background covariance if the target is sparsely distributed). Projecting $\boldsymbol{X}_{MNF}$ onto this MF vector results (in the same way as for CEM) in the MF abundance image (vector of size $N \times 1$),

$$\boldsymbol{x}_{MF} = \boldsymbol{X}_{MNF} \boldsymbol{w}, \tag{4.30}$$

23. Since $n_{ijk} = \frac{1}{2}(x_{ijk} - x_{(i-1)jk}) + \frac{1}{2}(x_{ijk} - x_{i(j-1)k}) = x_{ijk} - 0.5x_{(i-1)jk} - 0.5x_{i(j-1)k}$, and estimating the noise covariance implies squaring N, the covariance estimate must also be divided by $1 + 0.5^2 + 0.5^2 = 1.5$ [Boardman and Kruse, 2011].

containing the MF score of each pixel [Boardman and Kruse, 2011]. That is, $\boldsymbol{x}_{MF} = [x_{MF,1}, ..., x_{MF,N}]^T$.

iii) Mixture Tuning

The final step of the MTMF algorithm intends to improve the selectivity property of classical MF methods. In standard MF algorithms, *all* anomalies[24] returning a high MF score are detected as targets, even if their signatures may not resemble the true target signature. By including the abundance non-negativity constraint and the abundance sum constraint and considering the convex geometry of the LMM (discussed in Section 4.2.1), the MT step can automatically reject false positives [Boardman and Kruse, 2011].

Since a pure target pixel (according to the LMM) cannot exhibit any mixing, all variations must be explained by the noise. A background pixel, on the other hand, may form as any feasible combination of the $L-1$ background endmembers, which allows for variations equal to the background variation *plus* variations explained by noise. In this way, the mean target spectrum can be considered one vertex of the mixing $(L-1)$-simplex, while the background forms the opposite $L-2$ dimensional facet[25]. This means that all feasible mixtures between target and background are captured by an $(L-1)$-simplex [Boardman and Kruse, 2011].

The idea of the MT step is to reject all detections located *outside* this simplex, as they are infeasible mixtures of target and background, in this way rejecting anomalies that do not resemble the target signature. To be able to do this, the range of distributions (characterized by their means and standard deviations) describing the feasible mixtures between target and background are interpolated. Each MF score corresponds to a distribution, ranging from a score of zero (background distribution[26]) to a score of one (target distribution[27]). Figure 4.7 illustrates these distributions using transparent ellipsoids. Then, for each pixel, an *infeasibility* number (MT score) is calculated as the distance from the pixel to the mean of the distribution corresponding to the pixel's MF score (distance from pixel to the red solid line, along a score contour, in Figure 4.7), measured in number of standard deviations of the same distribution [Boardman and Kruse, 2011].

24. Data points not characterized by the covariance matrix in the whitening filter. If $P_b \gg P_t$, this is approximately the same as "data points not resembling the background".
25. A *face* of a n-simplex is a intersection between the simplex and a tangent hyperplane. Zero-dimensional faces are *vertices*, one-dimensional faces are *edges* and $(n-1)$-dimensional faces are called *facets* [Weisstein, n.d.-a]. That is, the facets of a n-simplex are the $(n-1)$-simplices obtained by removing one vertex.
26. Characterized by the mean background signature and background standard deviations.
27. Characterized by the target signature and unit standard deviations

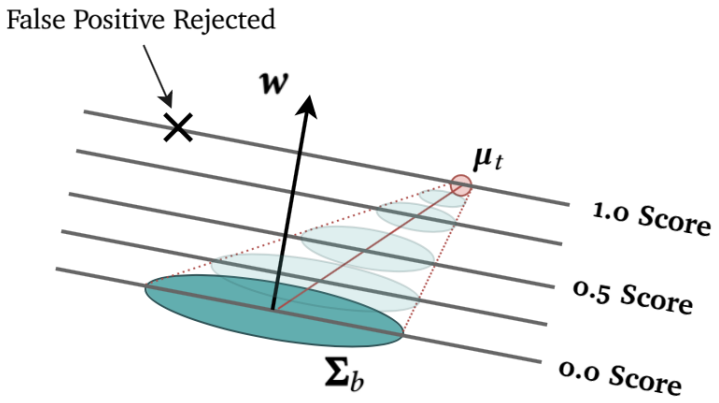

Figure 4.7: The MT step: The transparent ellipsoids represent the distributions linearly interpolated from the background distribution and target distribution. For each MF score, there is a corresponding distribution. The infeasibility number is calculated as the distance from the pixel to the mean of the distribution corresponding to the pixel's MF score (solid red line), measured in number of standard deviations of the same distribution. Modified from [Boardman and Kruse, 2011].

The interpolations of the distributions corresponding to each MF score rely on the MNF transform. In the MNF space, the noise is uncorrelated with unit variance (due to the noise whitening) *and* the variables are uncorrelated (due to the PCA). Since covariance matrices of uncorrelated variables are diagonal (covariences are zero) with elements (variances) equal to the eigenvalues, the interpolated values of the standard deviations (corresponding to the range of distributions) are easily accessed by linear interpolation between the unit standard deviation of the target and the square roots of the eigenvalues of the background. The interpolated standard deviation vector associated with each pixel is thus calculated as [Boardman and Kruse, 2011]

$$\boldsymbol{sd}_i = \boldsymbol{l}_{x_{MNF}}^{\frac{1}{2}} - x_{MF,i}(\boldsymbol{l}_{x_{MNF}}^{\frac{1}{2}} - \mathbf{1}), \tag{4.31}$$

where $\boldsymbol{l}_{x_{MNF}} = diagonal(\boldsymbol{\Lambda}_{MNF})$ is the *vector* whose elements are the eigenvalues of the covariance matrix estimated from the MNF data (here, $(\cdot)^{\frac{1}{2}}$ denotes the *element-wise* square root operator) and $\mathbf{1}$ represent the unit standard deviation vector of the noise. To constrain the values of $\boldsymbol{sd}$ to range from $\boldsymbol{l}_{x_{MNF}}^{\frac{1}{2}}$ (corresponding to the standard deviation of the background) to $\mathbf{1}$ (corresponding to the standard deviation of the noise), $\boldsymbol{x}_{MF}$ is floored at zero and ceilinged at one [Boardman and Kruse, 2011].

Further, since the target is assumed to be sparsely distributed in the scene, the mean correction in the MNF transform centers the background approximately

at the origin, such that the mean background signature is the zero vector. The interpolated mean vectors of the feasible mixture distributions are thus obtained by linear interpolation between the origin and the mean target signature. The interpolated mean vector corresponding to the MF score of the ith pixel is thus given by [Boardman and Kruse, 2011]:

$$\boldsymbol{m}_i = x_{MF,i}\boldsymbol{d}_{MNF}, \tag{4.32}$$

such that the distribution mean ranges from the zero vector for MF scores of zero (background) to $\boldsymbol{d}_{MNF}$ for MF scores of one (target).

The infeasibility number (MT score) is finally calculated for each pixel, $\boldsymbol{x}_{MNF,i} = [x_{MNF,i1}, \ldots, x_{MNF,ip}]$[28] by subtracting the interpolated mean and dividing *element-wise* by the interpolated standard deviation of the distribution corresponding to the pixel's MF score [Boardman and Kruse, 2011]:

$$x_{MT,i} = \left| \frac{\boldsymbol{x}_{MNF,i}^T - \boldsymbol{m}_i}{\boldsymbol{sd}_i} \right|_2^2. \tag{4.33}$$

By combining the information from the MT and the MF scores, it is thus possible to detect anomalies *and* decide whether they are feasible mixtures of background and target or not.

With that being said, it is not obvious *how* to combine the scores to obtain one *MTMF score*, and the issue is not adressed in [Boardman and Kruse, 2011]. However, to be able to use the MTMF method in an automatic detection system, this is necessary.

In this thesis, a simple heuristic method involving a decreasing sigmoid function is proposed to combine the MT and the MF score for each pixel:

$$x_{MTMF,i} = \frac{x_{MF,i}}{e^{a(x_{MT,i}-b)} + 1}, \tag{4.34}$$

where a and b are positive constants, determining the steepness and center position of the Sigmoid, respectively. By choosing a high a value, the Sigmoid becomes a step function such that MT values less than b get a response of one and MT values greater than b get a response of zero. In this way, it is possible to disregard samples with MT scores greater than b ($x_{MTMF} = 0$) while keeping samples with MT scores smaller than b unchanged ($x_{MTMF} = x_{MF}$). This is illustrated in figure 4.8. Note that the MTMF method reduces to pure MF when the b-value is set very high (compared to the maximum MT value).

28. The ith row of $\boldsymbol{X}_{MNF}$.

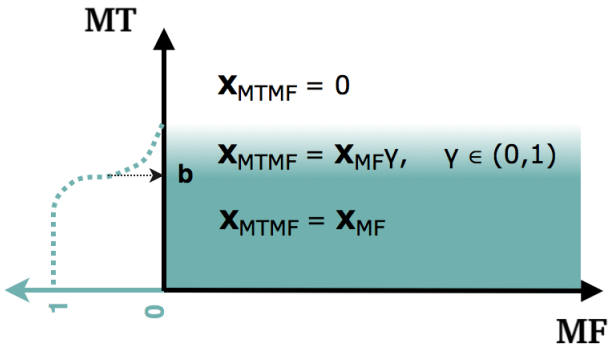

Figure 4.8: Combination of the MF score and the MT score to obtain the MTMF score. A high MTMF score will in this way indicate an anomaly that is a feasible mixture between the background and the target.

Remarks

- By the design of the algorithm, the target is characterized by one vector (the mean target signature). This allows for little spectral variation in the target.

- The estimated MF statistic is (like the CEM statistic) ideal when the covariance matrix is esimated based on background (target) pixels only [Manolakis and Shaw, 2002b].

- The noise estimation in the MNF transformation is basically a high-pass filter. Sharp edges in the scene will hence incorrectly contribute to the noise estimate. The higher the spatial frequency in the image, the poorer the noise estimate becomes, and the method should ideally be applied to homogeneous areas [Bioucas-Dias and Nascimento, 2008]. Further, noise estimation in the spatial domain is shown to be unstable in multiple studies, and regular noise (e.g. striping[29]) may be mistaken as signal [Gupta and Bajaj, 2018, Luo et al., 2016].

- For the MTMF method to yield better results than pure MF, there are two requirements upon the target and the background classes:

    1. The target class must have a sufficiently low spectral variability.

    2. There must exist false positives, i.e. samples in the background class

29. Striping is reffered to as a radiometric error which may appear in optical scanner systems. Differences in sensitivity of the detectors in the sensor can result in banding/striping in the image [Campbell, 2011].

not resembling the background.

A high spectral variability in the target class may lead to a lot of target samples having a high MT score: They do not resemble the target representative (target mean) and will thus *not* qualify as a mixture between target and background. Removing these samples will obviously not enhance the performance. Even if there exist false positives, they can not be removed (set to zero) without removing true positives as well.

If the background class consists of *one* material (with somewhat constant signature), the number of possible false positives will be low: all background samples are well represented by the sample covariance, and will thus get low MF scores. On the other hand, if the background class consist of multiple materials, and especially if one of the materials is over-represented, there might be more potential false alarms: If one of the background materials (say grass) is heavy represented compared to the other materials (for example asphalt and buildings), the sample covariance will best represent the grass, while the buildings and asphalt might get high MF scores (false positives).

Since the assumed pre-processing step in this thesis makes the detection problem a two-class classification problem, any possible enhancements due to the MT step is thus not expected to be very large.

5

Target Detection

This chapter starts by presenting the fundamentals of target detection. Then it describes the *constant miss rate* (CMR) detector, which will be used in the experiment to examine the algorithms described in the previous chapters. Finally, a short introduction to the receiver operating characteristic (ROC) curve, the area under the curve (AUC) and their roles in assessing the detection performance is given.

5.1 The Binary Detection Problem

The basic idea of target detection is to examine each pixel vector and, based on some *decision rule*, determine whether or not the target is present. To be able to form this decision rule and theoretically evaluate it, the first step is to assume some probability densities for the "target present" pixels and the "target absent" pixels. In this way, each pixel (observation) $\boldsymbol{x}$ is treated as a random variable, and the detecton problem is specified by the following competing hypothesis [Manolakis, 2016]:

$$\begin{aligned} H_0 : \quad & \boldsymbol{x} \sim p(\boldsymbol{x}|H_0) \quad \text{(Target Absent)}, \\ H_1 : \quad & \boldsymbol{x} \sim p(\boldsymbol{x}|H_1) \quad \text{(Target Present)}, \end{aligned}$$

which yield four potential outcomes, presented in Table 5.1.

Table 5.1: Possible detection outcomes. Modified from [Manolakis, 2016].

		State of Nature	
		H_0 true (target absent)	H_1 true (target present)
Decision	Choose H_0	Correct (Rejection)	Type 2 error (Miss)
	Choose H_1	Type 1 error (False Alarm)	Correct (Detection)

The decision rule is then defined by a test statistic

$$y = D(\boldsymbol{x}), \tag{5.1}$$

and a threshold, η such that if $y > \eta$ ($y < \eta$), H_1 (H_0) is accepted. The decision rule will thus divide the observation space into disjoint regions $\mathcal{R}_0$ and $\mathcal{R}_1$ such that [Manolakis, 2016]

$$\begin{aligned} \boldsymbol{x} \in \mathcal{R}_0 : &\quad \Rightarrow \quad \text{Accept} \quad H_0, \\ \boldsymbol{x} \in \mathcal{R}_1 : &\quad \Rightarrow \quad \text{Accept} \quad H_1. \end{aligned}$$

5.1.1 CMR Detection Algorithm

The concept of the CMR detection algorithm is equivalent to that of the more familiar *constant false alarm rate* (CFAR) detection algorithm. The difference is that the CMR detector estimates the probability density function (pdf) for the data under H_1 and not H_0. That is, it models the "target present" pixels and treats the background as unknown.

The reason for this choice is that the target (infected trees) is assumed to be more easily modeled. If the pre-processing step, extracting pixels that earlier have shown a "healthy-pine"-like signature, had been implemented, a CFAR detector based on modeling the healthy forest pixels could have been used. However, since the pre-processing step is replaced by a simple NDVI threshold operation, the background class is expected to contain more than healthy trees, leading to a greater spectral variability.

The idea of the CMR detector is to set the threshold such that the probability of miss,

$$P_M = \int_{\boldsymbol{x} \in \mathcal{R}_0} p(\boldsymbol{x}|H_1)P(H_1)d\boldsymbol{x}, \tag{5.2}$$

is equal to a specified constant. In the pixel space, this threshold is some closed surface in $\mathbb{R}^p$. However, after applying the machine learning techniques

discussed in the previous chapters, the pixel vectors are (in this thesis) mapped to $\mathbb{R}^1$ and the threshold is reduced to a single scalar[1] η. In the mapped space, $y = D(\boldsymbol{x}) \in \mathbb{R}^1$, the probability of miss is given by

$$P_M = \int_{-\infty}^{\eta} p(y|H_1)P(H_1)dy. \tag{5.3}$$

To be able to determine the threshold(s) for a given miss rate, the sampling distribution of $y = D(\boldsymbol{x})$ is thus required. In the next section, the sampling distributions of the output from the different detectors are discussed.

Exact Sampling Distributions

The statistical properties of a detector are determined by the probability distribution of its output (the test statistic) [Manolakis and Shaw, 2002b]. Since a test statistic is a function of the data, the probability distribution of the observations must be considered as well. When the observations are assumed to be random samples from an *unknown* distribution, the problem aggravates: The distribution of the data needs to be estimated from the samples, making the parameters random variables with their own pdfs. However, sometimes it is still possible to find the sampling distribution of the test statistic, if simple and tractable distributions are assumed for the random quantities involved.

The FLDA statistic, for instance, is dependent upon the estimation of the within-class covariance matrix and the mean vectors of both the target and the background. Assuming that the spectral intensities in the target and background pixels follow unique Gaussian distributions, the target and background models can be expressed as p-variate Gaussian distributions:

$$\boldsymbol{X}_i \sim \mathcal{N}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{5.4}$$

for $i = \{t, b\}$, indicating the target and background model.

The maximum likelihood estimate of the mean vector $\boldsymbol{\mu}$, based on a (general) p-variate Gaussian random sample $\mathcal{D} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n\}$, is given by [Johnson and Wichern, 2007]

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{\boldsymbol{x} \in \mathcal{D}} \boldsymbol{x}, \tag{5.5}$$

which follows a Gaussian distribution:

$$\hat{\boldsymbol{\mu}} \sim \mathcal{N}_p(\boldsymbol{\mu}, \tfrac{1}{n}\boldsymbol{\Sigma}). \tag{5.6}$$

1. In a two-tailed test, there will be two thresholds, i.e. two scalars.

The difference,

$$\hat{\boldsymbol{\mu}}_t - \hat{\boldsymbol{\mu}}_b, \tag{5.7}$$

assuming independence, is thus determined by

$$\begin{aligned} E\{\hat{\boldsymbol{\mu}}_t - \hat{\boldsymbol{\mu}}_b\} &= E\{\hat{\boldsymbol{\mu}}_t\} - E\{\hat{\boldsymbol{\mu}}_b\} \\ &= \boldsymbol{\mu}_t - \boldsymbol{\mu}_b \qquad (5.8) \\ \mathrm{Cov}\{\hat{\boldsymbol{\mu}}_t - \hat{\boldsymbol{\mu}}_b\} &= \mathrm{Cov}\{\hat{\boldsymbol{\mu}}_t\} + \mathrm{Cov}\{\hat{\boldsymbol{\mu}}_b\} \\ &= \frac{1}{n_t}\boldsymbol{\Sigma}_t + \frac{1}{n_b}\boldsymbol{\Sigma}_b, \qquad (5.9) \end{aligned}$$

which results in a Gaussian, $\mathcal{N}_p(\boldsymbol{\mu}_t - \boldsymbol{\mu}_b, \frac{1}{n_t}\boldsymbol{\Sigma}_t + \frac{1}{n_b}\boldsymbol{\Sigma}_b)$ distribution.

Further, the unbiased sample covariance estimate, based on the same random sample $\mathcal{D}$, is given by [Johnson and Wichern, 2007]

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{\boldsymbol{x}\in\mathcal{D}}(\boldsymbol{x} - \hat{\boldsymbol{\mu}})(\boldsymbol{x} - \hat{\boldsymbol{\mu}})^T, \tag{5.10}$$

and follows a Wishart distribution:

$$(n-1)\hat{\boldsymbol{\Sigma}} \sim W_p(n-1, \boldsymbol{\Sigma}). \tag{5.11}$$

Assuming that $P_t \ll P_b$, the FLDA detector statistic can be approximated as

$$D_{\mathrm{FLDA}}(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{w}_{\mathrm{FLDA}} \approx \boldsymbol{x}^T \boldsymbol{\Sigma}_b^{-1}(\boldsymbol{\mu}_t - \boldsymbol{\mu}_b), \tag{5.12}$$

where $\boldsymbol{x}$ is an observation drawn from the Gaussian target distribution. In this way, the $D_{\mathrm{FLDA}}(\boldsymbol{x})$ combines a Wishart, $W_p(n_b - 1, \boldsymbol{\Sigma}_b)$, random matrix, an Gaussian, $\mathcal{N}_p(\boldsymbol{\mu}_t - \boldsymbol{\mu}_b, \frac{1}{n_t}\boldsymbol{\Sigma}_t + \frac{1}{n_b}\boldsymbol{\Sigma}_b)$, random vector and an Gaussian, $\mathcal{N}_p(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$, random vector in the form

$$\begin{pmatrix} \text{Gaussian} \\ \text{random vector 1} \end{pmatrix}^T \left(\frac{\text{Wishart random matrix}}{\text{d.f.}}\right)^{-1} \begin{pmatrix} \text{Gaussian} \\ \text{random vector 2} \end{pmatrix},$$

where d.f. is short for degrees of freedom. This decomposition is inspired by a similar decomposition in [Johnson and Wichern, 2007].

Without any further assumptions, the distribution of this statistic is (to the author's knowledge) unknown. However, the distribution of the FLDA is a well explored problem. In [John, 1961] the exact distribution is derived under the assumption that $\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_b = \boldsymbol{\Sigma}$ is known, and in [Sigreaves, 1961] the exact distribution is expressed by an infinite series, assuming $n_t = n_b$. Multiple studies on the asymptotic distribution have also been carried out, including [Anderson, 1973], who studied the asymptotic distribution of the discriminant function when $n_t \to \infty$ and $n_b \to \infty$ [Neto and Dougherty, 2015]. A recent

study, [Bodnar et al., 2017], also examines the distribution properties of the discriminant function, however, also this one under the assumption of equal covariance matrices.

When it comes to the distribution of the MF statistic, it was determined exactly in [Richmond, 1996]. However, the MF statistics considered in this thesis are either a modified version (the ICEM statistic replaces the background covariance with a weighted covariance matrix) or is only part of the test statistic (the MTMF statistic combines the MT and the MF score).

The derivation of the sampling distributions of the detector statistics in Table 5.2 is a complicated problem and in this thesis the distributions will be estimated empirically by Monte Carlo simulations.

Table 5.2: Overview of the methods to be tested in this thesis.

Method	Mathematical Formulation
FLDA	$\boldsymbol{w}_{\text{FLDA}} = (P_t\boldsymbol{\Sigma}_t + P_b\boldsymbol{\Sigma}_b)^{-1}(\boldsymbol{\mu}_t - \boldsymbol{\mu}_b)$
SLDA	$\boldsymbol{w}_{\text{SLDA}} = \arg\min_{\boldsymbol{\theta}_k} \left\{\lvert Y\boldsymbol{\theta}_k - X\boldsymbol{w}_k\rvert_2^2\right\}$ s. t. $\frac{1}{N}\boldsymbol{\theta}_k^T Y^T Y\boldsymbol{\theta}_k = 1$ $\boldsymbol{\theta}_k^T Y^T Y\boldsymbol{\theta}_l = 0 \quad \forall \; l < k$
2D-Corr-NLS	$\boldsymbol{\alpha}_k = \arg\min_{\boldsymbol{\alpha}_k} \lvert M\boldsymbol{\alpha}_k - \boldsymbol{x}_k\rvert_2^2$ s. t. $\alpha_{ki} \geq 0 \;\; i = 1, \ldots, L$
ICEM	$\boldsymbol{w}_{\text{ICEM}} = \frac{\boldsymbol{\Sigma}_w^{-1}\boldsymbol{d}}{\boldsymbol{d}^T\boldsymbol{\Sigma}_w^{-1}\boldsymbol{d}}$, subscript w : weighted
MTMF	$\boldsymbol{w}_{\text{MF}} = \frac{\hat{\boldsymbol{\Sigma}}_{x_{MNF}}^{-1}\boldsymbol{d}_{MNF}}{\boldsymbol{d}_{MNF}^T\hat{\boldsymbol{\Sigma}}_{x_{MNF}}^{-1}\boldsymbol{d}_{MNF}}$ $x_{MT,i} = \left\lvert\frac{\boldsymbol{x}_{MNF,i}^T - \boldsymbol{m}_i}{s\boldsymbol{d}_i}\right\rvert_2^2$ $x_{MTMF,i} = \frac{x_{MF,i}}{e^{a(x_{MT,i}-b)}+1}$

Estimated Sampling Distributions: Monte Carlo

When the pdf of a test statistic is unknown, one alternative is to estimate the distribution via *Monte Carlo* simulations [Theodoridis, 2009]. The idea of Monte Carlo experiments is to generate S samples from the assumable data distribution, $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_S$ and feed them into the test statistic $y = D(\boldsymbol{x})$.

The histogram of the output values $y_1, y_2, ..., y_S$ will then approximate the unknown pdf with a level of accuracy determined by the sample size [Murphy, 2012].

In practice, to estimate the CMR threshold(s) for the different detector statistics, S samples form the assumed target distribution[2] are generated and evaluated by each statistic. From the design of the ICEM and the MTMF statistic, it is known that the target will achieve *higher* values than the background. The threshold corresponding to a specified miss rate of α is thus set at the nth smallest value such that $\frac{n}{S} = \alpha$ (Figure 5.1). For FLDA and SLDA, on the other hand, there is no way of knowing whether the target gets higher or lower values than the background, and a two-tailed test is required. This means that the thresholds are set at the nth smallest and the nth largest value such that $\frac{n}{S} = \frac{\alpha}{2}$.

