



**ON THE DEVELOPMENT OF ROBUST ANOMALY DETECTION
ALGORITHMS WITH LIMITED LABELED DATA**

DISSERTATION

Adam J. Messer, Major, USAF

AFIT-ENS-DS-17-S-040

**DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY**

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

DISTRIBUTION STATEMENT A. APPROVED FOR PUBLIC RELEASE;
DISTRIBUTION UNLIMITED.

The views expressed in this dissertation are those of the author and do not reflect the official policy or position of the United States Air Force, the Department of Defense, or the United States Government. This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

AFIT-ENS-DS-17-S-040

ON THE DEVELOPMENT OF ROBUST ANOMALY DETECTION ALGORITHMS
WITH LIMITED LABELED DATA

DISSERTATION

Presented to the Faculty
Department of Operational Sciences
Graduate School of Engineering and Management
Air Force Institute of Technology
Air University
Air Education and Training Command
in Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy in Operations Research

Adam J. Messer, B.A., M.S.

Major, USAF

September 2017

DISTRIBUTION STATEMENT A. APPROVED FOR PUBLIC RELEASE;
DISTRIBUTION UNLIMITED.

AFIT-ENS-DS-17-S-040

ON THE DEVELOPMENT OF ROBUST ANOMALY DETECTION ALGORITHMS
WITH LIMITED LABELED DATA

Adam J. Messer, B.A., M.S.
Major, USAF

Committee Membership:

Kenneth W. Bauer, Jr., Ph.D.
Chair

James W. Chrissis, Ph.D.
Member

Mark E. Oxley, Ph.D.
Member

Adedeji B. Badiru, Ph.D.
Dean, Graduate School of Engineering
and Management

Abstract

Anomaly Detectors are used in a number of multivariate applications to identify data that do not belong within a dataset. A significant problem in the development of anomaly detection algorithms is the insufficient availability of truth-labeled data. This can lead to instances where the algorithms might not generalize to real-world operation scenarios.

This research focuses on the sensitivity of anomaly detection algorithms to the anomalies present. First, a method to gradually modify the multivariate characteristics of the anomalies to approach the structure of the background is developed. In doing so, algorithms are tested with the expectation that gradual degradation in performance will occur as the anomalies become more difficult to differentiate from the background. This allows for a comparison between algorithms with both graphical and numerical techniques. The anomaly adaptation approach is extended into a robust parameter design framework to identify robust algorithm parameters that take into account the many possible anomalies present in extended operation scenarios.

Additionally, a novel method for the selection of anomaly thresholds from possibly anomaly-contaminated datasets is presented. This method is shown to have superior performance on both simulated and real-world data. The method is also used to create new versions of a number of popular anomaly detectors in Hyperspectral Images.

To my wife and parents.

Acknowledgments

I would be remiss to claim that this effort didn't come to fruition through much assistance of others.

First and foremost, Dr. Bauer, the value of his guidance through this whole process is unmeasurable. Whether it was reining in the leash or letting it spool out, he alone presumably kept me from a much longer research effort. I'd also like to acknowledge my research committee, as well as the other faculty, staff, and students of AFIT.

I would also like to thank my family, friends, and colleagues (especially those at USAFE/A9A) that convinced me coming back to school was the right choice. Especially those that called me crazy for considering the other paths life was offering at the time. Time will tell whether it was truly the right choice, but it's looking good so far.

A thanks to the many people that offered editing help. Though I'd like to note that any mistakes still contained herein are my own.

Adam J. Messer

Table of Contents

	Page
Abstract	iv
Dedication	v
Acknowledgments	vi
Table of Contents	vii
List of Figures	x
List of Tables	xii
I. Introduction	1
1.1 Research Contributions & Chapter Overview	2
II. Literature Review	4
2.1 Anomaly Detection	4
2.2 Dimensionality Reduction	5
2.2.1 Principal Component Analysis	6
2.2.2 Factor Rotation	8
2.2.3 Independent Component Analysis	8
2.2.4 Mixture of Factor Analyzers	9
2.3 Remote Sensing using Hyperspectral Imaging	9
2.4 Anomaly Detection Methods in Hyperspectral Imaging (HSI)	11
2.4.1 Topographical Anomaly Detector (TAD)	11
2.4.2 Blocked Adaptive Computationally Efficient Outlier Nominators (BACON)	12
2.4.3 RX and its Extensions	13
2.4.4 Support Vector Data Description	15
2.4.5 Methods using Component Analysis	16
2.4.6 Anomaly Detection in HSI using Random Fields	18
2.4.7 Spatially Dependent Methods	19
2.5 Quantifying Performance	22
2.5.1 Thresholding Methods	24
2.6 Collected Images	26