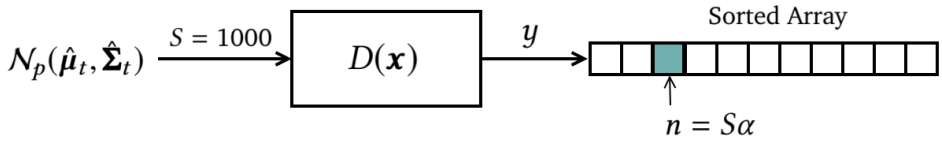

Figure 5.1: $S = 1000$ samples are generated from a Gaussian distribution (specified by the sample mean and sample covariance of the target samples in the training data) and fed into the test statistic. The output is sorted and the threshold is set to be the value of the nth smallest element.

5.2 Assessment

Since there is a trade-off between pushing the threshold one way to increase the number of detections, and pushing it the other way to avoid false alarms, it is necessary to examine both the probability of false alarm and the probability of detection to give an assessment of the detector's performance.

The receiver operating characteristics (ROC) curve is a plot of the probability of detection (5.13) against the probability of false alarm (5.14) as a function of the threshold [Manolakis, 2016]. In this way, the curve visualizes the trade-off at all possible values of the threshold and is a good descriptor of the performance.

2. In this thesis, the Gaussian distribution is used to model the target, and the samples are generated from a p-variate Gaussian distribution, specified by the sample mean and sample covariance of the target samples in the training data.

The probability of detection, P_D, and the probability of false alarm, P_{FA}, are given by:

$$P_D = \frac{n_{\text{detection}}}{n_{\text{detection}} + n_{\text{miss}}} \tag{5.13}$$

$$P_{FA} = \frac{n_{\text{false alarm}}}{n_{\text{false alarm}} + n_{\text{rejection}}}, \tag{5.14}$$

where $n_{\text{detection}}$ is the number of correctly detected targets, n_{miss} is the number of missed targets, $n_{\text{false alarm}}$ is the number of incorrect detections and $n_{\text{rejection}}$ is the number of samples correctly rejected by the detector.

Figure 5.2 shows the ROC curves from three different tests and indicates their relative performances. The closer the curve is to $(0, 1)$ (a detection rate of one without any false alarms), the better the performance. The "by chance" diagonal (dashed red line) indicates a test performance equivalent to random guessing: The number of detections increases at the same rate as the number of false alarms. This behavior witnesses about a detector that is unable to distinguish between target and background.

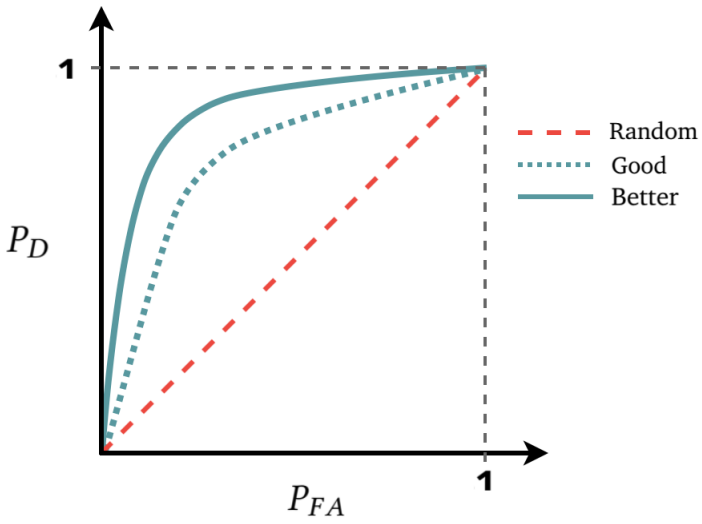

Figure 5.2: Example of ROC curves indicating the performances of three different tests. The best test gives the curve closest to $(0, 1)$ with the highest probability of detection achieved with the smallest probability of false alarm.

A standard way to summarize the ROC curve in one scalar is to consider the *area under the curve* (AUC) [Murphy, 2012]. The closer the area is to one,

the better the performance. Being threshold independent and numeric, the AUC is a very popular performance measure, although it has some potential shortcomings, examined in [Hand, 2009, Lobo et al., 2008].

Part II

Experiment

6 Method

The methods from the theory part are to be compared using two data sets; one fully ground truthed data set and one data set from Portugal showing trees, some of which are infected by PWN. This chapter starts by presenting these data sets and continues with describing the design and setup of the experiment.

6.1 Data Set 1: HYDICE

In order to compare the different methods described in the theory and explore the effects of spatial and spectral resolution in a satisfactory manner (obtain robust statistics[1]), the methods are first tested on a fully ground-truthed data set.

The "Urban" hyperspectral data set is published by the Geospatial Research Laboratory (U.S.)[2] and provided in MATLAB format with corresponding unmixing friendly ground truth (fractional abundance maps obtained in [Zhu et al., 2014c, Zhu et al., 2014b, Zhu et al., 2014a])[3].

1. Mean and variance of the performance.
2. Available at `http://www.agc.army.mil/Missions/Imagery/`
3. Available at `http://lesun.weebly.com/hyperspectral-data-set.html`

The data set was collected by the Hyperspectral Digital Imagery Collection Experiment (HYDICE) sensor in October 1995 and includes a 307×307 pixels image of two meters spatial resolution, sampled in 210 (162 after corrections) spectral channels throughout the visible and near infrared part of the spectrum [Zhu et al., 2014c]. The scene encompasses the outskirts of an urban area with multiple rows of houses, a big mall with a parking lot and a large field of grass and trees (Figure 6.1). Six disparate materials are identified and their corresponding abundance maps (ground truth) are displayed in Figure 6.2.

Figure 6.1: RGB composite of the original HYDICE image.

To be able to examine the effect of spectral resolution, a *semi-multispectral* image is generated by selecting and averaging over small intervals of channels in the areas of the blue, green, red, red-edge and near-infrared Sentinel-2 bands (resulting in five bands). In the same way, to explore the effect of spatial resolution, the hyperspectral and semi-multispectral images are resampled by averaging over 3^2 and 5^2 pixels to simulate spatial resolutions of six and ten meters, respectively. This gives a total of six HYDICE images, summarized in Table 6.1.

6.2 Data Set 2: Coruche

The second data set contains images from the Sentinel-2, RapidEye-2 and WorldView satellites (sensor information is summarized in Table 6.3), acquired over Couruche, Santarém, Portugal (depicted on the front page), in 2016.

The WorldView and RapidEye data are provided free of charge under ESA's

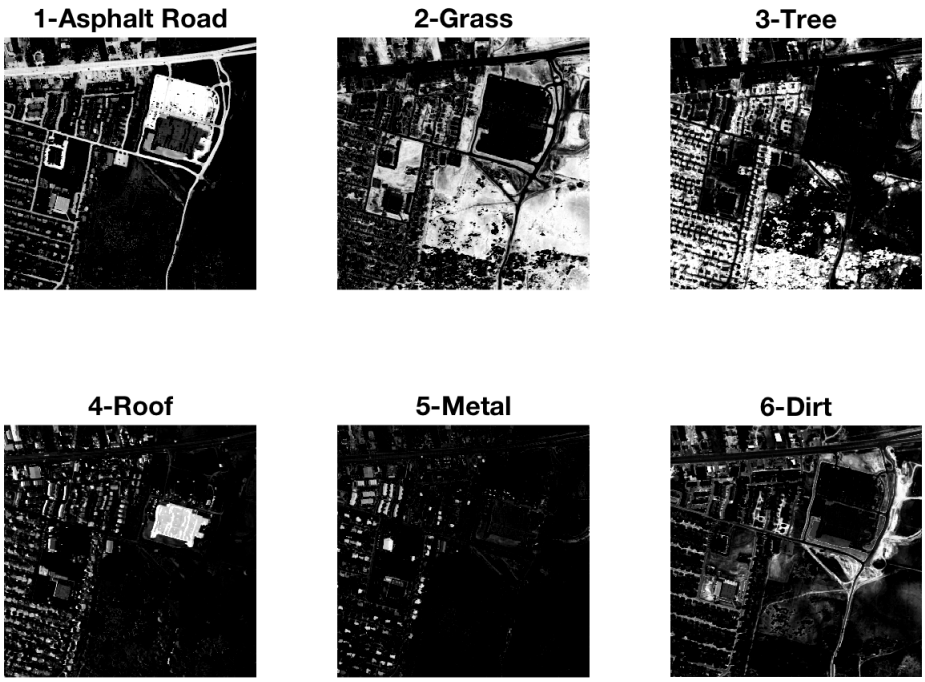

Figure 6.2: Ground truth: Abundance maps of the six identified disparate materials in the scene (obtained in [Zhu et al., 2014c, Zhu et al., 2014b, Zhu et al., 2014a]). A black pixel corresponds to zero abundance and a white pixel corresponds to 100% abundance of the material.

Table 6.1: The HYDICE data set. Based on the original HYDICE_h2 image, five "new" images are generated to simulate different spatial and spectral resolutions.

		Spatial Resolution		
		2 m	6 m	10 m
Spectral Resolution	Hyperspectral	HYDICE_h2	HYDICE_h6	HYDICE_h10
	Multispectral	HYDICE_m2	HYDICE_m6	HYDICE_m10

Third Party Mission scheme, project ID 34530, through an agreement between ESA and Arnoud Jochemsen, Science [&] Technology, as Principal Investigator of the ESA project "Silvisense: Forest Health Monitoring" (ESA contract no. 4000117722/16/INB), where UiT was a partner.

The Sentinel-2 data was downloaded from the Copernicus Open Access Hub (`https://scihub.copernicus.eu`), which provides free and open access to

Sentinel-1, Sentinel-2 and Sentinel-3 data.

- Sentinel-2: © ESA (2016).
- WorldView: © DigitalGlobe, Inc. (2016), provided by European Space Imaging.
- RapidEye: © (2016) BlackBridge S.àr.l. All rights reserved.

The Coruche region is one of the regions most affected by the PWN[4], and within an area of ~ 0.9 square kilometers, the coordinates of 78 infected trees are provided along with the images. However, it is important to emphasize that this is not necessarily all the infected trees within the area but rather registrations along a field transect, obtained in a ground survey in 2016. In the absence of available ground truth for the healthy-three class, a *semi-ground truth* data set containing 65 observations of healthy trees was manually collected from the WorldView image (based on the RGB and NDVI composites)[5].

The idea in this thesis is to compare the following:

- Sentinel-2 data *with* red edge bands (S7) and Sentinel-2 data *without* red-edge bands (S4).
- RapidEye-2 data *with* red-edge band (RE5) and RapidEye-2 data *without* red-edge band (RE4).
- WorldView data (WV), S4 and RE4.

In this way, it is possible to evaluate the importance of the Red-edge band and the effect of spatial resolution. The Coruche data sets are summarized in Table 6.2.

Since the red edge bands in the Sentinel-2 image have a spatial resolution of 20 meters, while the blue, green, red and near-infrared bands have a resolution of 10 meters, it is necessary to sharpen the red-edge bands. This is done in the Sentinel Application Platform (SNAP) using the plug-in processor Sen2Res [Brodu, 2017] (*after* doing the atmospheric correction[6]). The idea is to start with

4. For a long period of time, infected trees were not being disposed, allowing for the disease to spread in the area.
5. In fear of selecting infected trees for the "healthy three" class, the set is perhaps somewhat biased towards high-biomass trees in good lighting (yielding a very "healthy appearance" in the RGB and NDVI composites).
6. The atmosperic correction is done using the Sen2Cor [Louis et al., 2016] plug-in processor in SNAP.

Table 6.2: Summary of the data set from Coruche.

		Spatial Resolution		
		0.31 m	5 m	10 m
Spectral Resolution	w/ Red Edge Band	—	RE5	S7
	w/o Red Edge Band	WV	RE4	S4

the 10 m bands, separate the band-dependent reflectance information from the scene geometry (common to all the bands), and use this information to unmix the lower spatial resolution bands while preserving their band-dependent reflectance information [Brodu, 2017].

Also, since the images from the different sensors have different (and very large) sizes, a smaller region within the scene is selected as test image in this thesis (Figure 6.3).

Figure 6.3: Left: RGB composite of the test region of the WorldView[7]image. Right: RGB composites of the test region of the RapidEye-2[8] (top) and the Sentinel-2[9] (bottom) images. The red dots indicate infected trees.

7. © DigitalGlobe, Inc. (2016), provided by European Space Imaging.
8. © (2016) BlackBridge S.àr.l. All rights reserved.
9. © ESA (2016).

Remarks

- The WorldView image is pansharpened. The pansharpening process will introduce both spatial and spectral distortions in the resulting image, and the severity depends on the technique [Palsson et al., 2012]. The pansharpening technique used on the WorldView image is unknown and any distortions it may have caused are a potential sources of error.

- The Sen2Res sharpening operation will, in the same way as in the pansharpening operation, cause distortions in the sharpened bands and is also a potential source of error.

Table 6.3: Sensor information [SIC, n.d.-a, ESA, n.d.-a, SIC, n.d.-b].

Sensor	Band (CWL[10] $\pm \frac{1}{2}$BW[11] [nm], Spatial Resolution Nadir [m])	
RapidEye-2	Blue	(475 ± 35, 5)
	Green	(555 ± 35, 5)
	Red	(657.5 ± 27.5, 5)
	Red Edge	(710 ± 20, 5)
	NIR	(805 ± 45, 5)
Sentinel-2	Coastal Aerosol	(443.9 ± 27, 60)
	Blue	(496.6 ± 98, 10)
	Green	(560.0 ± 45, 10)
	Red	(664.5 ± 38, 10)
	Red Edge 1	(703.9 ± 19, 20)
	Red Edge 2	(740.2 ± 18, 20)
	Red Edge 3	(782.5 ± 28, 20)
	NIR	(835.1 ± 145, 10)
	Red Edge 4	(864.8 ± 33, 20)
	Water Vapour	(945.0 ± 26, 60)
	SWIR 1 (Cirrus)	(1373.5 ± 75, 60)
	SWIR 2	(1613.7 ± 143, 20)
	SWIR 3	(2202.4 ± 242, 20)
WorldView	Blue	(480 ± 30, 0.31)
	Green	(545 ± 35, 0.31)
	Red	(672.5 ± 17.5, 0.31)
	Red Edge	(850 ± 70, 0.31)

6.3 Experiment Design

All the supervised methods (all methods except the 2D-Corr-NLS) are evaluated using ROC curves based on their CMR detection results. Since the 2D-Corr-NLS method requires an image as input (it uses the spatial information to detect endmembers) and returns an unknown (determined by a clustering validity index) number of abundance fraction maps, a performance evaluation based on ROC curves is ill-suited[12]. The 2D-Corr-NLS method is thus treated separately and the performance is evaluated qualitatively by visual inspection and quantitatively for some selected outputs.

The experiment is divided in two parts; one for each data set (HYDICE and

10. Central Wavelength.
11. Bandwidth.
12. The identity and order of the abundance maps are impossible to know beforehand.

Coruche), and is organized as follows:

1. HYDICE images applied to
   (a) Supervised methods
   (b) Unsupervised method
2. Coruche images applied to
   (a) Supervised methods
   (b) Unsupervised method

In order to compare the supervised methods on the same basis for the two data sets, it is necessary to formulate some conditions:

- The amount and distribution (target/background) of labeled data points can not exceed 78 target points and 65 background points.
- The number of samples used to determine the threshold(s) in the Monte Carlo simulations should be the same in all experiments (to get the same accuracy), and is set to $S = 1000$.
- All test statistics should reduce the dimension of the data points to *one* dimension.

As mentioned in the introduction, an operational PWN detection system is assumed to be a twofold system with a pre-processing step to extract forest pixels (as to make the detection problem a two class problem, discriminating between healthy trees and infected trees). In this thesis, this is done by a simple NDVI threshold operation on the Coruche images[13]. To simulate this operation for the HYDICE data set, two classes are selected for the experiment. Because of their similar spectral signatures, trees and grass are chosen to represent the target and the background class, respectively. In practice, this is done by masking out pixels with abundances above 0.5[14] according to the fractional abundance maps of trees and grass (Figure 6.2). This choice implies that the target (background) class is a mix of pure and mixed target (background) pixels. This choice is made to complicate the problem and simulate more challenging

13. The thresholds are selected manually for each image in order to remove some of the non-forest pixels.
14. This goes for the original HYDICE image with 2 m spatial resolution. For the 6 m an 10 m spatial resolution images, the abundances must be $> 0.5 * \frac{1}{3^2}$ and $> 0.5 * \frac{1}{5^2}$, respectively.

conditions, that somewhat resemble the challenge in PWN detection.

6.3.1 Cross-Validation

Since the amount of labeled data is restricted to a maximum of 78 target samples and 65 background samples, a *Monte Carlo cross-validation* (or *repeated random sub-sampling validation*) procedure is chosen to evaluate the accuracy of each method's performance[15].

The idea is to repeatedly partition the labeled data points into a training and a test set and, for each split, compute the projection vector ($\boldsymbol{w}$) from the training set and validate the performance using the test set. The results from all the splits are then used to provide statistics (mean and variance[16]) of the detection and false alarm rate. Since detection performance depends on the threshold(s), the detector statistic should be evaluated at various thresholds, resulting in a ROC curve for each split, as illustrated in Figure 6.4. Then, the final performance measure of each method will be a mean ROC curve, together with accompanying information about the AUC. In addition, the upper and lower bounds of the ROC, corresponding to plus/minus one standard deviation in detection and false alarm rate are indicated by shading a region around the curve.

For the Coruche data, this will be the final performance measure. For the HYDICE data on the other hand, the 78+65 labeled points are randomly drawn from the original image *ten* times (with replacement[17]), and for each draw, the procedure described above is performed. In this way, by utilizing more of the ground truth, a more robust performance measure is obtained.

The ROC curves from the ten subsamples will produce the final mean ROC curve and the variance in detection rate and false alarm rate for the sub-samples are pooled.

In order to get enough data points to generate the ROC curves, the training sample/test sample ratio is set to 60/40. Also, for the methods that require

15. The *leave-one-out cross-validation* is disregarded, since it is undesirable to produce a ROC curve based on one test point. Further, because of the limited amount of training data (78 target samples + 65 background samples), it is not desirable to divide these into multiple subsamples as in a *K-fold cross-validation* procedure.
16. The variance is expected to be underestimated as no correction for the correlation between the splits is applied [Nadeau and Bengio, 2000]. Nevertheless, the values are still valid for comparison of the variability between the methods.
17. Because of the large amount of pixels to draw from, this is not expected to give much correlation between the sub-samples. However, random sampling *without* replacement would have removed this possibility entirely.

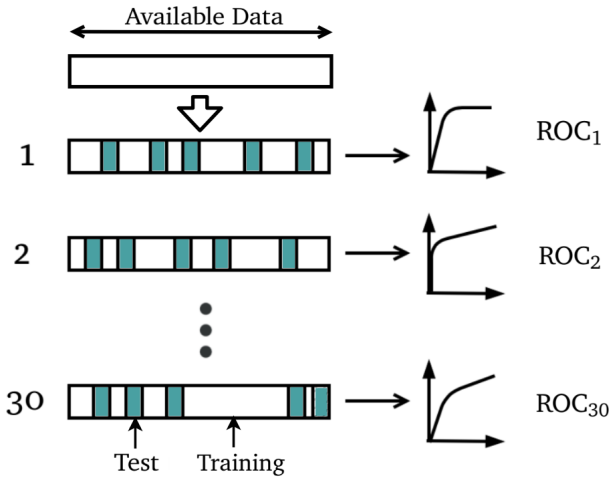

Figure 6.4: Monte Carlo cross-validation: The available data is repeatedly randomly divided into test samples (white) and training samples (green). Note: the random sampling is done such that the target/background ratio is equal for all splits. Modified from [Remesan and Mathew, 2014].

hyperparameters, the entire process is done for a range of parameter values to obtain the optimal value. This method is chosen because of the few labeled data points, in order to avoid splitting the training set further into training and validation sets.

Remark

- A disadvantage of the Monte Carlo cross-validation is that some samples may never appear in the test set while other smaples may never appear in the training set [Dubitzky et al., 2007].

6.4 Code

All algorithms, except the SLDA, have been implemented in MATLAB. For the SLDA algorithm, the `slda.m` code from the SpasSM MATLAB toolbox [Sjöstrand et al., 2012], developed at the Technical University of Denmark, is used.

7

Results and Discussion

In this chapter, the CMR detection results obtained by applying the different methods to the data sets are presented and discussed. All methods except for the 2D-Corr-NLS are evaluated on equal terms and presented with statistics on the performance. The 2D-Corr-NLS method is evaluated qualitatively by visual inspection and quantitatively for some selected outputs (abundance maps). The chapter is organized in the same way as the experiment, and the results are presented and discussed in the same order as the experiments were conducted.

7.1 Pt 1: HYDICE

Supervised Methods

For each HYDICE image, *ten* sub-samples of size 143 (78 target pixels + 65 background pixels) were randomly drawn.

For each of the ten sub-samples, the *mean* ROC curve (and variances of the detection rate and false alarm rate for each threshold) was estimated in a 30-split Monte Carlo cross-valedation procedure (illustrated in Figure 6.4). That is, for each of the 30 splits, the sub-samples were randomly divided into a training set (40 %) and a test set (60 %), and the projection vector ($\boldsymbol{w}$) was estimated from the training data. Based on $S = 1000$ samples generated from

a Gaussian distribution (specified by the sample mean and sample covariance of the target samples in the training data), the detector output distribution of the target class was estimated. From this distribution, 100 CMR thresholds from $\alpha = 0.01$ to $\alpha = 0.99$ were determined (Figure 5.1). For each threshold, the detection rate and false alarm rate of the test set were calculated, resulting in a ROC curve.

Thus, each sub-sample produced 30 ROC curves, from which one *mean* ROC curve was calculated (in addition to the variance of detection/false alarm rate at each threshold). In total, this gave ten mean ROC curves (for each image), which were used to compute the final mean ROC curves (with pooled variance form the 10 sub-samples), presented in this section.

Before the results are presented, some comments and considerations about the choices related to each method are made.

7.1.1 FLDA

The FLDA does not require any hyperparameters. However, in order to make the inversion of the within-class covariance matrix numerically stable, the dimensionality of the *hyperspectral* HYDICE images must be reduced. This is done via principal component analysis. To decide the optimal dimensionality, the mean AUC value (based on 300 ROC curves from 10 different sub-samples) is considered for different choices of dimensions for each image. The results are presented in Appendix C.1, and indicate that a dimensionality of four is ideal for all three hyperspectral images. The final results based on the optimal dimensions are presented in Figure 7.1.

7.1.2 SLDA

The `slda.m` function [Sjöstrand et al., 2012] requires two hyperparameters; the regularization parameter for the ℓ_2-penalty and *either* a *upper boundary* for regularization parameter for the ℓ_1-penalty *or* a parameter STOP defining the number of non-zero elements in $\boldsymbol{w}$. The STOP parameter is selected in this thesis. The optimal values for these parameters are found by evaluating the AUC (based on 300 ROC curves from 10 different sub-samples) for a range of values for each image (Appendix C.2). The results based on the optimal parameters are presented in Figure 7.2.

7.1.3 ICEM

For *hyperspectral* data, the ICEM algorithm is performed on a subset of orthogonally transformed data, e.g. a subset of the data in MNF space [Nielsen, 2001]. Based on the eigenvalues of the MNF transformed data, the dimension of the HYDICE_h2, HYDICE_h6 and HYDICE_h10 images were reduced to *three* dimensions, explaining 99.0%, 99.2% and 99.3% of the total variance, respectively.

The ROC curves of the ICEM method proposed in [Nielsen, 2001] (ICEM1) and the alternative ICEM algorithm that weights the background (ICEM2) are presented in Figure 7.3 and Figure 7.4, respectively.

7.1.4 MTMF

The MTMF algorithm requires two hyperparameters as input; the steepness a and the center position b of the sigmoid. a is set to one[1] and b is found by evaluating the mean AUC (based on 300 ROC curves from 10 different sub-samples) for different values of b. The search interval includes some values within the MT range of the training samples and *one* value greater than the maximum MT value (to measure the mean AUC based on pure MF, $x_{MTMF} = x_{MF}$, for reference). The results of the parameter search (Appendix C.3), indicates that information from the MT values improves the performance on all images except one (the performance on the HYDICE_h2 image is best when *all* MT scores get a sigmoid response of one). However, the improvements are quite small, as is expected in a two class-problem (discussed in Section 4.3.3). The only possible false positives are impure grass pixels containing another distinct component, making the pixels anomalies from the background *and* infeasible mixtures between target and background.

Further, based on the eigenvalues of the MNF transformed data, the dimension of *all* the HYDICE images were reduced to *three* dimensions, explaining 99% and $\sim$ 100% of the total variance for the hyperspectral and the multispectral data, respectively.

Also, since the experiment does *not* resemble a low target probability environment, the covariance matrix in the MF step is estimated from known background points in the training data (and not from all pixels). Furthermore, the MNF transformed data is "background-mean corrected"[2], in order to center

1. Because of the range of the interesting MT scores, this is sufficient to make the Sigmoid behave like a step function.
2. Based on the known background points in the training data.

the background class at the origin.

The final results based on the optimal values of b are presented in 7.5.

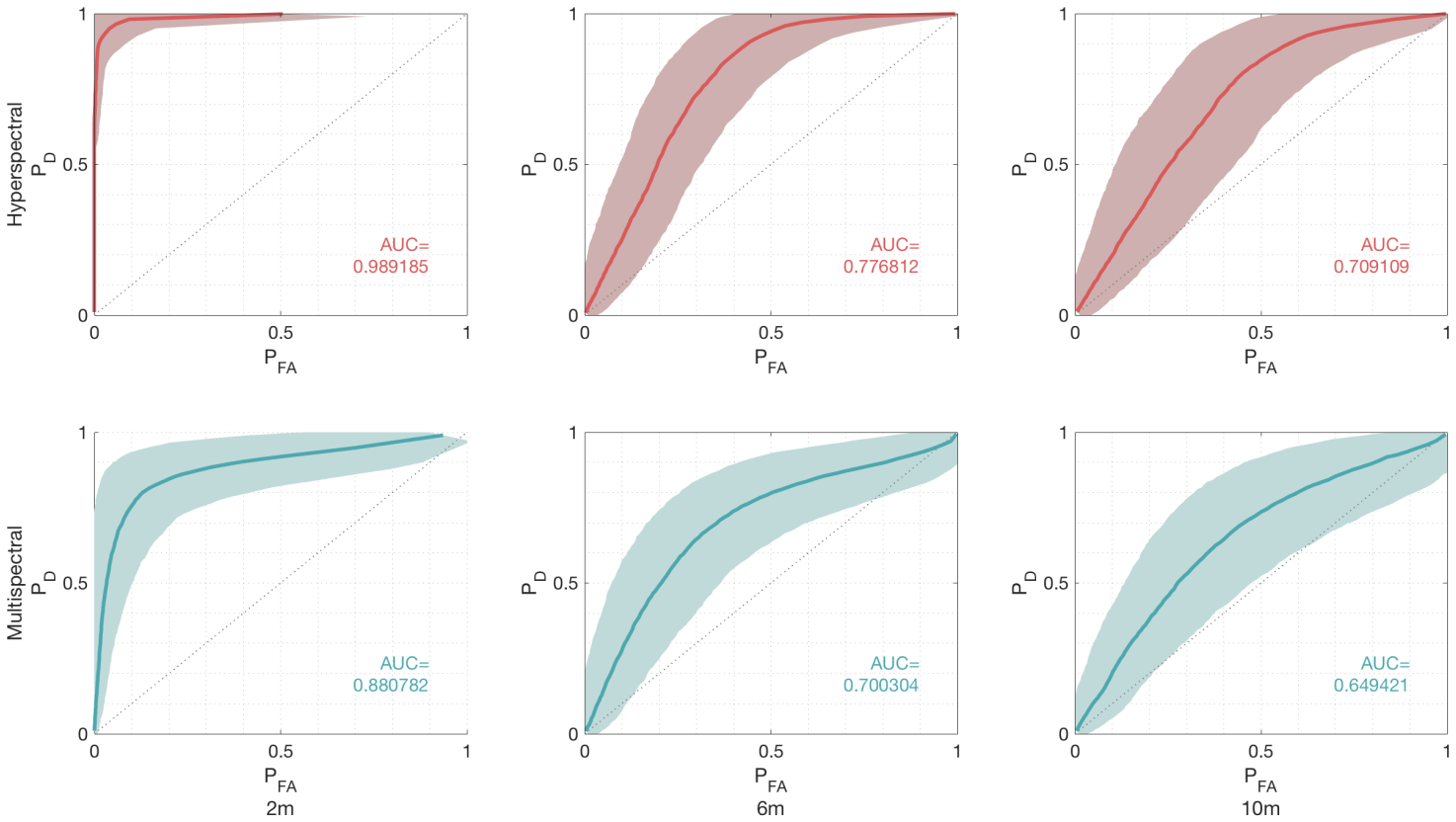

Figure 7.1: ROC curves based on the FLDA algorithm applied to the six HYDICE data sets with the optimal parameters (see Appendix C.1). Top: (from left) HYDICE_h2, HYDICE_h6, HYDICE_h10. Bottom: HYDICE_m2, HYDICE_m6, HYDICE_m10. The shaded regions indicate the plus/minus one standard deviation of P_D and P_{FA}.

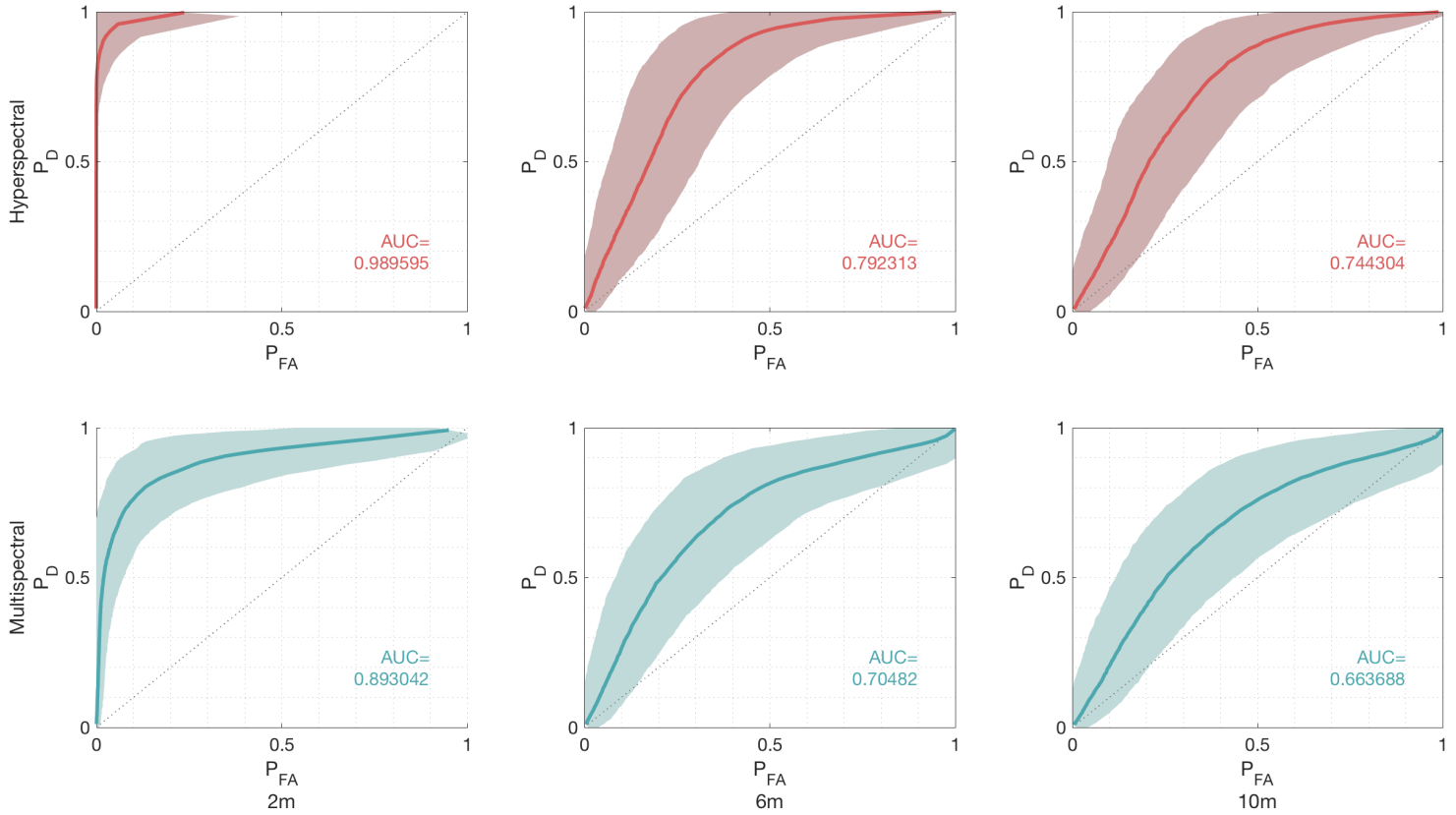

Figure 7.2: ROC curves based on the SLDA algorithm applied to the six HYDICE data sets with the optimal parameters (see Appendix C.2). Top: (from left) HYDICE_h2, HYDICE_h6, HYDICE_h10. Bottom: HYDICE_m2, HYDICE_m6, HYDICE_m10. The shaded regions indicates the plus/minus one standard deviation of P_D and P_{FA}.

.

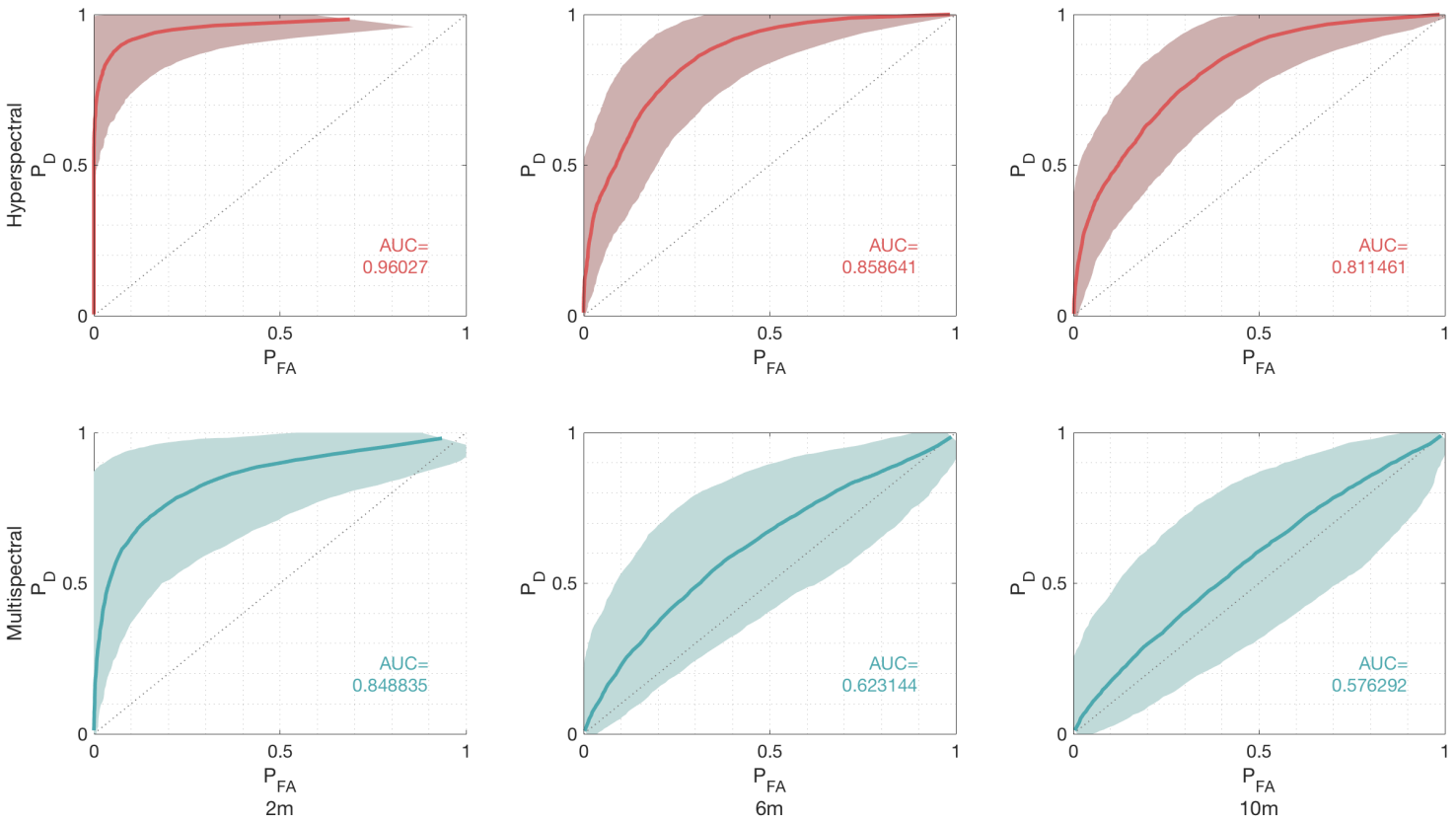

Figure 7.3: ROC curves based on the target weighted ICEM algorithm (ICEM1) applied to the six HYDICE data sets. Top: (from left) HYDICE_h2, HYDICE_h6, HYDICE_h10. Bottom: HYDICE_m2, HYDICE_m6, HYDICE_m10. The shaded regions indicate the plus/minus one standard deviation of P_D and P_{FA}.

.

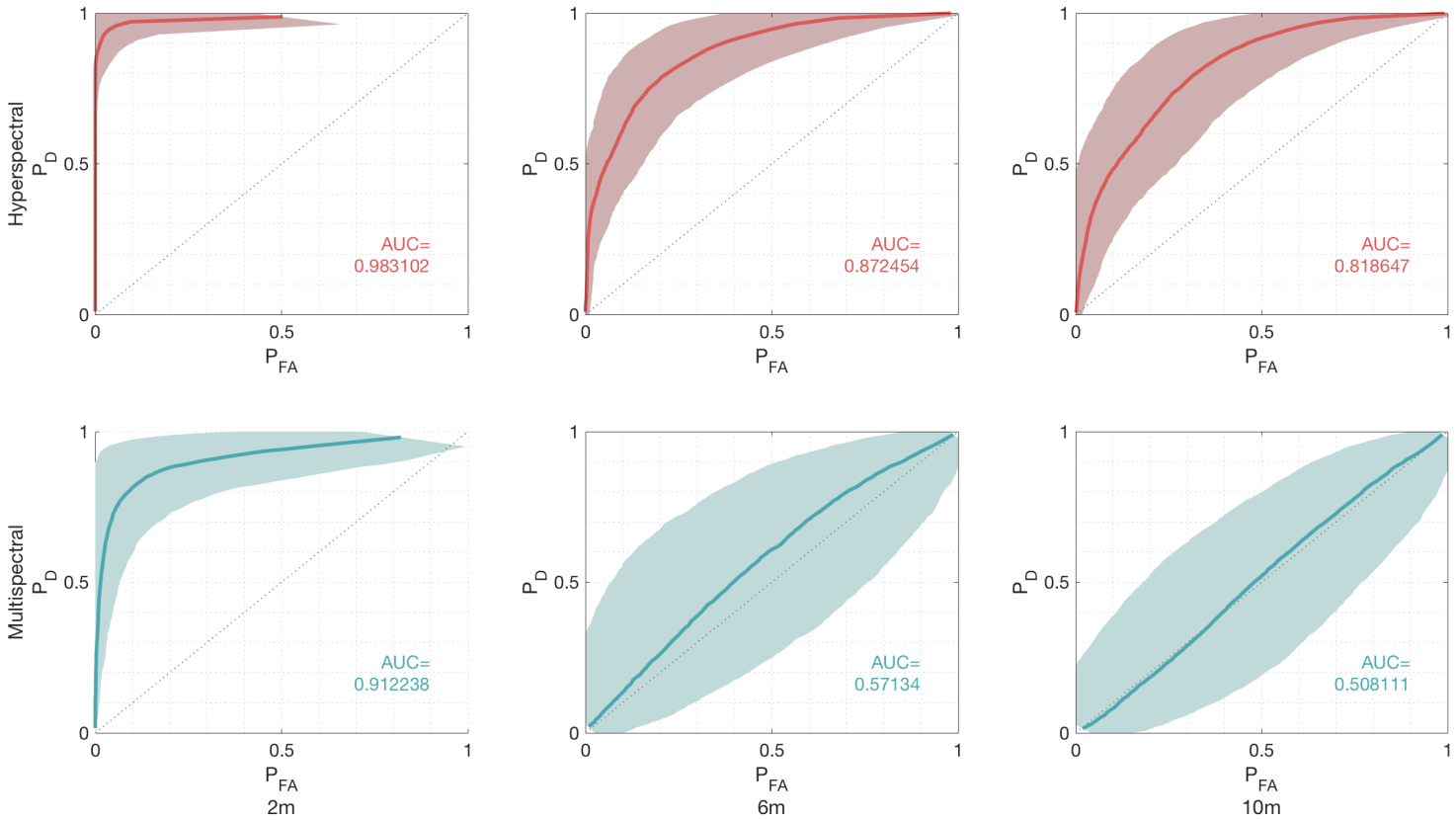

Figure 7.4: ROC curves based on the background weighted ICEM algorithm (ICEM2) applied to the six HYDICE data sets. Top: (from left) HYDICE_h2, HYDICE_h6, HYDICE_h10. Bottom: HYDICE_m2, HYDICE_m6, HYDICE_m10. The shaded regions indicate the plus/minus one standard deviation of P_D and P_{FA}.

.

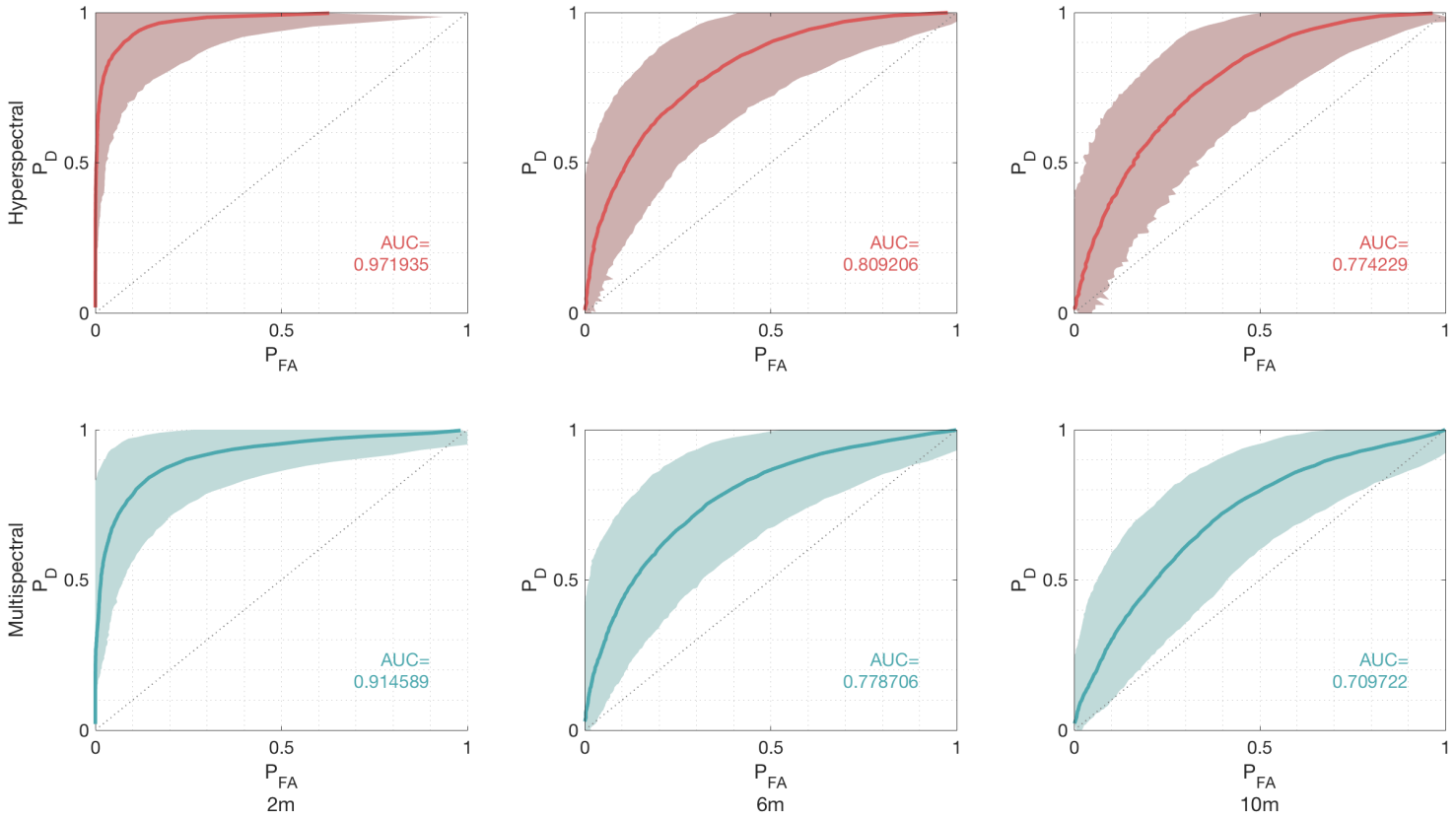

Figure 7.5: ROC curves based on the MTMF algorithm applied to the six HYDICE data sets with the optimal parameters (see Appendix C.3). Top: (from left) HYDICE_h2, HYDICE_h6, HYDICE_h10. Bottom: HYDICE_m2, HYDICE_m6, HYDICE_m10. The shaded regions indicate the plus/minus one standard deviation of P_D and P_{FA}.

.

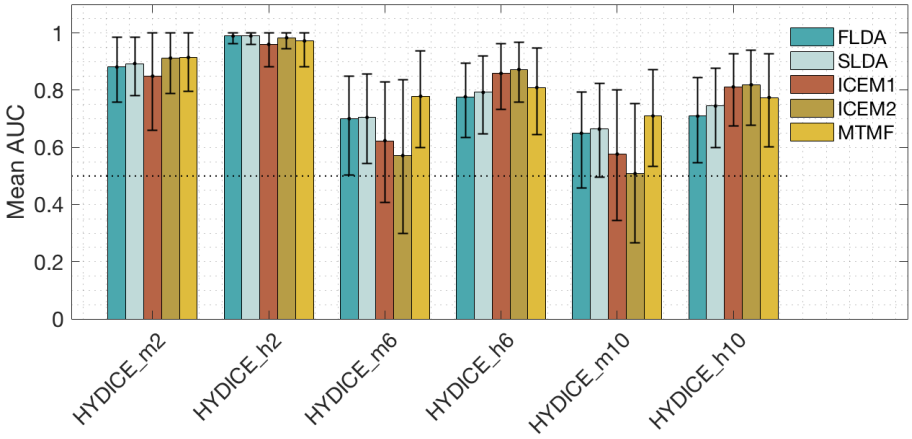

Figure 7.6: Summary of the mean AUC values for the different methods applied to all HYDICE images. The errorbars indicate the AUC corresponding to the ROC curves for plus/minus one standard deviation in detection rate and false alarm rate.

7.1.5 Discussion

The AUC values corresponding to the mean ROC curves in Figure 7.1-7.5 are summarized in the bar plot in Figure 7.6. The error bars represent the AUC values of the ROC curves for plus/minus one standard deviation in detection rate and false alarm rate. That is, the area under the upper (AUC^+) and lower (AUC^-) boundaries of the shaded regions in the ROC curves in 7.1-7.5. The difference $AUC^+ - AUC^-$ will be referred to as the variability of the performance.

The initial objective of this thesis was to explore the boundaries of the data, in terms of spatial and spectral resolution, for a selected set of methods in a detection problem. From the plot in Figure 7.6, three obvious observations are made on the general performance:

1. The hyperspectral data yields better results than the multispectral data for each method.

2. The performance of each method is better for the 2 m spatial resolution data than that of the 6 m spatial resolution data, which again is better than that of the 10 m spatial resolution data.

3. Overall, the variability of the performance decreases with increased spectral and spatial resolution.

Discussion of Observation 1

A closer look at observation 1 reveals that the effect of spectral resolution (difference in performance on hyperspectral and multispectral data) tends to slightly *decrease* with decreasing spatial resolution for the FLDA, SLDA and MTMF method. For the ICEM methods (ICEM1 and ICEM2), on the other hand, a decrease in spatial resolution from two to six or ten meters shows a considerable *increase* in the effect of spectral resolution: The AUC based on the multispectral data drops significantly when the spatial resolution is reduced to six and ten meters. This is illustrated in Figure 7.7 where the difference in mean AUC value between hyperspectral and multispectral data is displayed for all methods, at different spatial resolutions.