	Page
III. Method of Sensitivity Analysis in Anomaly Detection Algorithms for Hyper-spectral Images	28
3.1 Introduction	28
3.2 Background	28
3.2.1 Anomaly Detection Methods for HSI	29
3.2.2 Comparing Anomaly Detectors	30
3.2.3 Mahalanobis Distance	31
3.3 Methodology	32
3.3.1 Evaluation of Sensitivity	34
3.4 Results	35
3.4.1 Qualitative Results	38
3.4.2 Quantitative Results	39
3.5 Conclusion	44
IV. Identification of Robust Parameters using Mahalanobis Masking	45
4.1 Introduction	45
4.2 Robust Parameter Design	46
4.2.1 Multivariate Delta Method	48
4.2.2 Noise Framework for Robust Parameter Design (RPD) in Anomaly Detection	50
4.3 Mahalanobis Masking	52
4.4 Noise Framework Incorporating Mahalanobis Masking	54
4.4.1 Mahalanobis Masking as Adversarial Behavior	56
4.5 Measuring Performance	57
4.6 Images	57
4.7 RPD Experiments using Autonomous Global Anomaly Detector (AutoGAD)	57
4.7.1 Neural Network Structure	60
4.7.2 Optimal Settings	61
4.7.3 Test Image Performance	63
4.8 Conclusions	65
V. Truncated Extreme Value Anomaly Thresholding	66
5.1 Introduction	66
5.2 Background	69
5.2.1 Extreme Value Theory	69
5.2.2 Truncated Distributions	71
5.2.3 Goodness-of-Fit Tests	73
5.3 Methodology	78
5.3.1 Algorithm Overview	78

	Page
5.3.2 Power of GOF Tests for Detecting Anomalies	82
5.4 Results	83
5.4.1 Simulated Output	84
5.4.2 Detector Output from Hyperspectral Images	86
5.5 Conclusions	88
VI. Extreme Value Thresholding as Mitigation to Correlation Effects in Anomaly Detection	89
6.1 Introduction	89
6.2 Algorithms	90
6.2.1 Extreme (Ext) RX	91
6.3 Methodology	93
6.4 Results	94
6.5 Conclusions	97
VII. Conclusion	98
7.1 Original Contributions	98
7.2 Suggested Future Work	99
Appendix A: Derivation of Formulas	100
Appendix B: Quantitative Sensitivity Analysis Charts	109
Appendix C: List of Acronyms	119
Bibliography	121
Vita	130

List of Figures

Figure	Page
2.1 Plate Model for FA	7
2.2 Example of a hyperspectral image cube.	10
2.3 Windowing Techniques	15
2.4 AutoGAD four-phased approach [49].	17
2.5 ROC Curve Example	25
3.1 Spectral Signature Nearest Neighbor Example	32
3.2 Impact of Mahalanobis Masking on an anomalous pixel's spectral signature.	33
3.3 Notional Performance of Algorithms after Iterative Mahalanobis Masking. Each subfigure is the performance of the algorithm's on a different image.	40
4.1 Mahalanobis Masking creating variation of the original anomalous signals (x) both more (x^-) and less (x^+) difficult.	54
4.2 Impact of Control Variable x_2 at robust parameter setting.	62
4.3 Impact of Control Variable x_3 at robust parameter setting.	63
4.4 Impact of Control Variable x_4 at robust parameter setting.	63
4.5 Impact of Control Variable x_5 at robust parameter setting.	64
5.1 Example of Impact on Distribution Fit with and without accounting for truncation.	73
5.2 The three charts show the effect of ignoring the truncation of data when doing Goodness-of-Fit (GoF) testing.	77
5.3 Example Relative Placement of u , v , and CFAR estimates.	79
5.4 Flowchart depiction of Thresholding Algorithm.	80
5.5 Power of GoF tests on simulated Swerling type 0 data.	83
5.6 MSE of CFAR threshold estimate on simulated Swerling type 0.	85