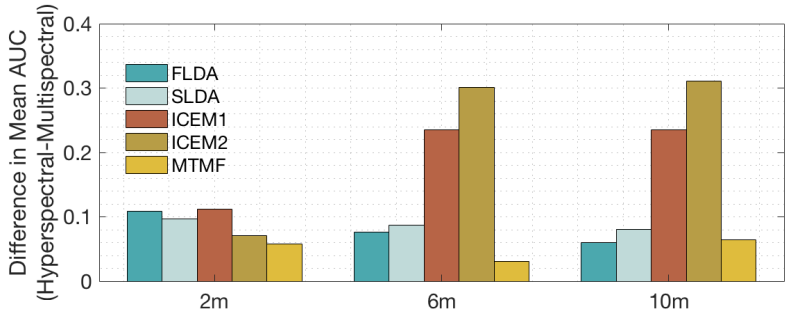

Figure 7.7: This graph displays the effect of spectral resolution: For each spatial resolution, the mean AUC based on the multispectral data is subtracted from that of the hyperspectral data.

An important factor that makes the ICEM methods stand out from the other supervised methods is that they only require information about the target. While the other supervised methods require information on both the target and the background[3], the ICEM methods only need information about the target pixels to run. This information is used to initialize the statistic, $\hat{\boldsymbol{w}}^{(1)}$. Then, the final statistic $\hat{\boldsymbol{w}}$ is calculated using the covariance estimate based on on *all* pixels, weighted[4] by the projection onto $\hat{\boldsymbol{w}}^{(1)}$. The significant drop in AUC may thus indicate that the multispectral information is *not* enough to weigh the correct pixels when the spatial resolution decreases, and that the ICEM methods require hyperspecral information to operate on images with coarse spatial resolution.

3. In a low-target probability scenario, the MTMF method would *not* have required information about the background (the background covariance would have been estimated from the entire scene). However, in this experiment, this is not the case and information about the background is required.
4. ICEM1 weights the target while ICEM2 weights the background.

Discussion of Observation 2

Moving over to observation 2, this effect is more clearly illustrated in Figure 7.8, where the mean AUC values for each method applied to the different images are plotted for the three spatial resolutions.

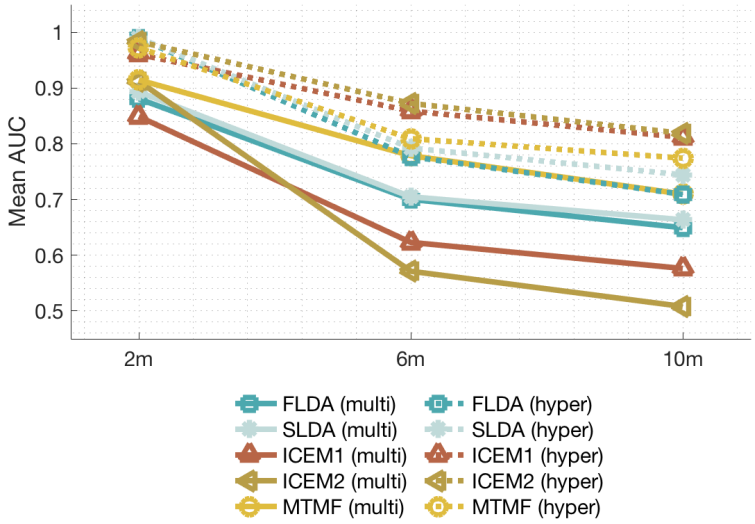

Figure 7.8: This graph displays the effect of spatial resolution on the mean AUC: For each method, for each spectral resolution, the mean AUC is plotted for different spatial resolutions.

Increasing the pixel size (reducing the spatial resolution) of an image is known to produce different effects. The most obvious consequence is that the amount of mixed pixels increases, making the discrimination between the classes a tougher problem [Campbell, 2011]. Another effect, known from many previous studies [Hsieh and Lee, 2000, Latty and Hoffer, 1981, Cushnie, 1987], is that the within-class variance may decrease. More detailed ground information (higher spatial resolution) may *not* automatically imply a higher overall classification accuracy. This is connected to the fact that an increase in spatial resolution usually leads to an increase in within-class variability, which is associated with a decrease in classification accuracy [Hsieh and Lee, 2000]. Accuracies of land-cover classification where the internal variability is high (such as forest covers) may thus improve by coarsening the spatial resolution and thereby decreasing the variance [Treitz et al., 1992, Cushnie, 1987].

The HYDICE image scene is composed of parcels of trees with grass in between. In the high resolution image, most of the pixels are *pure*, as opposed to the mid to low resolution images, where the number of mixed pixels dominates. From the results in Figure 7.8 it appears that the effect of increased amount

mixed pixels dominates and that any possible effect of decreased within-class variance is masked.

Discussion of Observation 3

Observation 3 is visualized in Figure 7.9. From this plot, it is clear that the performance variability decreases with increased spectral resolution (all dashed lines are located below the solid lines) and that *generally*, a decrease in spatial resolution yields an increase in variability. However, from 6 m to 10 m, the increase stagnates and for the multispectral data, the ICEM2 method even show a *decrease* in variability (the FLDA also shows a slight decrease).

Further, it is worth noting that the ICEM methods have the highest performance variability of all methods when applied to multispectral data, and among the lowest variabilities when applied to hyperspectral data. This supports the hypothesis that the information in the low (spatial) resolution multispectral images is insufficient to weight the correct pixels in the ICEM algorithms.

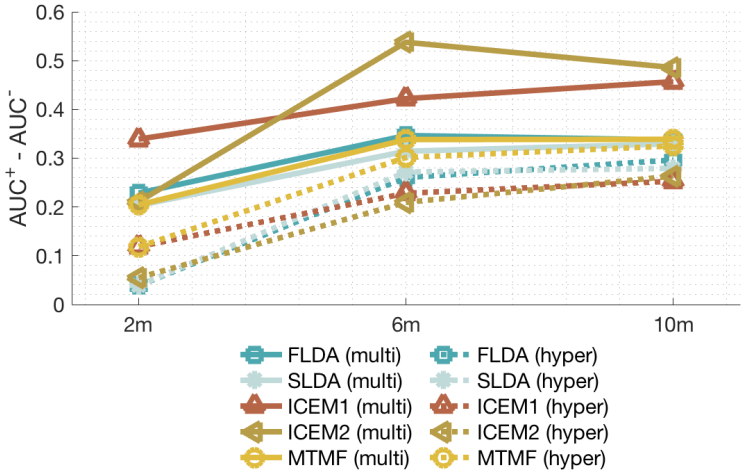

Figure 7.9: This graph displays the effect of spatial resolution on the perfomance variability: For each method, for each spectral resolution, the variability is plotted for different spatial resolutions.

Beyond the three general observations discussed this far, based on the theory of the methods, it is interesting to do a closer analysis of the differences between:

- The MTMF method and the ICEM2 method.
- The ICEM1 method and the ICEM2 method.

- The FLDA method and SLDA method.

Comparison: MTMF and ICEM2

A comparison of the MTMF and the ICEM2 method is interesting since the only difference between the MF step in MTMF and the ICEM2 method is the covariance estimate in

$$\boldsymbol{w} = \frac{\boldsymbol{\Sigma}^{-1}\boldsymbol{d}}{\boldsymbol{d}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{d}}, \tag{7.1}$$

where $\boldsymbol{w}$ is the projection vector and $\boldsymbol{d}$ is (the representative of) the target signature. The covariance estimate in MTMF is based on *known* background pixels while the covariance estimate in ICEM2 is based on *all* pixels, weighted by *one minus* the projection onto

$$\boldsymbol{w} = \frac{\hat{\boldsymbol{\Sigma}}_t^{-1}\boldsymbol{d}}{\boldsymbol{d}^T\hat{\boldsymbol{\Sigma}}_t^{-1}\boldsymbol{d}}, \tag{7.2}$$

where $\hat{\boldsymbol{\Sigma}}_t$ is the covariance estimate based on known target pixels. That is, the covariance in MTMF is based on known background pixels and the covariance estimate in ICEM2 is based on all pixels, where background pixels (as according to the projection onto (7.2)) are given more weight.

For the hyperspectral images, the ICEM2 method yields better results than the MTMF method for all spatial resolutions: the mean AUC is higher *and* the variability is smaller. This means that it is better to estimate the covariance matrix based on pixels *not* resembling the target instead of known background pixels. A possible explanation to this behavior could be that some of the labeled background pixels may be anomalies that actually do resemble the target. Basing the covariance estimate on pixels not resembling the target may thus remove the effect of these anomalies and generate a more "target-free" version of the covariance matrix than using the labeled points.

For the multispectral images, on the other hand, the MTMF yields better results than the ICEM2. This is probably related to the previous discussions about the ICEM methods not being able to weight the correct pixels in coarse multispectral images.

Comparison: ICEM1 and ICEM2

Disregarding HYDICE_m6 and HYDICE_m10[5], the ICEM2 method shows the superior performance on all images. However, the ICEM1 method follows closely. As discussed in Section 4.3.2, basing the covariance estimate on target pixels may be just as good as basing it on background pixels. The result will depend on the spread of the two classes. However, when the covariance structures are similar, the differences are not expected to be very large. Plotting the covariance matrices of the target (tree) and the background (grass) classes (Figure 7.10-7.11) shows that the covariance structures of the two classes look similar for all images, which is consistent with the discussion.

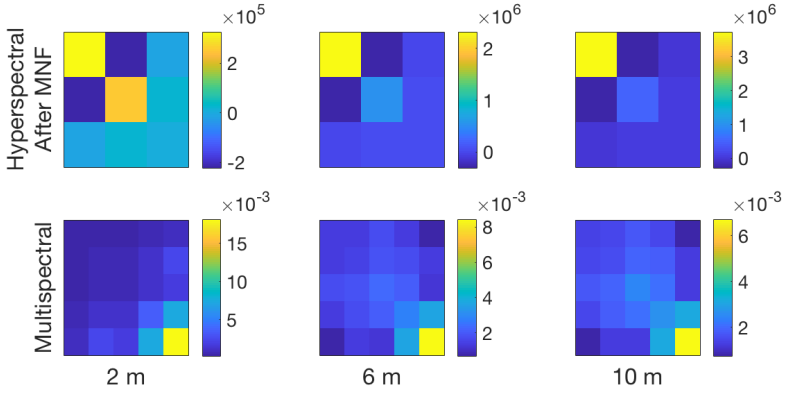

Figure 7.10: Covariance matrices of the target class of the different HYDICE images.

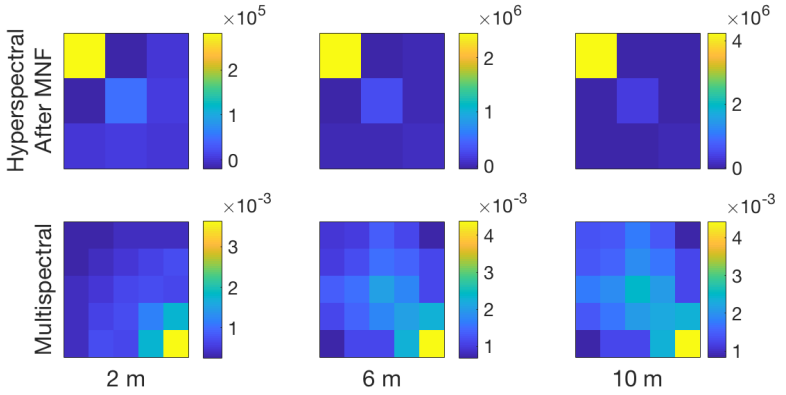

Figure 7.11: Covariance matrices of the background class of the different HYDICE images.

5. The application of both ICEM1 and ICEM2 to HYDICE_m6 and HYDICE_m10 yields very poor results and are irrelevant as there is hardly any discrimination.

Comparison: FLDA and SLDA

The two classical feature extraction methods, FLDA and SLDA, have quite similar performances (mean values and variability) for all data sets, SLDA being somewhat superior to the FLDA. The two methods show the highest AUC value (and smallest variablilty) of all methods when applied to the HYDICE_h2 data, but other than that they are at best mediocre (compared to the other methods).

As mentioned in Section 4.3.3, there are some concerns about the shift-difference method for noise estimation in the MNF transformation. Because of the way the 6 m and 10 m spatial resolution images are obtained (averaging over 3^2 and 5^2 pixels of the original 2 m image), the signal-to-noise ratio (SNR) is increased from that of the 2 m image. The same effect is experienced for a sensor with coarser spatial resolution: the instantaneous field of view is enlarged, more photons from each ground cell is received at the sensor and the SNR increases. In this way, the pixels, the and difference between pixels, of coarser spatial resolution images will mainly contain signal. However, by looking at the performance of the MTMF method (which applies the MNF transformation to all the images), the concerns about the MNF transform seems to be unnecessary.

All in all, there are two methods that stand out in these results, and that is the MTMF method and the ICEM2 method:

- The MTMF method yields the highest AUC values for all the multispectral images. Further, the AUC corresponding to the MTMF method applied to the *multispectral* HYDICE_m6 data actually *tangents* the AUCs corresponding to the FLDA method applied to the *hyperspectral* HYDICE_h6 data. The same is seen for the 10 m data. However, the variability of the MTMF method is greater in both cases.

- The ICEM2 method, despite its poor performance on the coarse resolution multispectral images, is the (overall) best method for the hyperspectral data. It shows the highest AUC values for for all methods when applied to the hyperspectral HYDICE_h6 and HYDICE_h10 images and it almost tangents SLDA and FLDA when applied to the HYDICE_h2 image.

Unupervised Method

7.1.6 2D-Corr-NLS

Since the 2D-Corr-NLS method utilizes the spatial information to automatically detect endmembers, it requires an image as input. Additionally, the size of the analysis window (the mask searching for single-source-zones) and the threshold determining whether or not a neighborhood is a single-source-zone[6] must be provided. Further, since the method is designed for multispectral imagery, no instructions on how to handle hyperspectral imagery is given. In this thesis, the dimension of the hyperspectral data is reduced to three, after applying the MNF transformation (as for the ICEM and MTMF methods).

If the analysis window is too large, it will not be able to detect the target as a single-source-zone. Because the target is distributed in small parcels within the image scene, the window size is set to 3×3 pixels.

In [Boardman and Kruse, 2011], the threshold is set to 0.992. However, in this thesis, the threshold is set much lower. Setting the the threshold very high (close to one) imposes strict purity requirements for the endmembers, allowing for little internal variation. Setting the threshold very low, on the other hand, can make any neighborhood qualify as a single-source zone. The thresholds for the HYDICE images were found by trial and error and set to 0.40 for HDICE_m2, HYDICE_h2, HYDICE_h6 and HYDICE_h10, 0.35 for HYDICE_m6 and 0.45 for HYDICE_m10.

One reason these thresholds have to be set so low is probably because vegetation is known to have high internal spectral variability.

A consequence of setting the thresholds low is that the method detects many unwanted endmembers and endmembers that are hard to identify. A set of identifiable abundance maps are thus selected (for each HYDICE image) to be presented in the following plots. The intensity in the abundance maps indicate the abundance of the specific endmember in each pixel (black: 0 % abundance, white: 100 % abundance).

6. The smallest element in the neighborhood cross-correlation vector must be greater than the threshold for the neighborhood to qualify as a single-source-zone.

Figure 7.12: RGB composite of the input image (2 m spatial resolution). The site consists of grass (light green), trees (dark green) and soil (brown/white).

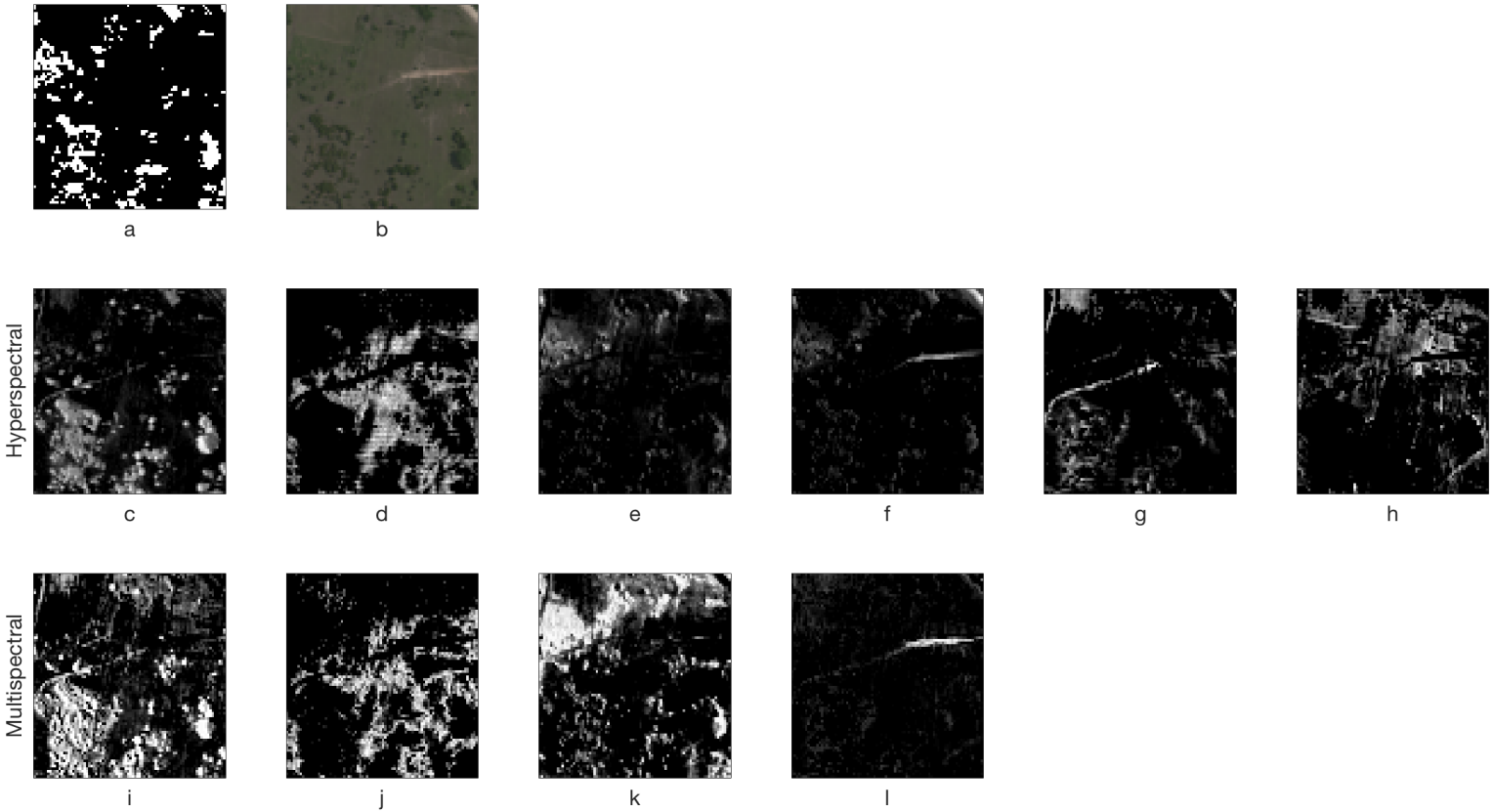

Figure 7.13: Resulting abundance maps from the 2D-Corr-NLS method applied to the HYDICE_h2 (c-h) and HYDICE_m2 (i-l) images. (a) represent the target ground truth (white = target) and (b) is the RGB composite of the input image. Note the striping in c)-h), which failed to be removed in the MNF process.

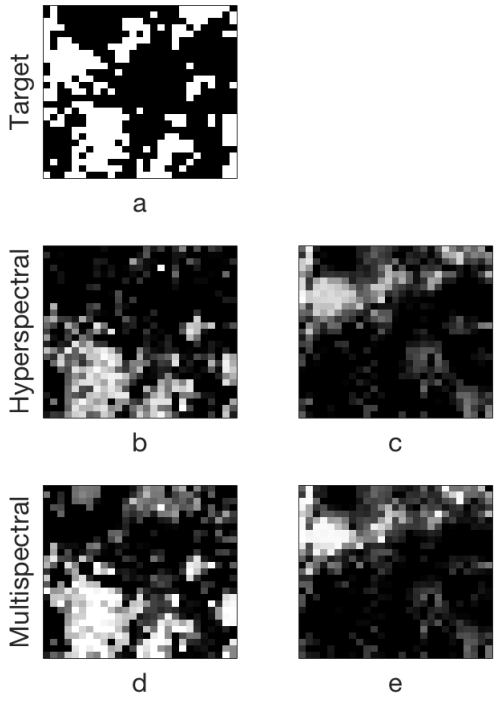

Figure 7.14: Two selected abundance maps representing the target class, obtained by applying the 2D-Corr-NLS method to the HYDICE_h6 (b-c) and HYDICE_m6 (d-e) images. (a) represent the target ground truth (white = target).

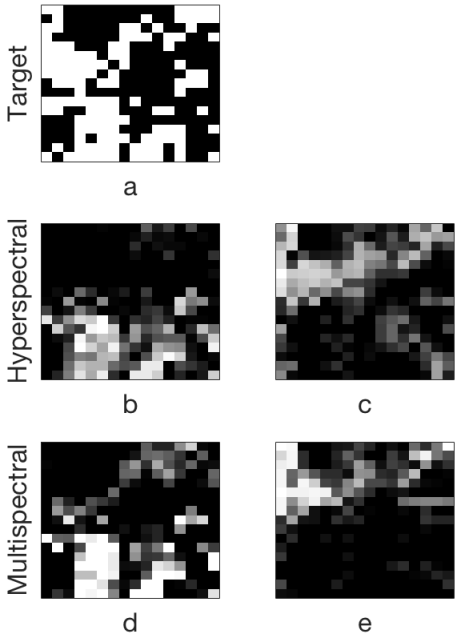

Figure 7.15: Two selected abundance maps representing the target class, obtained by applying the 2D-Corr-NLS method to the HYDICE_h10 (b-c) and HYDICE_m10 (d-e) images. (a) represent the target ground truth (white = target).

7.1.7 Discussion

The 2D-Corr-NLS algorithm was applied to the six HYDICE images, with varying results (presented in Figure 7.13-7.15).

For the HYDICE_h2 image and the HYDICE_m2 image, six and four abundance maps (Figure 7.13) were selected among the outputs. Based on the ground truth, there are *three* types of materials within the test scene: trees, grass and dirt. From the plots in Figure 7.13 it is clear that the 2D-Corr-NLS algorithm was able to detect all three endmembers, however, split into multiple sub-endmembers:

- Figure 7.13 c) and i) represent trees.
- Figure 7.13 d)-e) and j)-k) represent green grass.
- Figure 7.13 f)-h) and l) represent dirt.

The splitting of the endmembers bears witness of spectral variability within the endmembers (as defined by the ground truth).

In order to quantitatively evaluate the method, the two images representing trees (the target) were tested in a simple detector using 100 thresholds from zero to one. This resulted in two ROC curves, presented in Figure 7.16.

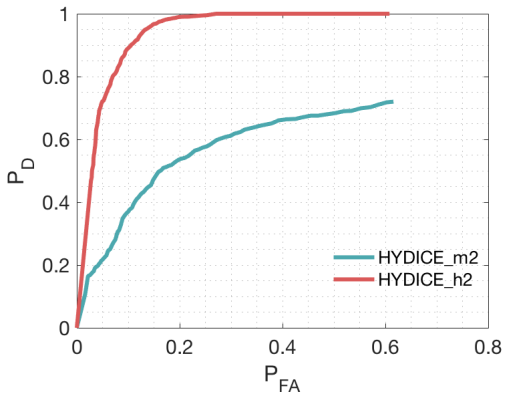

Figure 7.16: Resulting ROC curve from evaluating Figure 7.13 c) (red) and Figure 7.13 i) (green) for thresholds from zero to one.

Based on this plot, it is evident that *all* pixels containing target material had a non-zero value in Figure 7.13 c) (the target endmember detected from the HYDICE_h2 image). A detection rate of one is reached at a false alarm rate of only ~ 0.24 (red ROC curve). In the abundance map in Figure 7.13 i) (the

target endmember detected from the HYDICE_m2 image), on the other hand, some of the target material is not present at all, preventing the green ROC curve to reach a detection rate of one. A closer look at Figure 7.13 reveals that some of the trees are mistaken for grass in the multispectral case (the bright spot in the upper left corner in k) corresponds with the bright spot in the upper left corner of the ground truth image, a)).

The 2D-Corr-NLS algorithm was also applied to the coarser spatial resolution HYDICE images. However, no thresholds were found that would capture the target class (trees) within *one* endmember. This is shown in Figure 7.14-7.15 where two selected abundance maps (output) representing the target class are presented for each input image.

This result may be explained by the fact that the target is more sparsely distributed in the upper part of the test scene, compared to that in the lower part. In this way, when the resolution decreases, the target pixels in the upper part of the image get more influenced by the background. In addition, the background in the upper part of the scene is more green (grass) and the background in the lower part of the scene is more brown (dirt). The differences in background material is also seen in Figure 7.13 e) and g).

These results further supports the discussion in Section 4.3.1 about the algorithm being sensitive to spectral variability in the background *and* variations in the proportions of target and background in each pixel.

7.2 Pt 2: Coruche

For each Coruche image, the 143 known data points (78 target pixels + 65 background pixels) were used to produce a mean ROC curve, based a 30-split Monte Carlo cross-validation procedure (illustrated in Figure 6.4). That is, for each of the 30 splits, the sub-samples were randomly divided into a training set (40 %) and a test set (60 %), and the projection vector ($\boldsymbol{w}$) was estimated from the training data. Based on $S = 1000$ samples generated from a Gaussian distribution (specified by the sample mean and sample covariance of the target samples in the training data), the detector output distribution of the target class was estimated. From this distribution, 100 CMR thresholds from $\alpha = 0.01$ to $\alpha = 0.99$ were determined (Illustrated in Figure 5.1). For each threshold, the detection rate and false alarm rate of the test set were calculated, resulting in a ROC curve.

The mean of these 30 ROC curves are presented below, for each image, for each method. Also, the upper and lower bounds on the ROC curves, corresponding

to plus/minus one standard deviation in detection and false alarm rate, are indicated by shading a region above and below the mean curves.

Before the results are presented, some comments and considerations about the choices related to each method are made.

Supervised Methods

7.2.1 FLDA

The FLDA does not require any hyperparameters. The results are presented in Figure 7.17.

7.2.2 SLDA

The optimal values for the parameters (STOP and γ) are found by evaluating the AUC for a range of values for each image. Based on the results (Appendix C.2), the following choices were made:

- For the WorldView data, the STOP parameter is set to *two* (picking out the red and the near infrared band), and the ℓ_2 parameteris set to *one*.
- For the RapidEye data, the STOP parameter is set to its maximum for both RE4 and RE7 (causing non-sparse solutions) and the ℓ_2 parameter is set to $1e20$.
- For the Sentinel data, the STOP parameter is set to *one* (picking out the red band) for both S4 and S7, thereby making the ℓ_2 parameter insignificant.

The results based on the optimal parameters are presented in Figure 7.18.

7.2.3 ICEM

The ICEM methods do not require any hyperparameters. The results are presented in Figure 7.19 and 7.20.

7.2.4 MTMF

The parameters of the MTMF method are found in the same way as for the HYDICE data, by evaluating the mean AUC for different values of b (Appendix C.3).

The MT values are found to *not* improve the performance on any of the Coruche images. That is, b is set higher than the maximum MT for all images, making the MTMF algorithm reduce to pure MF. A possible explanation to this behavior could be that the internal variability in the target class is too high, as discussed in Section 4.3.3. A high spectral variability in the target class may lead to a lot of target samples having a high MT score. It is thus impossible to remove (set to zero) false positives without also removing true positives. In addition, the MT range varied more between the sub-samples for the Coruche data, compared to that of the HYDICE data, and it was difficult to find *one* good value for b.

Further, based on the eigenvalues of the MNF transformed data, the dimension of the Coruche images were reduced to *two* (WV, RE4, S4) and *three* (RE5, S7) dimensions, explaining > 99% of the total variance.

The final results are presented in 7.21.

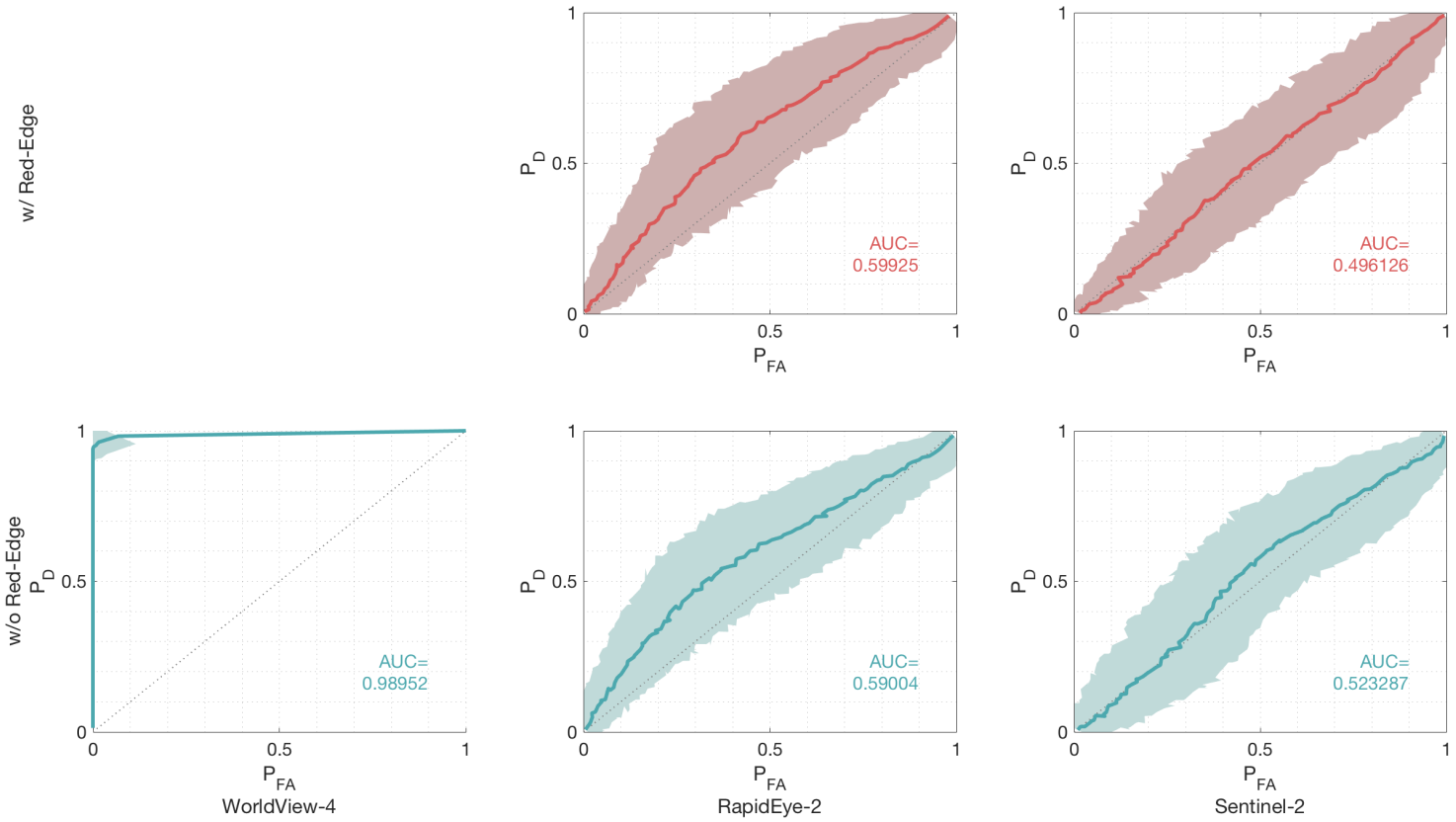

Figure 7.17: ROC curves based on the FLDA algorithm applied to the five Coruche images. Top: (from left) RE5, S7. Bottom: WV, RE4, S4. The shaded regions indicate the plus/minus one standard deviation of P_D and P_{FA}.

.

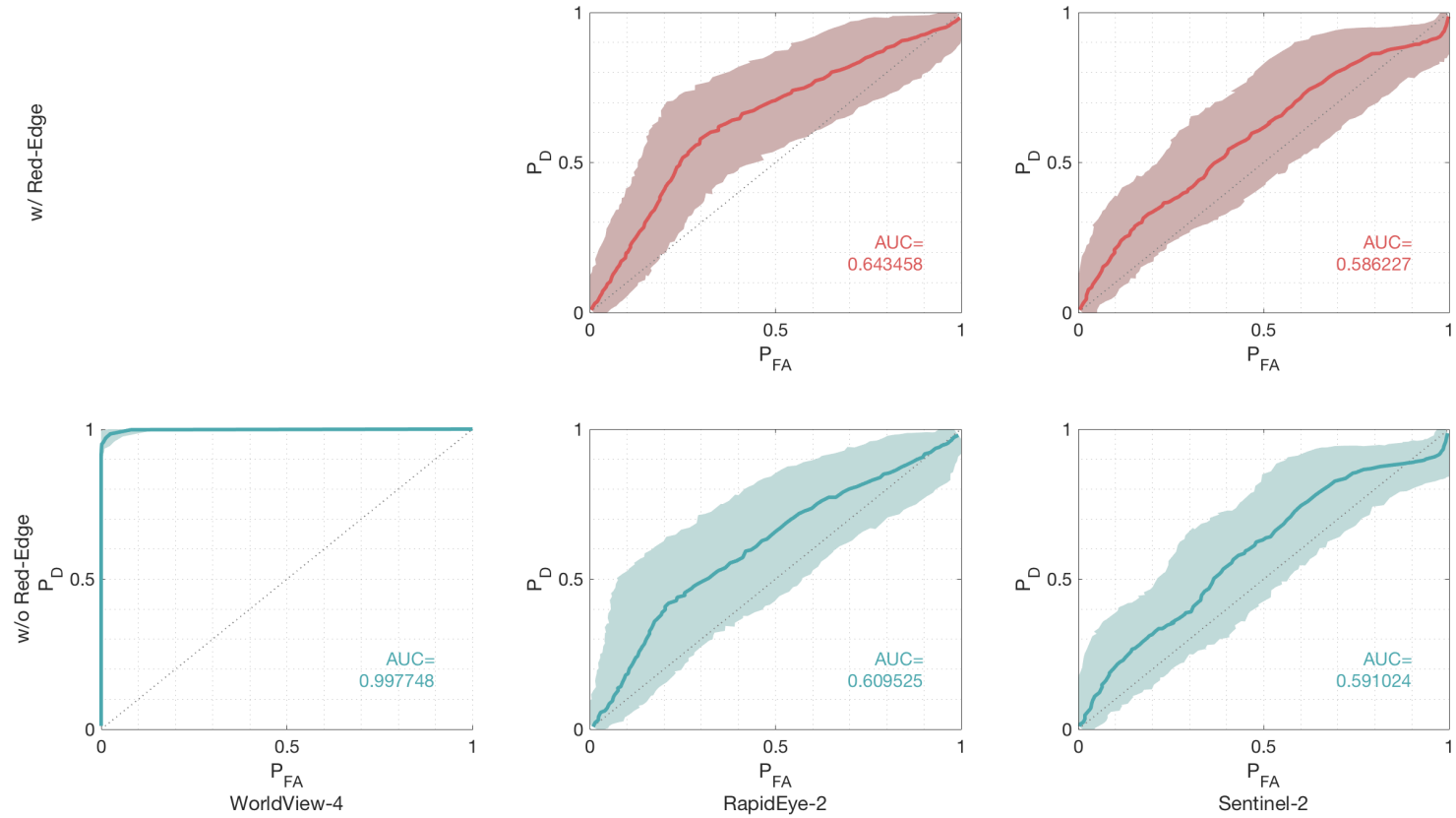

Figure 7.18: ROC curves based on the SLDA algorithm applied to the five Coruche images for optimal parameters (see Appendix C.2). Top: (from left) RE_5, S_7. Bottom: WV, RE_4, S_4. The shaded regions indicate the plus/minus one standard deviation of P_D and P_{FA}.

.

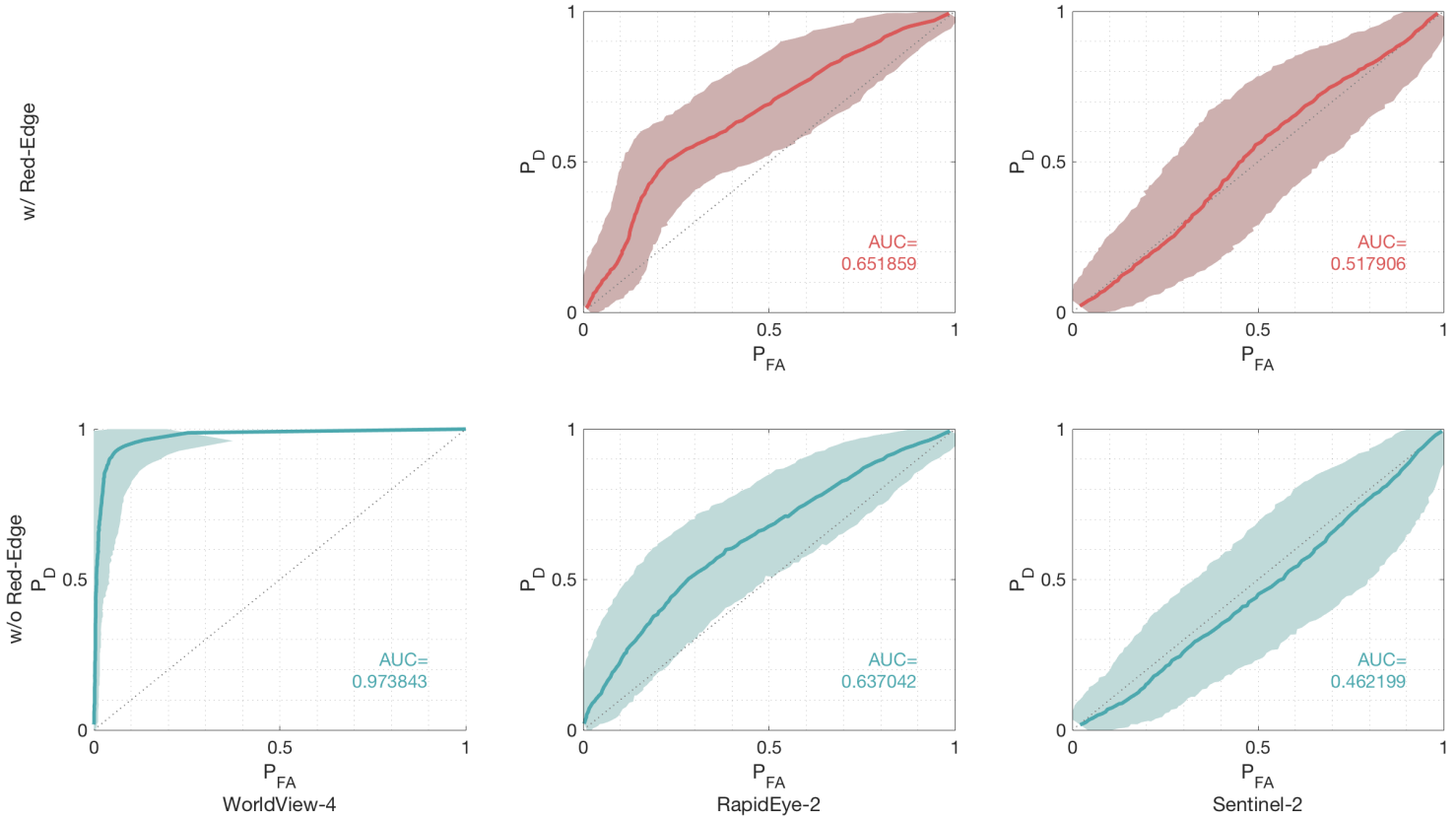

Figure 7.19: ROC curves based on the target weighted ICEM algorithm (ICEM1) applied to the five Coruche images. Top: (from left) RE5, S7. Bottom: WV, RE4, S4. The shaded regions indicate the plus/minus one standard deviation of P_D and P_{FA}.

.

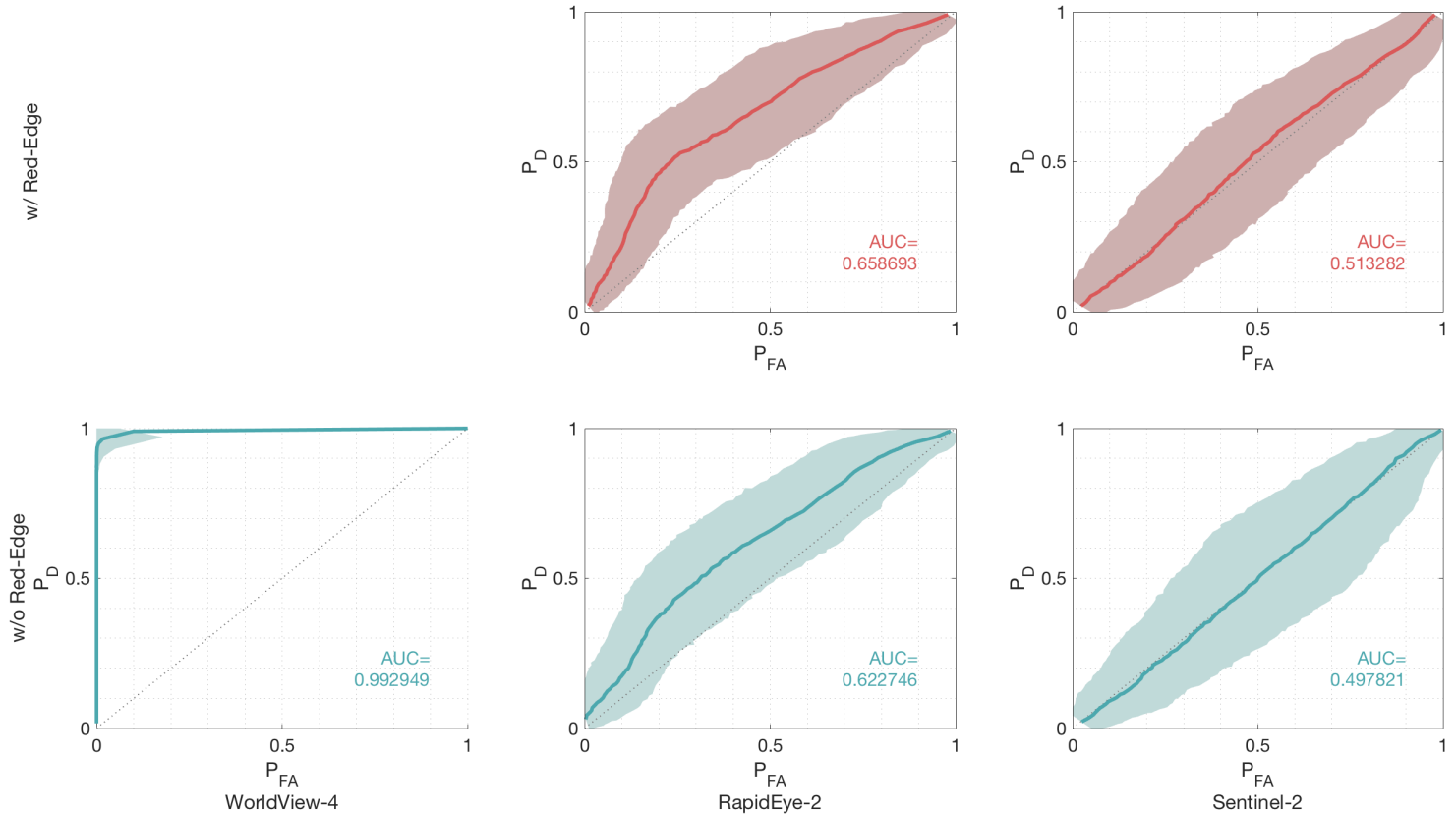

Figure 7.20: ROC curves based on the background weighted ICEM algorithm (ICEM2) applied to the five Coruche images. Top: (from left) RE5, S7. Bottom: WV, RE4, S4. The shaded regions indicate the plus/minus one standard deviation of P_D and P_{FA}.

.

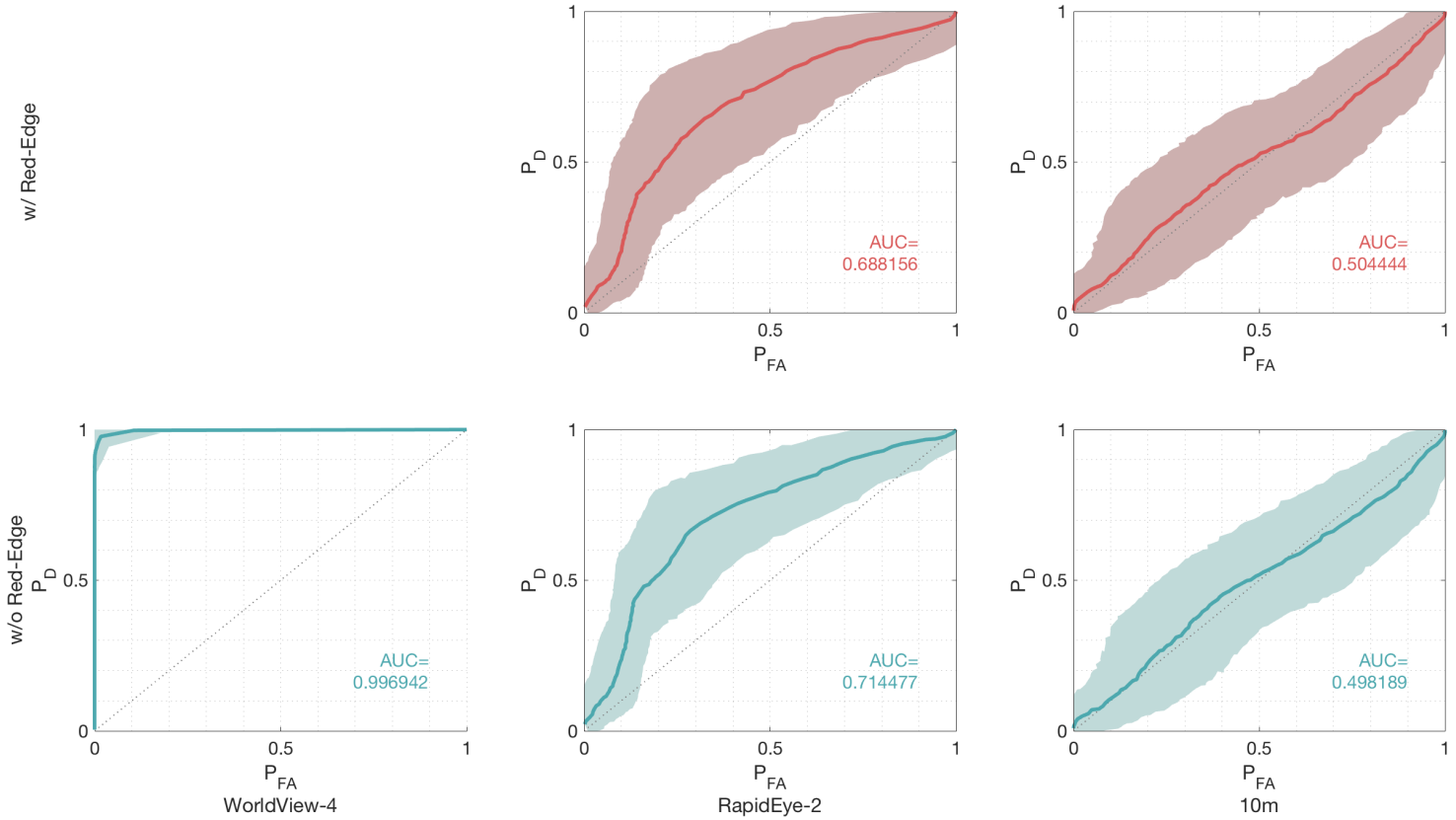

Figure 7.21: ROC curves based on the MTMF algorithm applied to the five Coruche images. Top: (from left) RE5, S7. Bottom: WV, RE4, S4. The shaded regions indicate the plus/minus one standard deviation of P_D and P_{FA}.

.

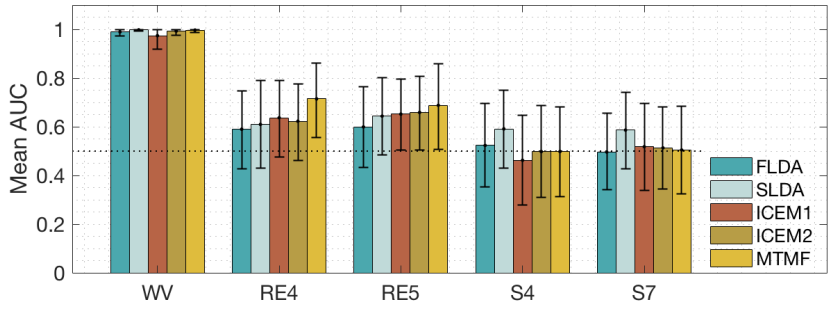

Figure 7.22: Summary of the AUC values for the different methods applied to all Coruche images. The errorbars indicate the AUC corresponding to the ROC curves for plus/minus one standard deviation in detection rate and false alarm rate.