Figure	Page
5.7 MSE of threshold estimate on RX scores from HYDICE images.	88
6.1 Sample of RX scores from LRX on ARES5F.	92
B.1 Mahalanobis Masking on RX	110
B.2 Mahalanobis Masking on HLRX	111
B.3 Mahalanobis Masking on LRX	112
B.4 Mahalanobis Masking on ILRX	113
B.5 Mahalanobis Masking on TAD	114
B.6 Mahalanobis Masking on AutoGAD	115
B.7 Mahalanobis Masking on GFAAD	116
B.8 Mahalanobis Masking on MultiPCA	117
B.9 Mahalanobis Masking on SVDD	118

List of Tables

Table	Page
2.1 Confusion Matrix Layout.	22
2.2 HYDICE Image Properties.	27
3.1 Initial TPF Performance of Algorithms.	37
3.2 Initial FPF Performance of Algorithms.	37
3.3 Mean TPF of Algorithms across 500 iterations of Masking.	41
3.4 Mean FPF of Algorithms across 500 iterations of Masking.	41
3.5 Variance TPF of Algorithm across 500 iterations of Masking.	42
3.6 Mean Squared Error of TPF of Algorithm across 500 iterations of Masking. . .	42
4.1 Calculated noise variables of images from Mindrup <i>et al.</i> calculated on the training and test image sets [69].	58
4.2 Training images with new calculated noise variables.	59
4.3 AutoGAD parameters and bounds.	60
4.4 Recommended robust operating points for AutoGAD as determined without masking and with Mahalanobis Masking.	61
4.5 Test images with new calculated noise variables.	64
4.6 Results of robust operating points on test images.	65
5.1 tGOF Test Statistics	75
5.2 tGOF Test Statistics	78
6.1 Spectral bands removed due to indication of atmospheric absorption.	93
6.2 FPF Performance on Images at Desired FAR of 1%.	95
6.3 TPF Performance on Images at Desired FAR of 1%.	95
6.4 FPF Performance on Images at Desired FAR of 5%.	95
6.5 TPF Performance on Images at Desired FAR of 5%.	95

Table	Page
6.6 FPF Performance on Images at Desired FAR of 10%.	96
6.7 TPF Performance on Images at Desired FAR of 10%.	96
6.8 Variance of FPF at Desired FAR of 10% (Bolded value represent statistical significance at a p-value of 0.05).	97

ON THE DEVELOPMENT OF ROBUST ANOMALY DETECTION ALGORITHMS WITH LIMITED LABELED DATA

I. Introduction

The information age has seen the development of processes that surpass the human capability to carry them out. To this end, a number of approaches using computational techniques to aid in the processes have been developed. The effects are evident in application areas from internet search engines and medical imaging to the presence of robots in many industrial production lines. One of the most significant areas of application for computational techniques is that of anomaly detection. Anomaly detection is concerned with finding the subset of a dataset that does not conform to expected behavior, similar to its kin: outlier detection and novelty detection [22]. Primarily, anomalies can be placed in two over-arching categories that stem from the reason for their anomalousness. The first, which is often the uninteresting category, comes from measurement error in the data. For example, if the height of 30 people are collected but the decimal point for one measurement is neglected, labeling said point as 19 meters instead of 1.9 meters, the data point will be anomalous. However, finding the anomaly only tells us our data was collected incorrectly; not something significant about the behavior underlying the dataset. The second source for anomalies is out-of-class behavior and is normally the category that most interests researchers since these are truly displaying unexpected behavior. Detecting this unexpected behavior is the objective of anomaly detection. However, it is often not possible to discern between the two categories without an expert opinion or additional information. The field of anomaly detection is applied in a number of areas to include

medical monitoring, machine fault detection, and meteorological phenomena. This dissertation is focused on anomaly detection via of multivariate statistics.