7.2.5 Discussion

The assessment results for the supervised methods applied to the Coruche data (Figure 7.17-7.21) are summarized in Figure 7.22 in the same way as for the HYDICE experiment.

The idea of this experiment was to explore the effect of:

1. The information in the red-edge bands.

2. Spatial resolution.

The effect of the red-edge bands is better visualized by plotting the difference in mean AUC for the methods applied to the data with and without the red-edge bands (Figure 7.23). As can be seen from this plot, the additional information in the red-edge bands does not make much of a difference. For all cases except the FLDA method applied to the Sentinel data and the MTMF method applied to the RapidEye data, there is a small increase in mean AUC. However, compared to the variability of the performances these are *not* significant changes and may be random variations due to variability in Monte Carlo cross-validation processes.

One possible reason for the red-edge bands not showing any effect in the Sentinel images could be that the signals were attenuated in the Sen2Res processing (where the 20 m red-edge bands were sharpened to 10 m). However, since the red-edge bands do not make any difference in the RapidEye data either, an equally important factor might be the signal strength. Compared to the spatial resolution of the Sentinel and RapidEye sensors, the target might

be too small to contribute significantly to the signal.

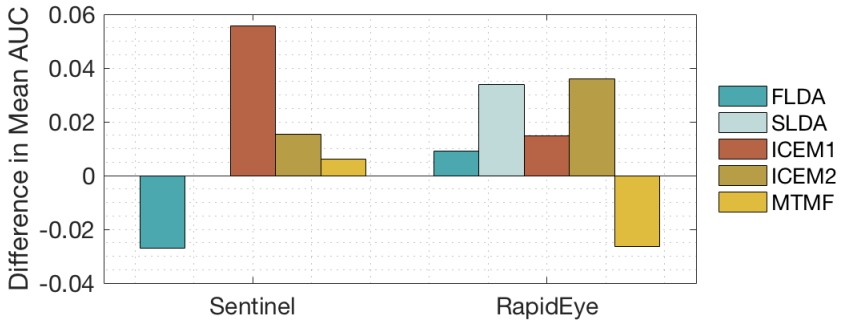

Figure 7.23: This graph displays the effect of the red-edge bands. The mean AUC corresponding to the data *without* the red-edge band(s) is subtracted from the mean AUC corresponding to the data *with* the red-edge band(s) for all methods applied to the Sentinel and the RapidEye data. Note: The SLDA applied to the Sentinel data is removed as it only uses information from the red band.

The effect of spatial resolution shows the same trend as for the HYDICE images: the higher the spatial resolution, the better the performance. For the coarse Sentinel images, the mean AUC value is around 0.5 for all methods except the SLDA, indicating *no* discrimination between target and background (the ROC curves follow the "by chance" diagonal). For the RapidEye images, the performances are somewhat better, but the methods still does not manage to discriminate between the target and background in a satisfactory manner. For the WorldView image, on the other hand, *all* methods are capable of discriminating between the target and the background almost perfectly.

It is interesting to note that, out of all methods applied to the Sentinel data, it is the SLDA, which only uses the information in the red band to discriminate between the target and background, that shows the best performance. Further, for the RapidEye data, the MTMF method shows the superior performance (just like in the HYDICE experiment).

Plotting the results of the methods applied to the actual test sites of the Sentinel and the RapidEye images is meaningless because of the poor performance on the ground-truth data. However, because of the promising assessment results for the application to the WorldView image, the result of applying the methods to this image are plotted in Figure 7.25-7.29 for three different CMR values.

Before applying the methods, an NDVI threshold operation was performed. Figure 7.24 shows an RGB composite of the test site (left) and the mask used to extract the green pixels (right). The mask was obtained by thresholding the NDVI image at the highest threshold possible without loosing any of the

14 known targets in the scene. Also, the known targets within the scene were removed from the training data before applying the methods.

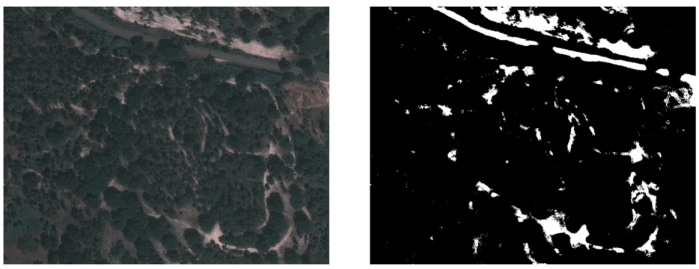

Figure 7.24: Left: RGB composite of the test scene. © DigitalGlobe, Inc. (2016), provided by European Space Imaging. Right: Mask used to extract the green pixels. Note: The colormap was inverted for visual appearance (black = 1, white = 0). That is, the black pixels were extracted.

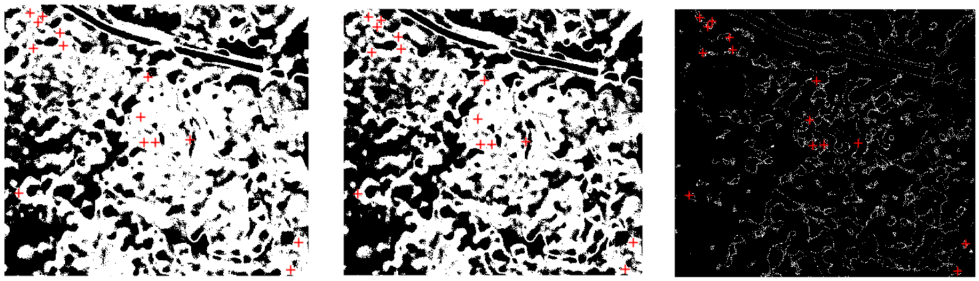

Figure 7.25: FLDA applied to the test site and evaluated by a CMR detector for three different thresholds. Left: CMR = 0.01. Middle: CMR = 0.05. Right: CMR = 0.5. The red markers indicate infected trees and the white is detection. The prior probabilities of the target and background classes were set to 0.02 and 0.98, respectively.

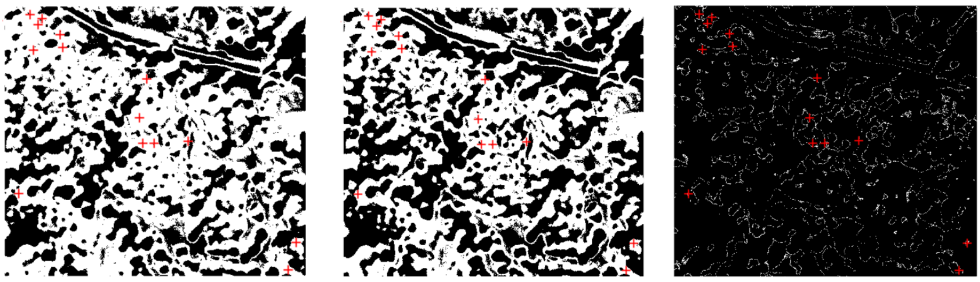

Figure 7.26: SLDA applied to the test site and evaluated by a CMR detector for three different thresholds. Left: CMR = 0.01. Middle: CMR = 0.05. Right: CMR = 0.5. The red markers indicate infected trees and the white is detection.

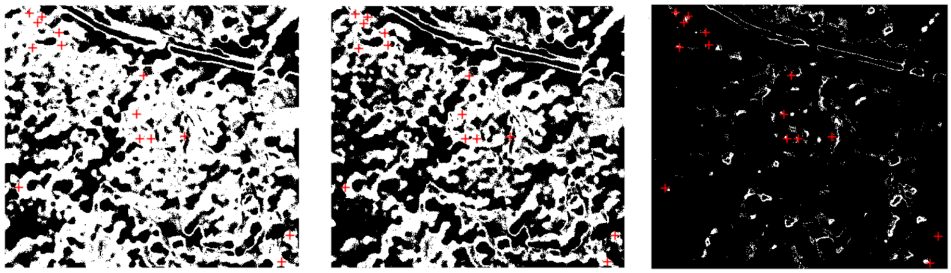

Figure 7.27: ICEM1 applied to the test site and evaluated by a CMR detector for three different thresholds. Left: CMR = 0.01. Middle: CMR = 0.05. Right: CMR = 0.5. The red markers indicate infected trees and the white is detection.

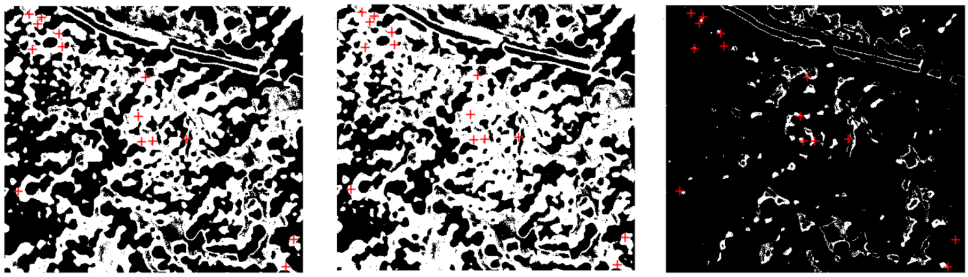

Figure 7.28: ICEM2 applied to the test site and evaluated by a CMR detector for three different thresholds. Left: CMR = 0.01. Middle: CMR = 0.05. Right: CMR = 0.5. The red markers indicate infected trees and the white is detection.

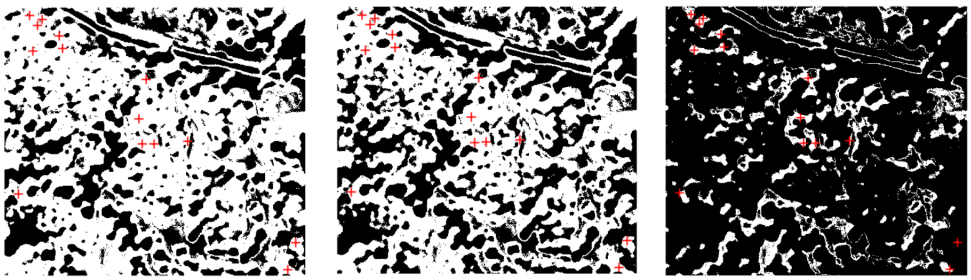

Figure 7.29: MTMF applied to the test site and evaluated by a CMR detector for three different thresholds. Left: CMR = 0.01. Middle: CMR = 0.05. Right: CMR = 0.5. The red markers indicate infected trees and the white is detection.

Comparing these plots to the RGB composites in Figure 6.3, it is apparent that the information in the WorldView bands of the provided ground truth is *not* enough to isolate infected trees in the image, at least not using the supervised methods tested in this thesis.

For a CMR of 0.01, it seems like all methods detect (more or less) everything except dense, healthy looking vegetation. As the CMR increases to 0.05, the number of detections decreases, shrinking the white regions to comprise less of the sparsely distributed vegetation as well. As the CMR reaches 0.5, the detections seems to be a type of shadowed, sparsely distributed vegetation mixed with soil. For FLDA and SLDA, only the outlines of these shadows are detected, therefore turning the detector into an edge detector.

There *is* a correlation between the detections and the known targets in the scene. However, there are too many false alarms. Some of the shadows do not contain trees, making them impossible to qualify as true positives. An attempt was made to set the CMR even higher, however, it did not help as it only lead to the rejection of (known) true positives.

There are several possible explanations for this result. One reason might be that the pansharpening process has attenuated the signal from the infected trees, and given the resolution enhanced pixels an appearance more similar to the average surroundings. Another possible explanation is connected to the accuracy in co-registration of image and ground-truth. When the target *and* the pixel sizes become very small, this accuracy becomes crucial. Perhaps (some of) the coordinates do not coincide perfectly with the infected trees, either because of inaccuracy in the measurements or as an effect of the georeferencing. There is, however, also the possibility that any possible discriminant signal is *not* sampled in the WorldView bands, which means that the classes overlap.

Unsupervised Method

7.2.6 2D-Corr-NLS

For the 2D-Corr-NLS method, the entire test scene was given as input, and some selected abundance maps (outputs) are presented for the WorldView (Figure 7.30), RapidEye (Figure 7.31) and Sentinel (Figure 7.32) images[7]. The red markers indicate infected trees and the intensity indicate the abundance of the endmember.

7. Since the results based on the Sentinel images and the RapidEye images with and without red-edge bands were very similar, only the results based on the images *with* red-edge information are presented.

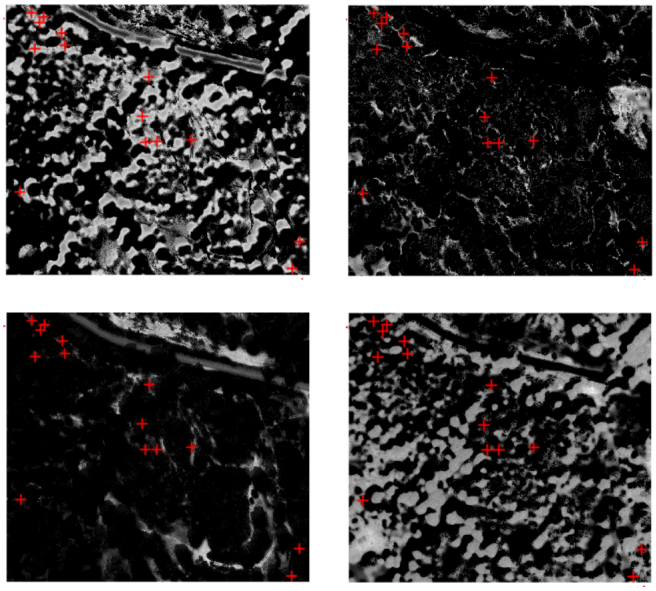

Figure 7.30: Four selected outputs from applying the 2D-Corr-NLS algorithm to the WorldView image. Upper left: Identified as shadows. Upper right: Identified as dirt. Lower left: Identified as dirt roads + asphalt. Lower right: identified as vegetation.

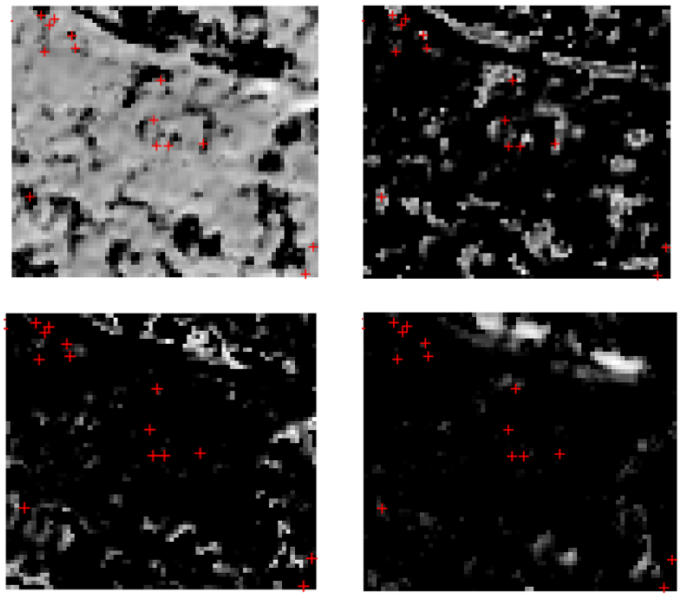

Figure 7.31: Four selected outputs from applying the 2D-Corr-NLS algorithm to the RapidEye image (with red-edge band). Upper left: Identified as vegetation. The remaining abundance maps are identified as dirt and dirt roads.

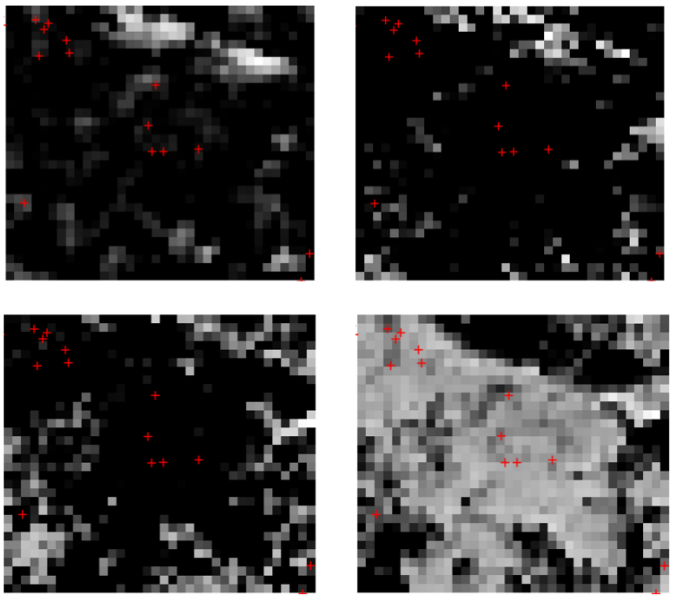

Figure 7.32: Four selected outputs from applying the 2D-Corr-NLS algorithm to the Sentinel image (with red-edge bans). Lower right: Identified as vegetation. The remaining abundance maps are identified as dirt and dirt roads.

7.2.7 Discussion

The application of the 2D-Corr-NLS method to the HYDICE images showed that the method was sensitive to the spatial resolution and to distribution of target and background pixels. The results from the application to the Coruche images further supports this.

Starting with the high resolution WorldView image (Figure 7.30), the result shows that the method is capable of detecting and distinguishing between the different main elements in the image scene: shadows (upper left), dirt (upper right), dirt roads (lower left) and vegetation (lower right). However, the method failed to detect infected trees as an isolated endmember, and the red markers seem to blend in with the shadows. An attempt was also made to force the algorithm to search for a higher number of endmembers, without success. It only lead to a splitting of the existing endmembers and producing additional noise-like abundance maps.

For the RapidEye and Sentinel images (Figure 7.31 and Figure 7.32, respectively), on the other hand, the method is unable to distinguish between dirt and dirt roads, and shadows are not detected as an endmember. The four "best" clusters (according to the criterion) represent vegetation and dirt/dirt roads. The dirt/dirt road class is divided into three endmembers, just like the target class were in the coarse HYDICE images.

As the spatial resolution decreases, areas with densely distributed targets will still (to some degree) yield pure pixels. In contrast, areas with sparsely distributed targets will produce mixed pixels. In this way, the target gets split into separate endmembers.

8

Conclusion and Future Work

This thesis has examined three different spectral unmixing methods (ICEM, MTMF and 2D-Corr-NLS) and two classical feature extraction methods (FLDA and SLDA) to use in a detection problem. In addition, based on an analysis of the ICEM method, an alternative ICEM method (ICEM2) is suggested and examined. Further, a method to combine the MT and the MF step in the MTMF method is proposed. Each method has been presented, discussed and tested on two different data sets. The performances of the supervised methods have been quantitatively evaluated using a constant miss rate detector, while the unsupervised method has been evaluated qualitatively.

Based on the results presented in the previous chapter, the following findings and observations were made:

- FLDA: The FLDA method shows the best result when applied to the HYDICE_h2 image, where it (almost) tangents the SLDA method. Other than that, the performance is at best mediocre.

- SLDA: The SLDA method shows the best performance (highest AUC and smallest variability) of all methods when applied to the high spatial high spectral resolution HYDICE_h2 image. Further, it is the only method that manages to discriminate (at all) between target and background in the

Sentinel image.

- ICEM: The ICEM methods require high spatial resolution *or* high spectral resolution in order to achieve good results. ICEM1 and ICEM2 show quite similar performances on all data sets, ICEM2 being somewhat superior.

- MTMF: The method is designed for hyperspectral data, but turns out to yield good results for multispectral imagery as well. Moreover, it shows the highest AUC values of all methods when applied to the multispectral HYDICE images.

- 2D-Corr-NLS: The method is designed for multispectral imagery, but appears to be quite sensitive to variability in background *and* variations in the proportions of target and background material, as the spatial resolution decreases.

8.1 Conclusion

The aim of this thesis was to explore the viability of PWN detection by Sentinel-2 data, based on signal processing and machine learning techniques such as spectral unmixing and classical feature extraction methods.

The two main research questions presented in the introduction were the following:

- *Is the spatial resolution of the Sentinel-2 data a limitation to successful use of the selected methods for PWN detection?*

- *Is the spectral resolution of the Sentinel-2 data a limitation to successful use of the selected methods for PWN detection, and do the red-edge bands help the discrimination?*

The answers to these questions are, based on the experiments in this thesis, found to be *yes* and *maybe*.

All the supervised methods managed to discriminate (almost) perfectly between the labeled target and background samples in the WorldView image. That is, based on the information in the blue, green, red and infrared bands, the methods could discriminate between the target and the background. Since both the RapidEye and the Sentinel data have equivalent bands, but still did not manage to discriminate between target and background, this suggests that it is the spatial resolution that is the major constraint. However, as seen

when the methods were applied directly to the WorldView test site (not the labeled samples), the information in the bands was insufficient to isolate the target. Hence, the spectral resolution *may* also be a limiting factor in the low resolution images.

This study could not find the red-edge bands to help the discrimination. However, that does *not* imply that they do not. The lack of effect may be due to the coarse resolution and the small size of the targets: Any possible discriminating signals in the red-edge bands may have been too weak.

Hence, the *spatial* resolution of the Sentinel-2 data is concluded to be too low for the selected methods in order to be useful for PWN detection. However, based on the experiments in this thesis, it is not possible to tell whether or not the *spectral* resolution of the Sentinel-2 data is a limiting factor.

Limitations

Several limitations to this study need to be acknowledged. Three possible sources of error are:

- The sample sizes of healthy and infected trees are small. As the sample size decreases, the sample is less likely to ensure a representative distribution for the population.

- The samples for the background class are manually sampled from the WorldView data and is probably somewhat biased towards very healthy looking trees in good lighting. Further, since the pre-processing step is replaced with a NDVI threshold operation, these samples will not cover the entire background class.

- The sharpening of the WorldView image and the red-edge bands in the Sentinel-2 image may have distorted the information in the pixel vectors, which may have affected the results.

Also, due to the lack of high spatial resolution images *with* red-edge information, the effect of the red-edge bands could not be investigated in a satisfactory manner.

8.2 Future Work

There are many possible improvements to the specific methods used in this thesis. Some suggestions are listed below:

- The mentioned pre-processing step, extracting pixels that *previously* have shown a "healthy-pine"-like signature could be implemented and investigated.

- Since estimating the probability densities numerically allows for the possibility of assuming more "complicated" (compared to the Gaussian) distributions without aggravating the problem, other distributions could be tested to model the target class in the Monte Carlo simulations. However, since the FLDA and MF statistics are *ideal* under the assumption of Gaussian data[1], bell-shaped distributions are favorable. A Gamma distribution with a large shape parameter is thus interesting. With the pixel intensities being purely positive, the positive support of the Gamma is also beneficial.

- The MTMF method showed the best potential for the multispectral low resolution images in the HYDICE experiment. However, because the MT range may vary between sub-samples (as seen in the Coruche experiment), it would be interesting to, in stead of combining the MT and the MF step using a sigmoid, try to remove a *percentage* of the samples, corresponding to the greatest MT values.

Another approach, that has not been considered in this thesis, is the use of multi-temporal data. By considering a group of pixels that (at time step one) contain healthy pines, one can follow the evolution of the signals over time to find pixels where the signal diverges from that of the healthy trees. The multi-temporal information could be used to filter out the signature of both healthy and infected trees, in this way enabling the use of the Neyman-Pearson detector *and* take into account the phenological cycle.

However, equally important is further research on multispectral unmixing. Most studies in the field of spectral unmixing have only focused on hyperspectral imagery. Though, for operational systems where the number of available bands are *tens* instead of *hundreds*, the challenges and premises of the unmixing are somewhat different. More research on the topic of multispectral unmixing is

1. When the target and background classes follow Gaussian distributions, and the covariance matrices are equal, the likelihood ratio detector reduces to the FLDA. Depending on how the threshold is set, this gives an ideal detector in the Bayes (minimum probability of error) or Neyman-Pearson (maximum probability of detection for a given probability of false alarm) sense [Manolakis and Shaw, 2002b].

thus needed. This could include inventing new algorithms, specially designed for multispectral data, as well as modifying known hyperspectral algorithms to fit the multispectral problem. Hereunder, studies examining the limit between hyperspectral and multispectral data may be useful. To this purpose, the hyperspectral data from the Joint Research Centre campaign (mentioned in the introduction) could be interesting.