Hyperspectral Imaging (HSI) is one of the most promising areas of anomaly detection application to emerge in the past few decades. In HSI, an image, or scene, is comprised of the electromagnetic spectral signatures of its spatial representation. This can be thought of as an extension from a normal color photograph, which normally is measured in three spectral bands (red, green, and blue), but instead the hyperspectral image measures across contiguous spectral bands from near infrared to the visible color spectrum. The additional information collected allows for much higher discrimination between materials in the photograph. Anomaly detection in HSI is an active field of research with a number of different approaches in the literature. This dissertation concentrates on pure, unsupervised anomaly detection approaches that operate with no *a priori* knowledge of anomalous spectral signatures. An unsupervised approach allows for anomaly detection in a wider range of operating conditions provided by background adaptability.

1.1 Research Contributions & Chapter Overview

The focus of this dissertation is on the techniques and approaches taken to develop quality anomaly detection algorithms. The first goal is to provide an approach for gleaning additional knowledge about anomaly detectors when facing scarcity of truth-labeled training data. The second goal is to extend this new knowledge to the selection of algorithm parameters in a systematic fashion. The third goal is to establish an anomaly thresholding method that efficiently returns a desired performance characteristic to algorithms.

Chapter 2 presents common literature to subsequent chapters such as a background on statistical techniques used, HSI, and anomaly detection.

In Chapter 3, a novel approach for conducting sensitivity analysis on anomaly detectors is presented. In the field of anomaly detection, the availability of truth-labeled

data for performance evaluation can be severely limited. By creating variations of this real-world data, one can characterize the performance of algorithms in progressively more difficult datasets, thus highlighting any sensitivity to the anomalies' characteristics.

In Chapter 4, the work of Chapter 3 is further developed by extending previous work on noise framework development for Robust Parameter Design (RPD) approaches to HSI. This allows for the identification of algorithm parameters for algorithms in a manner that uses an extended dictionary of anomalies.

Chapter 5 presents Truncated Extreme Value Anomaly Thresholding (TEVAT), an improved method for anomaly threshold identification from the output of anomaly detectors. This method is based on the field of Extreme Value Theory.

In Chapter 6, the TEVAT algorithm is used to create new embellishments on a family of common anomaly detection algorithms in HSI. In doing so, the aim is to return Constant False Alarm Rate (CFAR) performance to the algorithm which can be lost due to spatial correlation effects.

In the final chapter closes with conclusions and suggestions for future research.

II. Literature Review

In this chapter, much of the literature that is common and pertinent to the subsequent chapters is presented. This chapter covers the background on anomaly detection, Hyperspectral Imaging (HSI), algorithms for anomaly detection in HSI, performance measurement approaches, and anomaly thresholding. However, literature that pertains only to individual chapters is contained therein. This includes much of the background on Mahalanobis Distance in Chapter 3, Robust Parameter Design in Chapter 4, and Extreme Value Theory, Truncated Distributions, and determination of Goodness-of-Fit for Distributions in Chapter 5.

2.1 Anomaly Detection

At its core, anomaly detection is meant to detect and identify patterns, objects, or observations that do not belong within the greater majority of the dataset [22]. Similar to outlier detection and novelty detection, anomaly detection has applications in a number of areas such as cyberspace intrusion detection, financial fraud detection, and medical monitoring [22]. There have been a number of excellent survey papers on anomaly, outlier, and/or novelty detection (see [22, 79]).

While approaches to anomaly detection may be similar, application areas differ in the characteristics that make up an anomaly. For example, in cyberspace intrusion detection repeated requests for network access may either be a sign of an intrusion or just a very active user, whereas in medical monitoring the absence of a regular signal may be cause for concern. When viewed from a statistical standpoint, anomaly detection is concerned with the nature of the background data in order to statistically discern the behavior of anomalies. Thus, once a statistical understanding of the background is gained, anomaly detection is disconnected from any explicit application.