Part III

Appendices

A

Whitening Transformations

Whitening, or spherizing, is a linear transformation where the goal is to convert the a random vector $\boldsymbol{x} \in \mathbb{R}^p$ to a new random vector $\boldsymbol{y} \in \mathbb{R}^p$ with unit variance in all directions. That is,

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}, \quad \boldsymbol{W} \in \mathbb{R}^{p \times p} \tag{A.1}$$

such that the new covariance matrix, $\boldsymbol{\Sigma}_y = \boldsymbol{I}$ [Stark and Woods, 2014]. $\boldsymbol{W}$ is called the *whitening matrix*, and since

$$\boldsymbol{\Sigma}_y = \boldsymbol{W}\boldsymbol{\Sigma}_x\boldsymbol{W}^T = \boldsymbol{I} \leftrightarrow \boldsymbol{W}\boldsymbol{\Sigma}_x\boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{W} \tag{A.2}$$

is satisfied for all $\boldsymbol{W}$ such that

$$\boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{\Sigma}_x,^{-1} \tag{A.3}$$

there are infinitely many valid matrices $\boldsymbol{W}$[1] [Kessy et al., 2018].

A.1 PCA Whitening

One of the most common whitening matrices is revealed by considering the eigen-decomposition of the covariance matrix corresponding to the original

1. Whitened data will remain whitened after any rotation (multiplication by any orthogonal matrix) [Hyvarinen et al., 2009].

data. The eigendecomposition can be written as

$$\boldsymbol{\Sigma}_x \boldsymbol{E} = \boldsymbol{E}\boldsymbol{\Lambda}, \tag{A.4}$$

where $\boldsymbol{E}$ is the matrix whose columns are eigenvectors of $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Lambda}$ is the diagonal matrix containing the associated eigenvalues. By some simple matrix multiplications, this can be rewritten as

$$\begin{aligned}
\boldsymbol{\Sigma}_x^{-1} &= \boldsymbol{E}\boldsymbol{\Lambda}^{-1}\boldsymbol{E}^{-1} \\
&\overset{2}{=} \boldsymbol{E}\boldsymbol{\Lambda}^{-1}\boldsymbol{E}^{T} \\
&= (\boldsymbol{E}\boldsymbol{\Lambda}^{-\frac{1}{2}})(\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{E}^{T}) \\
&= (\boldsymbol{E}\boldsymbol{\Lambda}^{-\frac{1}{2}})(\boldsymbol{E}\boldsymbol{\Lambda}^{-\frac{1}{2}})^{T}.
\end{aligned} \tag{A.5}$$

From this expression, it is clear that

$$\boldsymbol{W} = \boldsymbol{E}\boldsymbol{\Lambda}^{-\frac{1}{2}} \tag{A.6}$$

is a valid whitening matrix, and this is called the *principal component analysis* (PCA) *whitening* [Kessy et al., 2018].

The interpretation of the PCA whitening transform is a combination of rotation and scaling of the variables: the variables are orthogonalized by rotation, using the eigenvectors, and scaled by the square root of the eigenvalues to obtain unit variance [Kessy et al., 2018].

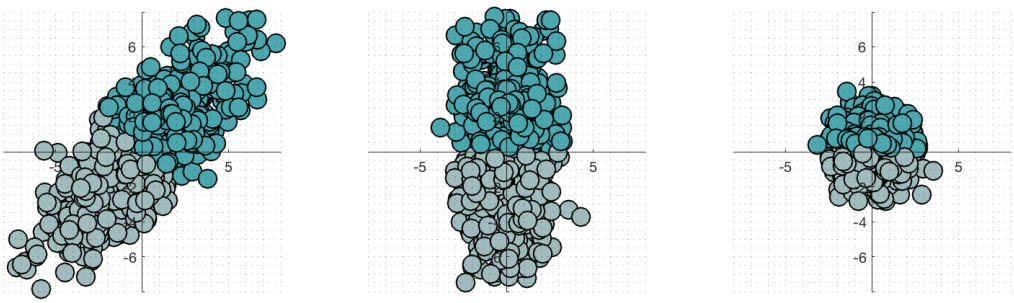

Figure A.1: PCA whitening (different colors to show the rotation). Left: Original data. Middle: Data rotated by multiplying by $\boldsymbol{E}$. Right: Data scaled and rotated by multiplying by $\boldsymbol{E}\boldsymbol{\Lambda}^{-\frac{1}{2}}$.

2. Since covariance matrices are always symmetric, their eigenvectors are orthogonal and $\boldsymbol{E}^{-1} = \boldsymbol{E}^{T}$.

A.2 ZCA Whitening

Another important whitening matrix is obtained when $\boldsymbol{W}$ is forced to be symmetrical, that is when

$$\boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{W}^2. \tag{A.7}$$

Then it follows (by substituting (A.7) into (A.3)) that

$$\boldsymbol{W} = \boldsymbol{\Sigma}^{-\frac{1}{2}} \tag{A.8}$$

is a valid whitening matrix [Bell and Sejnowski, 1997]. This is called the *zero-phase component analysis* (ZCA) *whitening* (also called the *Mahalanobis whitening*) and consists in rescaling only [Kessy et al., 2018].

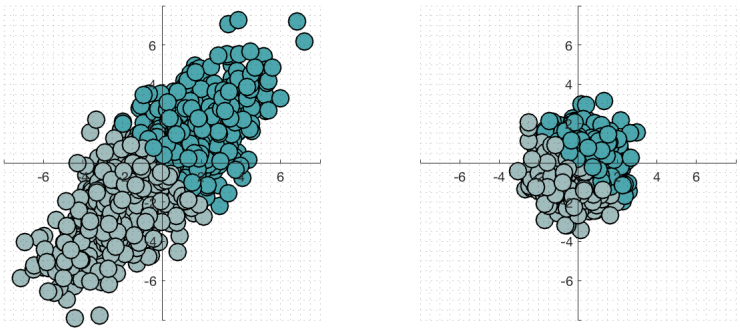

Figure A.2: ZCA whitening (different colors to show the lack of rotation). Left: Original data. Right: Data scaled by multiplying by $\boldsymbol{\Sigma}^{-\frac{1}{2}}$.

B

Principal Component Analysis

The *Principal Component Analysis* (PCA) is a linear, unsupervised transformation of the data, where the goal is to remove information redundancies. The measure of information redundancy is in the PCA defined as the statistical measure covariance. The aim is thus to find the transformation matrix, $\boldsymbol{A}$ in the transformation

$$\begin{aligned} \boldsymbol{y} = \boldsymbol{A}^T \boldsymbol{x}, \quad & \boldsymbol{y} \in \mathbb{R}^\ell \\ & \boldsymbol{A} \in \mathbb{R}^{\ell \times \ell}, \end{aligned} \tag{B.1}$$

such that the transformed variables (transformed features), y_i, are *uncorrelated*, i.e. linearly independent.

Since the correlation matrix of an uncorrelated vector is diagonal, the correlation matrix of $\boldsymbol{y}$ must be forced to become a diagonal matrix. The correlation matrix of $\boldsymbol{y}$ can be written as

$$\boldsymbol{\Sigma}_y = E\left\{\boldsymbol{y}\boldsymbol{y}^T\right\} = E\left\{\boldsymbol{A}^T \boldsymbol{x}\boldsymbol{x}^T \boldsymbol{A}\right\} = \boldsymbol{A}^T \boldsymbol{\Sigma}_x \boldsymbol{A}, \tag{B.2}$$

where $\boldsymbol{\Sigma}_x$ is the correlation matrix of the original feature vector, $\boldsymbol{x}$. This matrix is always symmetric and positive semidefinite, which implies that its eigenvectors, $\boldsymbol{e}_i$, are orthogonal and eigenvalues, $\boldsymbol{\lambda}_i$, non-negative. Its eigendecomposition

can be written as

$$\boldsymbol{\Sigma}_x \boldsymbol{E} = \boldsymbol{E}\boldsymbol{\Lambda}, \tag{B.3}$$

where $\boldsymbol{E}$ is the matrix with the eigenvectors as columns and $\boldsymbol{\Lambda}$ is the diagonal matrix with the eigenvalues on the diagonal. By a simple right multiplication of $\boldsymbol{E}^{-1} = \boldsymbol{E}^T$ we get that

$$\boldsymbol{\Sigma}_x = \boldsymbol{E}\boldsymbol{\Lambda}\boldsymbol{E}^T, \tag{B.4}$$

which, by substitution into (B.2) gives

$$\boldsymbol{\Sigma}_y = \boldsymbol{A}^T \boldsymbol{E}\boldsymbol{\Lambda}\boldsymbol{E}^T \boldsymbol{A}. \tag{B.5}$$

The choice $\boldsymbol{A} = \boldsymbol{E}$ makes the correlation matrix of $\boldsymbol{y}$ equal to $\boldsymbol{\Lambda}$, which is diagonal, and thus the transformed variables uncorrelated [Theodoridis, 2009].

C

Parameter Tuning

C.1 FLDA

In order to make the inversion of the within-class covariance matrix numerically stable, the dimensionality of the hyperspectral HYDICE images are reduced via PCA. To decide the optimal dimensionality after the PCA, the mean AUC value (based on 300 ROC curves) is considered for different dimensions for each image. The results are presented in the following plots.

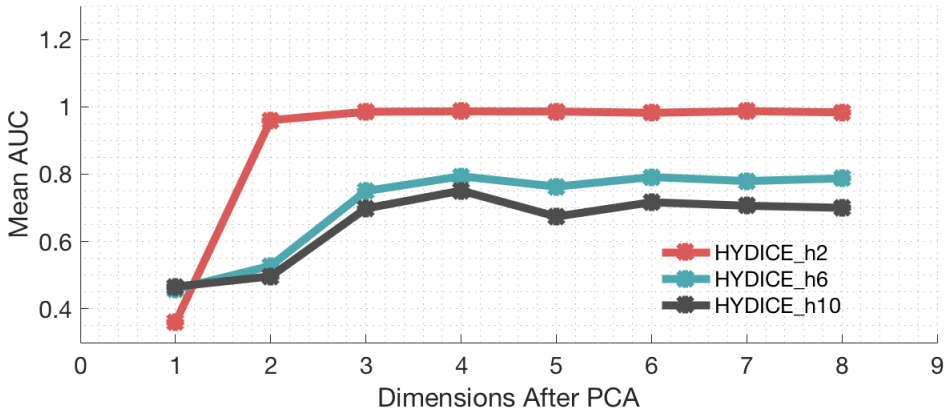

Figure C.1: Mean AUC values. Reducing the dimension to *four* (before applying FLDA) yields the highest mean AUC for all the hyperspectral HYDICE images.

C.2 SLDA

The `slda.m` function requires the regularization parameter for the ℓ_2-penalty and *either* a *upper boundary* for regularization parameter for the ℓ_1-penalty *or* a parameter STOP defining the number of non-zero elements in $\boldsymbol{w}$ as input.

The optimal values for the parameters are found by evaluating the AUC for 30 ROC curves (in ten random samples for the HYDICE data), for different values of the parameters. The results for the different images are shown in the following plots.

C.2.1 HYDICE_m2

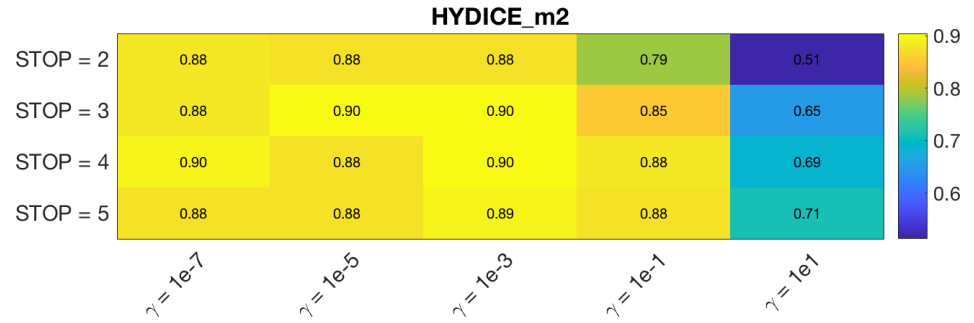

Figure C.2: Mean AUC values. From these plots, the parameters $STOP = 4$ and $\gamma = 1e-3$ is chosen for the HYDICE_m2 data set.

C.2.2 HYDICE_h2

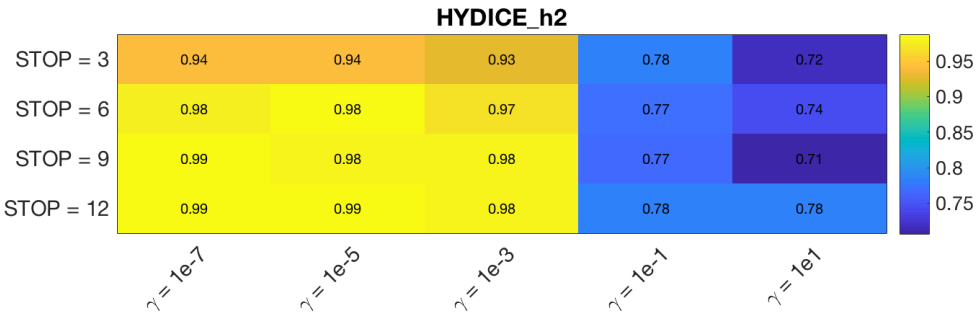

Figure C.3: Mean AUC values. From these plots, the parameters $STOP = 12$ and $\gamma = 1e-5$ is chosen for the HYDICE_h2 data set.

C.2.3 HYDICE_m6

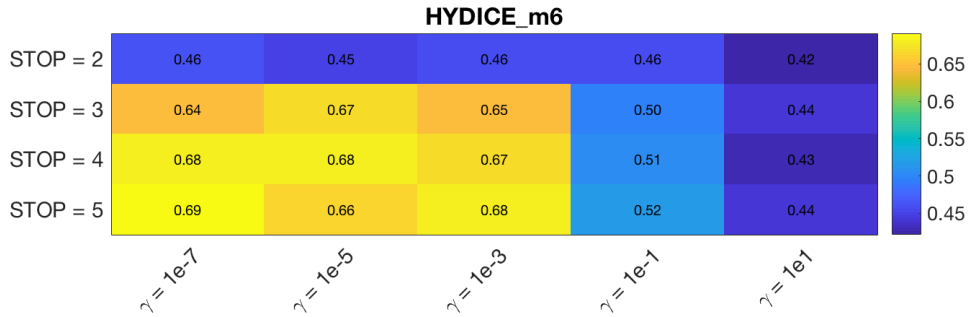

Figure C.4: Mean AUC values. From these plots, the parameters $STOP = 5$ and $\gamma = 1e-7$ is chosen for the HYDICE_m6 data set.

C.2.4 HYDICE_h6

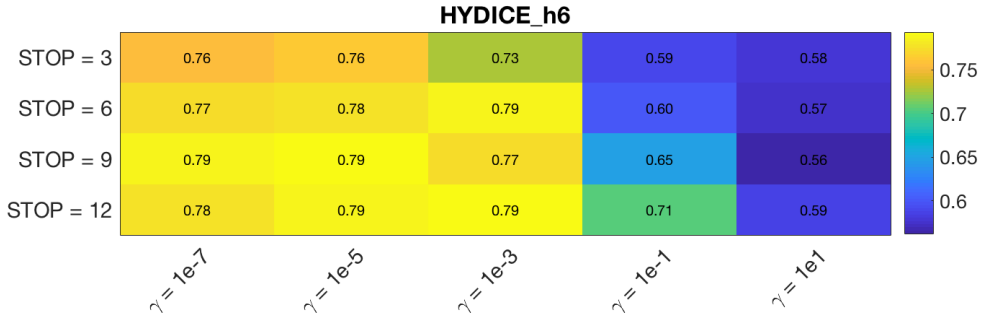

Figure C.5: Mean AUC values. From these plots, the parameters $STOP = 9$ and $\gamma = 1e-5$ is chosen for the HYDICE_h6 data set.

C.2.5 HYDICE_m10

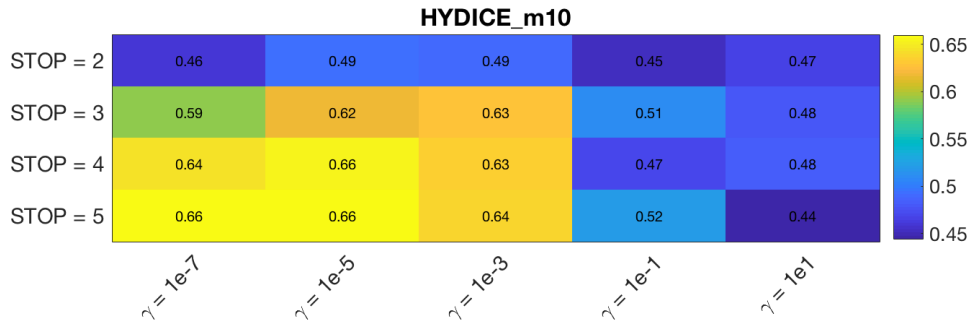

Figure C.6: Mean AUC values. From these plots, the parameters $STOP = 5$ and $\gamma = 1e-5$ is chosen for the HYDICE_m10 data set.

C.2.6 HYDICE_h10

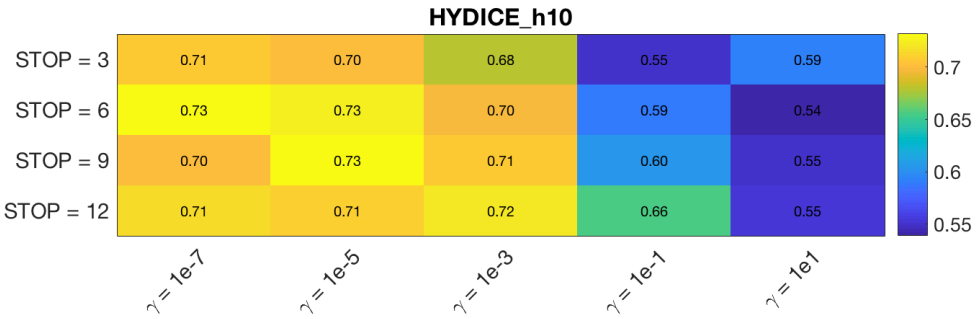

Figure C.7: Mean AUC values. From these plots, the parameters $STOP = 6$ and $\gamma = 1e-5$ is chosen for the HYDICE_h10 data set.

C.2.7 S4

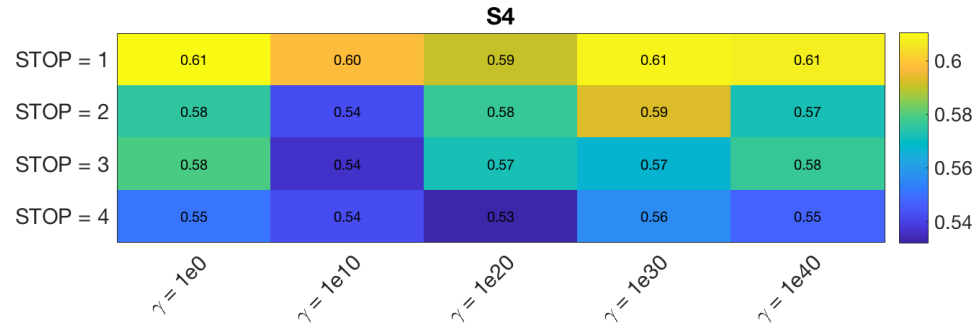

Figure C.8: Mean AUC values. From these plots, the parameters $STOP = 1$, making the γ insignificant for the S4 data set.

C.2.8 S7

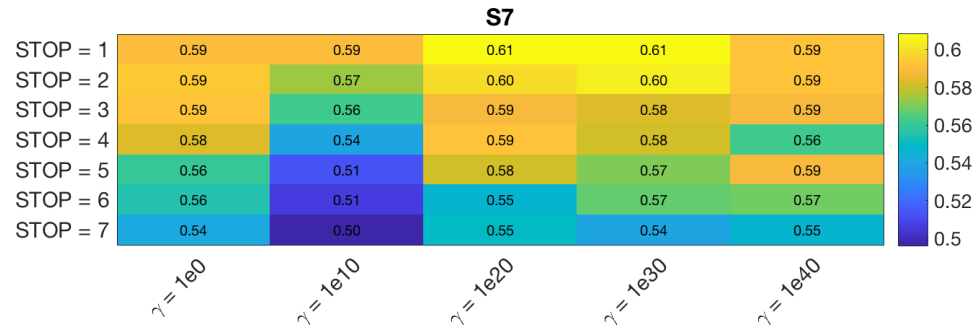

Figure C.9: Mean AUC values. From these plots, the parameters $STOP = 1$, making the γ insignificant for the S7 data set.

C.2.9 RE4

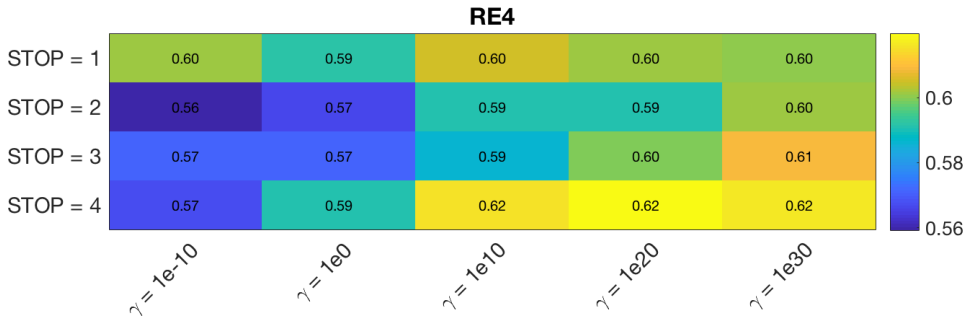

Figure C.10: Mean AUC values. From these plots, the parameters $STOP = 4$ and $\gamma = 1e20$ is chosen for the RE4 data set.

C.2.10 RE5

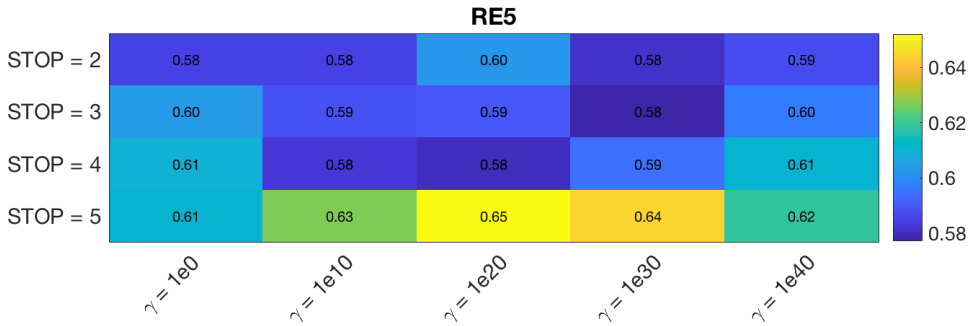

Figure C.11: Mean AUC values. From these plots, the parameters $STOP = 5$ and $\gamma = 1e20$ is chosen for the RE5 data set.

C.2.11 WV

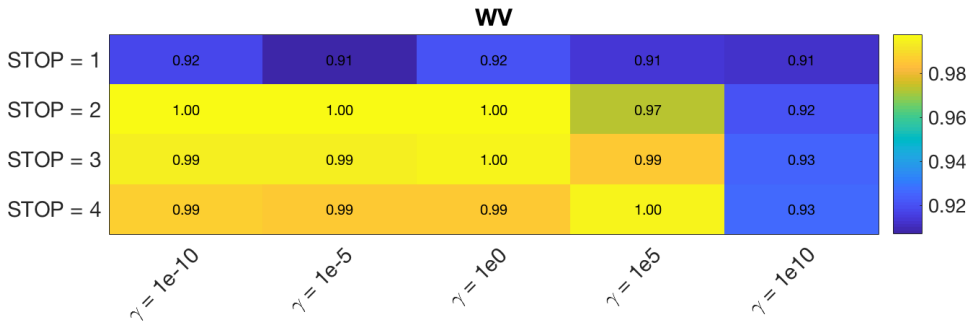

Figure C.12: Mean AUC values. From these plots, the parameters $STOP = 2$ and $\gamma = 1e0$ is chosen for the WV data set.