An important distinction between anomaly detection algorithms is that of supervised, semi-supervised, and unsupervised methods [22]. In supervised methods, algorithms are allowed to use truth-labeled data, indicating both background and anomalous classes, to train, or set parameter values. This is often problematic in application areas where labeling data is difficult, time-consuming, or expensive. Similarly, semi-supervised methods are provided with a dataset known to only contain background data. This allows algorithms to first train on uncontaminated data to establish true behavior characteristics of the background. The third category—unsupervised methods—are not provided with truth labels at all, exclusively using datasets that may already contain anomalies to determine the statistical properties of the background. These rely on processing steps within the algorithm to differentiate the anomalies from the background and offer greater applicability to data with background behavior that changes between datasets.

2.2 Dimensionality Reduction

While the rapid increase of available data facilitates an ability to conduct research in many new areas, the sheer amount of data can overwhelm classical methods. For example, a linear least squares Maximum Likelihood Estimation (MLE) approach to regression is a classical method with many desirable characteristics such as representation as orthogonal projections into subspaces, statistically testable coefficients, and a guaranteed solution; however, the closed-form solution of the coefficients requires a matrix inversion operation that can be unwieldy when the data being investigated is very large. Additionally, for certain statistical properties to hold, the least squares method requires uncorrelated input vectors. This requirement is not often present in data collected outside a controlled experiment, nor when the data are spatially or temporally correlated.

A family of methods that can help overcome this is known as Factor Analysis. Factor Analysis (FA) is interested in finding a simpler structure explaining the observed data. This serves three principal purposes [29]:

1. Reducing the overall number of independent variables that can then be used in subsequent methods in a way that retains most of the variance in the data;
2. Providing a means for analyzing data when the size of the dataset is such that direct analysis is beyond comprehension;
3. Testing a hypothesis about the generating characteristics of the data.

The first two purposes of FA are the most applicable in our research. For example, in a hyperspectral image containing tens of thousands of pixels, each with upwards of 250 spectral band measurements, most statistical methods for parameter estimation are outpaced. The application of FA type techniques for dimensionality reduction using methods such as Principal Component Analysis (PCA) and Independent Component Analysis (ICA) can overcome the size and correlation structure of the data. To describe the methods of FA, we must first establish a baseline structure. The general structure of FA assumes we have some i -th observation of a latent random vector, z_i of size k , giving rise to our observable random vector, x_i , of size d . The linear relationship between z_i and x_i is given by some matrix W of size $d \times k$ where $k < d$. The covariance of our k factors is represented by the diagonal matrix Ψ . The plate diagram for this FA model is shown in Figure 2.1.

The distributions of z_i and x_i give rise to different forms of FA. The most straight forward are Gaussian distributions where

$$p(z_i) = \mathcal{N}(z_i | \mu_0, \Sigma_0). \quad (2.1)$$

2.2.1 Principal Component Analysis.

Although not completely confined as an FA method, PCA may, to some extent, be viewed within the family of FA methods as an initial step in Exploratory Factor Analysis (EFA). The main assumptions that shifts the FA model from the above

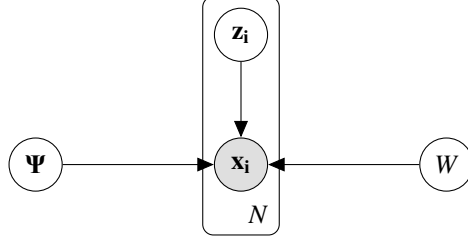


Figure 2.1: Plate Model for FA with observed variable $\mathbf{x}_i$ as it arises from components $\mathbf{z}_i$ with mixing matrix W . The covariance of the observations Ψ is normally assumed to be a diagonal matrix indicating no correlation between the N observations. Adapted from [71].

generalized presentation to that of a PCA model are (1) that the diagonal covariance matrix of the observations, Ψ , is zero and (2) that the projection matrix $\mathbf{W}$ is orthogonal. With these two assumptions, FA is identical to PCA. There are a number of ways to view PCA, each presented and discovered via different viewpoints in the literature [71], the most valuable for our use is that of Singular Value Decomposition (SVD). The goal of SVD is to decompose a matrix C of size $d \times d$ into three components, V, S , and V^T , such that

$$C = V^T \Lambda V \quad (2.2)$$

where Λ is a diagonal matrix of size $d \times d$, and V is a orthonormal matrix of size $d \times d$. For the purposes of PCA, the matrix C can be either the correlation matrix or covariance matrix of the centered data, $\hat{\mathbf{X}}_c = \mathbf{X} - \hat{\mu}_X$, from the original data matrix, $\mathbf{X}$. By sorting the diagonal elements of Λ —the eigenvalues, or characteristic values, of C , $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$ —and the corresponding rows of V , k -dimensional linear subspaces of our data can be selected in ways that retain the maximum amount of variance of X , where $k < d$, by selecting the first k values in Λ .