C.3 MTMF

The MTMF score requires two parameters, a and b. a is set to 1 while the optimal values for b are found by evaluating the area under the ROC curve for different values of the parameters. The following plots were produced (the error bars represent the difference between the AUC of the upper and the lower boundaries of the ROC, corresponding to plus/minus one standard deviation in detection rate and false alarm rate):

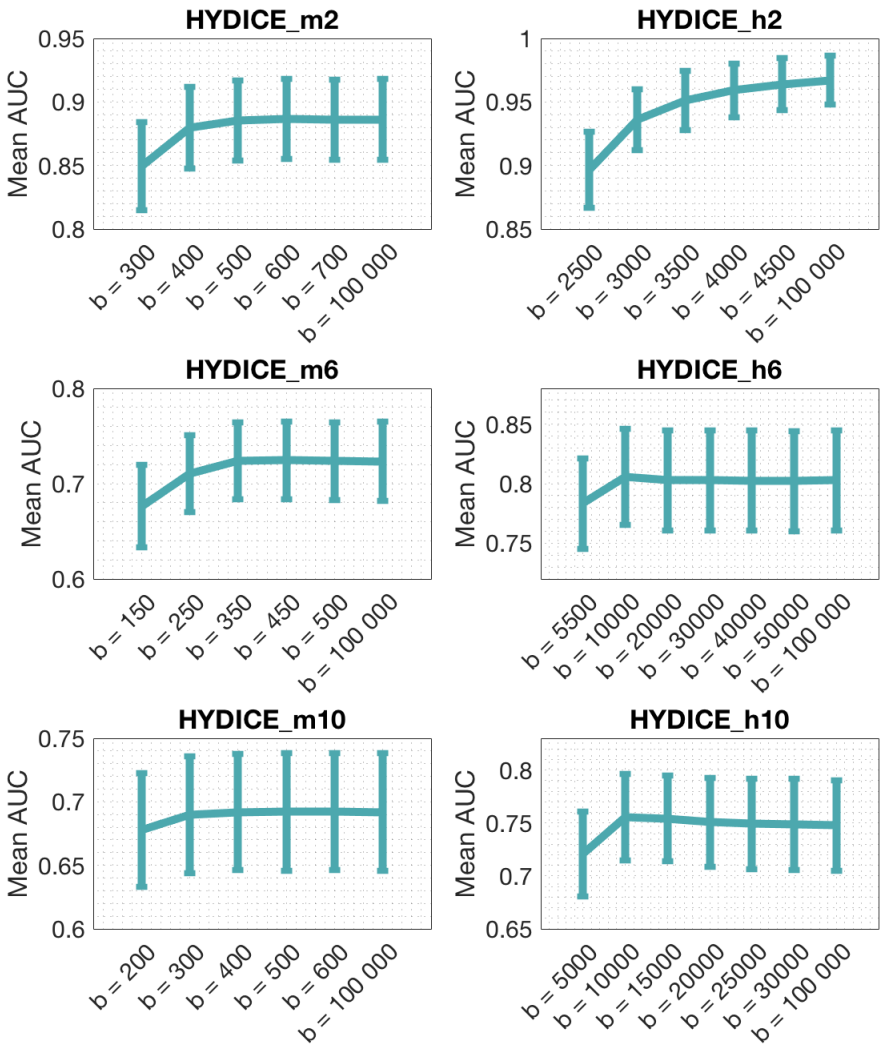

Figure C.13: Mean AUC values for different values of b. Based on these plots, b is set to 600, 100 000,450,10 000, 500 and 10 000 for the HYDICE_m2, HYDICE_h2, HYDICE_m6, HYDICE_h6, HYDICE_m10 and HYDICE_h10 images, respectively.

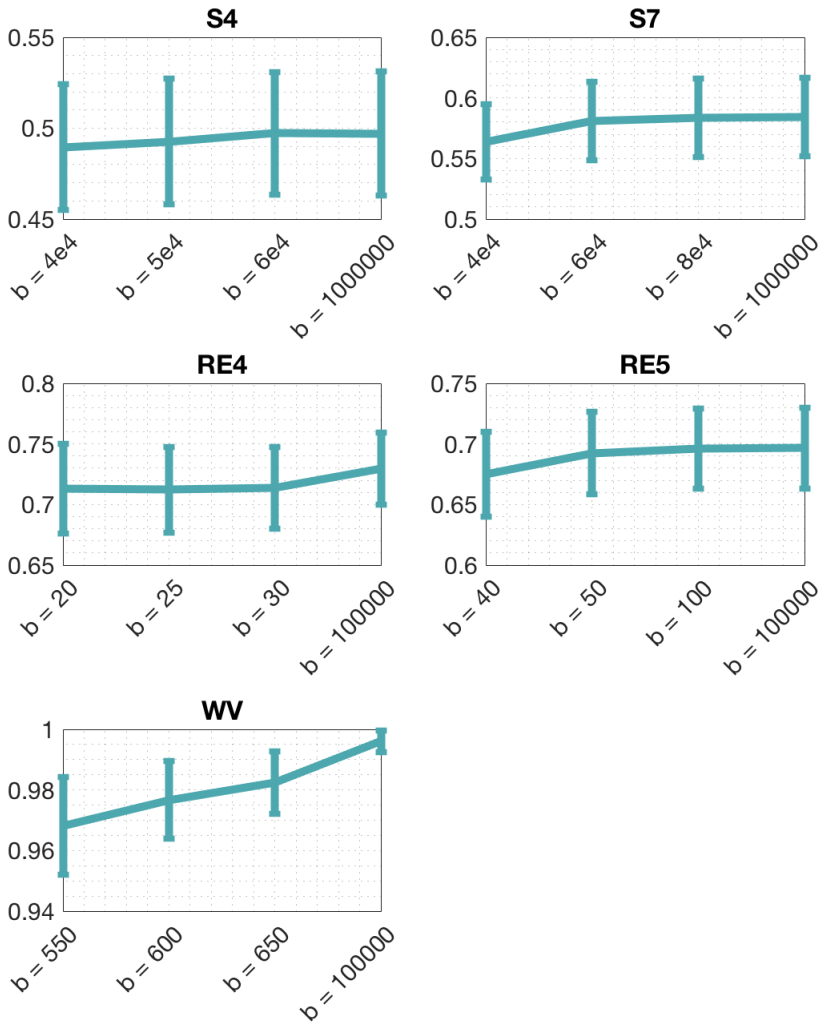

Figure C.14: Mean AUC values for different values of b. Based on these plots, b is set to 1000000 for *all* images, reducing the MTMF method to MF.

Bibliography

[Adelabu et al., 2014] Adelabu, S., Mutanga, O., and Adam, E. (2014). Evaluating the impact of red-edge band from rapideye image for classifying insect defoliation levels. *ISPRS Journal of Photogrammetry and Remote Sensing*, 95:34–41.

[Anderson, 1973] Anderson, T. W. (1973). An asymptotic expansion of the distribution of the studentized classification statistic w1. *The Annals of Statistics*, pages 964–972.

[Atkins et al., 2017] Atkins, S., Einarsson, G., Ames, B., and Clemmensen, L. (2017). Proximal methods for sparse optimal scoring and discriminant analysis. *arXiv preprint arXiv:1705.07194*.

[Beck, 2015] Beck, P. (2015). *The feasibility of detecting trees affected by the pine wood nematode using remote sensing*. Publications Office, Luxembourg.

[Bell and Sejnowski, 1997] Bell, A. J. and Sejnowski, T. J. (1997). The "independent components" of natural scenes are edge filters. *Vision research*, 37(23):3327–3338.

[Bezdek et al., 1984] Bezdek, J. C., Ehrlich, R., and Full, W. (1984). Fcm: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, 10(2-3):191–203.

[BIO IS, 2011] BIO IS, B. (2011). Disturbances of eu forests caused by biotic agents, final report prepared for european commission (dg env).

[Bioucas-Dias and Nascimento, 2008] Bioucas-Dias, J. M. and Nascimento, J. M. (2008). Hyperspectral subspace identification. *IEEE Transactions on Geoscience and Remote Sensing*, 46(8):2435–2445.

[Bioucas-Dias et al., 2013] Bioucas-Dias, J. M., Plaza, A., Camps-Valls, G., Scheunders, P., Nasrabadi, N., and Chanussot, J. (2013). Hyperspectral remote sensing data analysis and future challenges. *IEEE Geoscience and remote*

sensing magazine, 1(2):6–36.

[Bioucas-Dias et al., 2012] Bioucas-Dias, J. M., Plaza, A., Dobigeon, N., Parente, M., Du, Q., Gader, P., and Chanussot, J. (2012). Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches. *IEEE journal of selected topics in applied earth observations and remote sensing*, 5(2):354–379.

[Bishop, 2006] Bishop, C. (2006). *Pattern recognition and machine learning*. Springer, New York.

[Boardman, 1998] Boardman, J. W. (1998). Leveraging the high dimensionality of aviris data for improved sub-pixel target unmixing and rejection of false positives: mixture tuned matched filtering. In *Summaries of the seventh JPL Airborne Geoscience Workshop, JPL Publication, 1998*, volume 97, pages 55–56. NASA Jet Propulsion Laboratory.

[Boardman and Kruse, 2011] Boardman, J. W. and Kruse, F. A. (2011). Analysis of imaging spectrometer data using n-dimensional geometry and a mixture-tuned matched filtering approach. *IEEE Transactions on Geoscience and Remote Sensing*, 49(11):4138–4152.

[Boas et al., 2016] Boas, D. A., Pitris, C., and Ramanujam, N. (2016). *Handbook of biomedical optics*. CRC press.

[Bodnar et al., 2017] Bodnar, T., Mazur, S., Ngailo, E., and Parolya, N. (2017). Discriminant analysis in small and large dimensions. *arXiv preprint arXiv:1705.02826*.

[Brodu, 2017] Brodu, N. (2017). Super-resolving multiresolution images with band-independent geometry of multispectral pixels. *IEEE Transactions on Geoscience and Remote Sensing*, 55(8):4610–4617.

[Brondsted, 2012] Brondsted, A. (2012). *An introduction to convex polytopes*, volume 90. Springer Science & Business Media.

[Caliński and Harabasz, 1974] Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27.

[Campbell, 2011] Campbell, J. (2011). *Introduction to remote sensing*. Guilford Press, New York.

[Chuvieco, 2009] Chuvieco, E. (2009). *Fundamentals of satellite remote sensing*.

CRC press.

[Clemmensen et al., 2011] Clemmensen, L., Hastie, T., Witten, D., and Ersbøll, B. (2011). Sparse discriminant analysis. *Technometrics*, 53(4):406–413.

[Cooper, 1995] Cooper, L. N. (1995). *How We Learn, how We Remember: Toward an Understanding of Brain and Neural Systems: Selected Papers of Leon N. Cooper*, volume 10. World scientific.

[Cunningham et al., 2008] Cunningham, P., Cord, M., and Delany, S. J. (2008). Supervised learning. In *Machine learning techniques for multimedia*, pages 21–49. Springer.

[Cushnie, 1987] Cushnie, J. L. (1987). The interactive effect of spatial resolution and degree of internal variability within land-cover types on classification accuracies. *International Journal of Remote Sensing*, 8(1):15–29.

[Deville and Puigt, 2007] Deville, Y. and Puigt, M. (2007). Temporal and time-frequency correlation-based blind source separation methods. part i: Determined and underdetermined linear instantaneous mixtures. *Signal Processing*, 87(3):374–407.

[Dobigeon et al., 2014] Dobigeon, N., Tourneret, J.-Y., Richard, C., Bermudez, J. C. M., McLaughlin, S., and Hero, A. O. (2014). Nonlinear unmixing of hyperspectral images: Models and algorithms. *IEEE Signal Processing Magazine*, 31(1):82–94.

[Dubitzky et al., 2007] Dubitzky, W., Granzow, M., and Berrar, D. P. (2007). *Fundamentals of data mining in genomics and proteomics*. Springer Science & Business Media.

[Eitel et al., 2011] Eitel, J. U., Vierling, L. A., Litvak, M. E., Long, D. S., Schulthess, U., Ager, A. A., Krofcheck, D. J., and Stoscheck, L. (2011). Broadband, red-edge information from satellites improves early stress detection in a new mexico conifer woodland. *Remote Sensing of Environment*, 115(12):3640–3646.

[Elachi, 2006] Elachi, C. (2006). *Introduction to the physics and techniques of remote sensing*. Wiley-Interscience, Hoboken, N.J.

[ESA, n.d.-a] ESA (n.d.-a). *Radiometric Resolutions*. `https://earth.esa.int/web/sentinel/user-guides/sentinel-2-msi/resolutions/radiometric` [Accessed: 20.03.18].

[ESA, n.d.-b] ESA (n.d.-b). *SENTINEL-2*. https://sentinel.esa.int/web/sentinel/missions/sentinel-2 [Accessed: 20.04.18].

[Franke et al., 2006] Franke, J., Heinzel, V., and Menz, G. (2006). Assessment of ndvi-differences caused by sensor specific relative spectral response functions. In *Geoscience and Remote Sensing Symposium, 2006. IGARSS 2006. IEEE International Conference on*, pages 1138–1141. IEEE.

[Gupta and Bajaj, 2018] Gupta, S. and Bajaj, C. (2018). *Efficient Clustering-Based Noise Covariance Estimation for Maximum Noise Fraction*.

[Hand, 2009] Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the roc curve. *Machine learning*, 77(1):103–123.

[Hastie, 2009] Hastie, T. (2009). *The elements of statistical learning : data mining, inference, and prediction*. Springer, New York.

[Hastie et al., 1995] Hastie, T., Buja, A., and Tibshirani, R. (1995). Penalized discriminant analysis. *The Annals of Statistics*, pages 73–102.

[Hastie et al., 1994] Hastie, T., Tibshirani, R., and Buja, A. (1994). Flexible discriminant analysis by optimal scoring. *Journal of the American statistical association*, 89(428):1255–1270.

[Hastie et al., 2015] Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with Sparsity : the lasso and generalizations*. CRC Press LLC, Boca Raton.

[Heinz et al., 2001] Heinz, D. C. et al. (2001). Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery. *IEEE transactions on geoscience and remote sensing*, 39(3):529–545.

[Hsieh and Lee, 2000] Hsieh, P.-F. and Lee, L.-C. (2000). Effect of spatial resolution on classification error in remote sensing. In *Geoscience and Remote Sensing Symposium, 2000. Proceedings. IGARSS 2000. IEEE 2000 International*, volume 1, pages 171–173. IEEE.

[Hubert et al., 2005] Hubert, M., Rousseeuw, P. J., and Vanden Branden, K. (2005). Robpca: a new approach to robust principal component analysis. *Technometrics*, 47(1):64–79.

[Hyvarinen et al., 2009] Hyvarinen, A., Hurri, J., and Hoyer, P. O. (2009). *Natural image statistics : a probabilistic approach to early computational vision*.

Springer-Verlag, London.

[John, 1961] John, S. (1961). Errors in discrimination. *The Annals of Mathematical Statistics*, pages 1125–1144.

[Johnson and Wichern, 2007] Johnson, R. and Wichern, D. (2007). *Applied Multivariate Statistical Analysis*. Applied Multivariate Statistical Analysis. Pearson Prentice Hall.

[Karoui et al., 2012] Karoui, M. S., Deville, Y., Hosseini, S., and Ouamri, A. (2012). Blind spatial unmixing of multispectral images: New methods combining sparse component analysis, clustering and non-negativity constraints. *Pattern Recognition*, 45(12):4263–4278.

[Kay, 1999] Kay, J. W. (1999). *Statistics and neural networks : advances at the interface*. Oxford University Press, Oxford New York.

[Keshava and Mustard, 2002] Keshava, N. and Mustard, J. F. (2002). Spectral unmixing. *IEEE signal processing magazine*, 19(1):44–57.

[Kessy et al., 2018] Kessy, A., Lewin, A., and Strimmer, K. (2018). Optimal whitening and decorrelation. *The American Statistician*, pages 1–6.

[Kim et al., 2001] Kim, D.-J., Park, Y.-W., and PARK, D.-J. (2001). A novel validity index for determination of the optimal number of clusters. *IEICE Transactions on Information and Systems*, 84(2):281–285.

[Kuhn and Johnson, 2013] Kuhn, M. and Johnson, K. (2013). *Applied predictive modeling*, volume 26. Springer.

[Kuroda, 2008] Kuroda, K. (2008). *Physiological Incidences Related to Symptom Development and Wilting Mechanism*, pages 204–222. Springer Japan, Tokyo.

[Latty and Hoffer, 1981] Latty, R. S. and Hoffer, R. M. (1981). Computer-based classification accuracy due to the spatial resolution using per-point versus per-field classification techniques. In *LARS Symposia*, page 448.

[Lobo et al., 2008] Lobo, J. M., Jiménez-Valverde, A., and Real, R. (2008). Auc: a misleading measure of the performance of predictive distribution models. *Global ecology and Biogeography*, 17(2):145–151.

[Louis et al., 2016] Louis, J., Debaecker, V., Pflug, B., Main-Knorn, M., Bieniarz, J., Mueller-Wilm, U., Cadau, E., and Gascon, F. (2016). Sentinel-2 sen2cor: L2a processor for users. In *Proceedings Living Planet Symposium 2016*, pages

1–8. Spacebooks Online.

[Luo et al., 2016] Luo, G., Chen, G., Tian, L., Qin, K., and Qian, S.-E. (2016). Minimum noise fraction versus principal component analysis as a preprocessing step for hyperspectral imagery denoising. *Canadian Journal of Remote Sensing*, 42(2):106–116.

[Ma et al., 2014] Ma, W.-K., Bioucas-Dias, J. M., Chan, T.-H., Gillis, N., Gader, P., Plaza, A. J., Ambikapathi, A., and Chi, C.-Y. (2014). A signal processing perspective on hyperspectral unmixing: Insights from remote sensing. *IEEE Signal Processing Magazine*, 31(1):67–81.

[Manolakis, 2016] Manolakis, D. (2016). *Hyperspectral imaging remote sensing : physics, sensors, and algorithms*. Cambridge University Press, Cambridge, United Kingdom.

[Manolakis et al., 2003] Manolakis, D., Marden, D., and Shaw, G. A. (2003). Hyperspectral image processing for automatic target detection applications. *Lincoln laboratory journal*, 14(1):79–116.

[Manolakis and Shaw, 2002a] Manolakis, D. and Shaw, G. (2002a). Detection algorithms for hyperspectral imaging applications. *IEEE signal processing magazine*, 19(1):29–43.

[Manolakis et al., 2014] Manolakis, D., Truslow, E., Pieper, M., Cooley, T., and Brueggeman, M. (2014). Detection algorithms in hyperspectral imaging systems: An overview of practical algorithms. *IEEE Signal Processing Magazine*, 31(1):24–33.

[Manolakis and Shaw, 2002b] Manolakis, D. G. and Shaw, G. A. (2002b). Directionally constrained or constrained energy minimization adaptive matched filter: Theory and practice. In *Imaging Spectrometry VII*, volume 4480, pages 57–65. International Society for Optics and Photonics.

[Murphy, 2012] Murphy, K. (2012). *Machine learning : a probabilistic perspective*. MIT Press, Cambridge, Mass.

[Nadeau and Bengio, 2000] Nadeau, C. and Bengio, Y. (2000). Inference for the generalization error. In *Advances in neural information processing systems*, pages 307–313.

[NASA, n.d.] NASA (n.d.). Measuring vegetation (ndvi and evi). `https://earthobservatory.nasa.gov/Features/MeasuringVegetation/measuring_vegetation_2.php` [Accessed: 19.11.17].

[NEA, 2017] NEA (2017). Invasive alien species. http://www.environment.no/topics/biodiversity/species-in-norway/invasive-alien-species/ [Accessed: 10.03.18].

[Neto and Dougherty, 2015] Neto, U. M. B. and Dougherty, E. R. (2015). *Error estimation for pattern recognition*. John Wiley & Sons.

[Nielsen, 2001] Nielsen, A. A. (2001). Spectral mixture analysis: Linear and semi-parametric full and iterated partial unmixing in multi-and hyperspectral image data. *Journal of Mathematical Imaging and Vision*, 15(1-2):17–37.

[Ortiz et al., 2013] Ortiz, S. M., Breidenbach, J., and Kändler, G. (2013). Early detection of bark beetle green attack using terrasar-x and rapideye data. *Remote Sensing*, 5(4):1912–1931.

[Palsson et al., 2012] Palsson, F., Sveinsson, J. R., Benediktsson, J. A., and Aanaes, H. (2012). Classification of pansharpened urban satellite images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(1):281–297.

[Pissanetzky, 1984] Pissanetzky, S. (1984). *Sparse Matrix Technology-electronic edition*. Academic Press.

[Preparata and Shamos, 2012] Preparata, F. P. and Shamos, M. I. (2012). *Computational geometry: an introduction*. Springer Science & Business Media.

[Remesan and Mathew, 2014] Remesan, R. and Mathew, J. (2014). *Hydrological data driven modelling: a case study approach*, volume 1. Springer.

[Resmini et al., 1997] Resmini, R., Kappus, M., Aldrich, W., Harsanyi, J., and Anderson, M. (1997). Mineral mapping with hyperspectral digital imagery collection experiment (hydice) sensor data at cuprite, nevada, usa. *International Journal of Remote Sensing*, 18(7):1553–1570.

[Richmond, 1996] Richmond, C. D. (1996). Derived pdf of maximum likelihood signal estimator which employs an estimated noise covariance. *IEEE Transactions on Signal Processing*, 44(2):305–315.

[Seber and Lee, 2012] Seber, G. A. and Lee, A. J. (2012). *Linear regression analysis*, volume 329. John Wiley & Sons.

[SIC, n.d.-a] SIC, S. (n.d.-a). *RapidEye Satellite Sensors (5m)*. https://www.satimagingcorp.com/satellite-sensors/other-satellite-sensors/rapideye/ [Accessed: 20.03.18].

[SIC, n.d.-b] SIC, S. (n.d.-b). *WorldView-4 Satellite Sensor (0.31m)*. https://www.satimagingcorp.com/satellite-sensors/geoeye-2/ [Accessed: 20.03.18].

[Sigreaves, 1961] Sigreaves, R. (1961). Some results on the distribution of the w-classifications statistics. *Studies in Item Selection and Prediction*, pages 241–261.

[Sjöstrand et al., 2012] Sjöstrand, K., Clemmensen, L. H., Larsen, R., and Ersbøll, B. (2012). Spasm: A matlab toolbox for sparse statistical modeling. *Journal of Statistical Software Accepted for publication*.

[Stark and Woods, 2014] Stark, H. and Woods, J. (2014). *Probability and Random Processes with Applications to Signal Processing: International Edition*. Pearson Education Limited.

[Sun et al., 2014] Sun, K., Geng, X., Ji, L., and Lu, Y. (2014). A new band selection method for hyperspectral image based on data quality. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(6):2697–2703.

[Theodoridis, 2009] Theodoridis, S. (2009). *Pattern recognition*. Academic Press, Burlington, MA London.

[Treitz et al., 1992] Treitz, P. M., Howarth, P. J., Suffling, R. C., and Smith, P. (1992). Application of detailed ground information to vegetation mapping with high spatial resolution digital imagery. *Remote Sensing of Environment*, 42(1):65–82.

[Weisstein, n.d.-a] Weisstein, E. W. (n.d.-a). *Face*. http://mathworld.wolfram.com/Face.html [Accessed: 20.04.18].

[Weisstein, n.d.-b] Weisstein, E. W. (n.d.-b). *Simplex*. http://mathworld.wolfram.com/Simplex.html [Accessed: 02.04.18].

[Witten and Tibshirani, 2011] Witten, D. M. and Tibshirani, R. (2011). Penalized classification using fisher's linear discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):753–772.

[Zare and Ho, 2014] Zare, A. and Ho, K. (2014). Endmember variability in hyperspectral analysis: Addressing spectral variability during spectral unmixing. *IEEE Signal Processing Magazine*, 31(1):95–104.

[Zhu et al., 2014a] Zhu, F., Wang, Y., Fan, B., Meng, G., and Pan, C. (2014a).

Effective spectral unmixing via robust representation and learning-based sparsity. *CoRR*, abs/1409.0685.

[Zhu et al., 2014b] Zhu, F., Wang, Y., Fan, B., Meng, G., Xiang, S., and Pan, C. (2014b). Spectral unmixing via data-guided sparsity. *CoRR*, abs/1403.3155.

[Zhu et al., 2014c] Zhu, F., Wang, Y., Xiang, S., Fan, B., and Pan, C. (2014c). Structured sparse method for hyperspectral unmixing. *ISPRS Journal of Photogrammetry and Remote Sensing*, 88:101–118.

[Zou and Hastie, 2005a] Zou, H. and Hastie, T. (2005a). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

[Zou and Hastie, 2005b] Zou, H. and Hastie, T. (2005b). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.