Estimators from SVD give rise to matrices that can then be shown to minimize reconstructive error for the PCA model [71]. Reconstruction error is the error in the

estimation of a data point, $\hat{X}_c$, using only the principal component vector, V_k ,

$$\|\hat{X}_c - V_k^T Y\|^2 \quad (2.3)$$

using the nearest estimate, Y , formed using its principal component scores

$$Y = \hat{X}_c V_k = (X - \mu_X) V_k. \quad (2.4)$$

The properties of SVD allow for estimation of the projection matrix, V_k , once we have an estimate of C without performing iterations with the entire dataset.

2.2.2 Factor Rotation.

During EFA, PCA is often a first step toward finding an appropriately-sized subspace to project further. Whereas PCA identifies the sorted orthogonal factors that explain the most variance, we can find factors with different characteristics through a process known as Factor Rotation [29]. For example, a popular approach known as VariMax rotation aims to maximize the amount of variance for each variable explained by a single factor. Using the covariance or correlation matrix, VariMax maximizes the squared loadings matrix while retaining the orthogonality of the components [50]. There are other methods for factor rotation (see [71]).

2.2.3 Independent Component Analysis.

Another approach to factor analysis is known as Independent Component Analysis (ICA). In ICA, a number of independent signals are sought to be unmixed from the noisy, mixed signal. This departs from PCA in that there is a stronger requirement that components be independent and not simply uncorrelated. It is often used in blind source signal recovery where the signals are being mixed and additive noise is present [25]. FastICA is a common method for finding the ICA rotation matrix and the signals [45]. There are additional methods that are more appropriate in certain scenarios, such as additive Gaussian noise [101].

2.2.4 Mixture of Factor Analyzers.

Typically, traditional factor analysis assumes the signal has a constant mean vector, which is often removed during a whitening process. However, there are situations where subsets of data present differing mean vectors. In these situation, a mixture of factors model is appropriate, where each factor contributing to the observed signals is allowed to have a different mean vector. The mixture of factor analyzers approach prevents misrepresentation of subpopulations in the data as factors in a single dataset. To find our source signals, a mixture of factor analyzers is used to determine both the mean vectors and the correlation structures of the submodels.

2.3 Remote Sensing using Hyperspectral Imaging

Remote sensing is the use of signals taken from a distance, such as signals from satellites or high-flying aircraft, to generate information. One subset of remote sensing involves the use of Hyperspectral Imaging (HSI). HSI is best understood by comparing it to simple color images taken by a digital camera. A color image is made up of three separate intensity images measured across the color bands corresponding to red, green, and blue. The intensity values in these component images represent the intensity of the color for each pixel. Therefore, a pixel is a representation of the spatial area from which the signal was collected. When the images are mixed together, a color image is created. HSI is similar to this, but the number of intensity images is much greater with each of them representing a much narrower spectral band. The spectral bands normally sampled can range from infrared to near ultraviolet (about 400-2500 nanometers (nm) wavelengths) with upwards of 250 separate intensity images. A useful visualization and data structure for HSI is known as an image cube. An image cube is three-dimensional with two dimensions being spatial and the third spectral. Any image—or in the geospatial application, map— can be selected from a particular spectral band, represented by the map

to the left in Figure 2.2. Alternatively, selecting a certain pixel, or spatial location, the spectral signature can be retrieved, as in the right pull out in Figure 2.2.

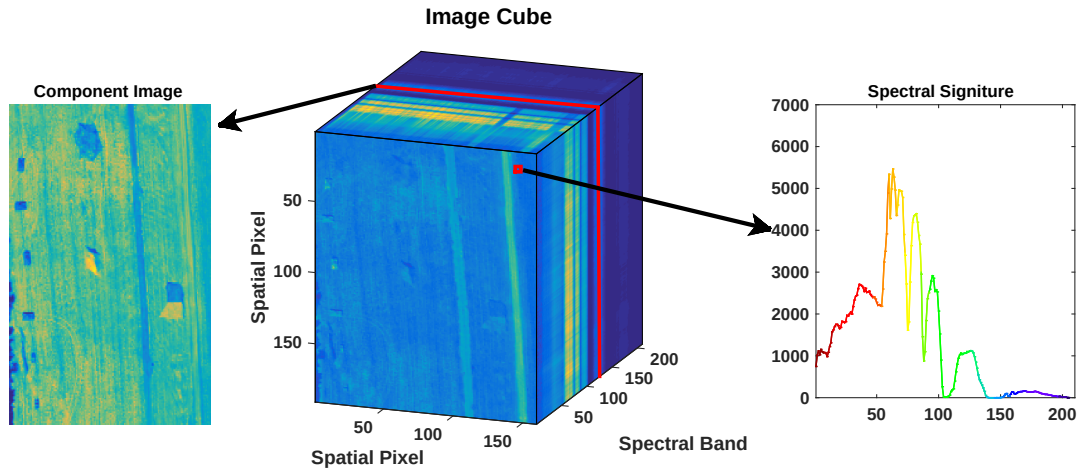


Figure 2.2: Example of a hyperspectral image cube.

As the spatial and spectral resolution of an HSI sensor are increased, a number of new application areas are made available. For example with increased spectral resolution, specific materials can often be determined, whereas, in a normal color image only the color of objects can be identified. This is known as spectral matching and can be a valuable application in surveillance activities. Examples of its application are numerous [38].

Another benefit of HSI is detecting the presence of regions of interest that do not appear to belong in the imaged area. This has applications in the automation of object identification. In this manner a large amount of images can be analyzed with minimal input from a human. By creating such a process, vast amounts of imaging data can be reduced significantly for subsequent, in-depth analysis by a human analyst. This greatly increases the value of HSI in remote sensing, with even greater proliferation of sensors as a by-product of decreased cost.

2.4 Anomaly Detection Methods in HSI

Anomaly detection in HSI is performed through a number of varying methods, each with their own characteristics. Perhaps the largest bifurcation of methods is the distinction between detectors using the entire image for analysis, known as global detectors, and those using only small windows around pixels to declare anomalies, known as local detectors. Having the statistical properties of the spectral signatures of the entire image to aid in the declaration of anomalies, provides global detectors an advantage as this provides a far more accurate depiction of the background distribution, or non-anomalous, pixels. However, this comes at an increased computational burden, apparent during the analysis of large images.

Local detectors use a limited window around each pixel to compare spectral signatures instead. This greatly reduces the computational requirement if the per pixel computational requirement is low, but in regards to image size the per pixel nature of local detectors may also drive an exponential growth in computational complexity. Local detectors are more susceptible to regions of increased variance that cover up the presence of anomalies since they lack the information of the entire scene [92].

In the following sections, an overview is presented of a selection of prominent anomaly detection algorithms for HSI. These are only a selection of the multitude of algorithms in the literature, but they provide a general indication of the different approaches discussed in later chapters.

2.4.1 *Topographical Anomaly Detector (TAD).*

The Topological Anomaly Detector (TAD) is a global anomaly detector developed on the assumption that points in the background class should display similar signatures. Thus, when measured with a distance function, the background should display lower distances between points due to their similarity than those of an anomaly [8]. In a graph theoretic form of the data let G_r denote the graph where each vertex represents a pixel. An

edge between two pixels occurs if the distance between two pixels is less than some $r > 0$. As the value of r increases, the number of connected pixels is expected to increase and G_r to become denser in its connections. For an appropriate value of r , anomalies should display lower connectiveness than background pixels. However, finding this value of r is dependent on the image and sensor characteristics. A slight modification to the metric uses the mean distance to the closest subset of pixels, for example the third, fourth, and fifth closest pixels [9]. This provides computational improvements over the former method since, by exploiting the triangle inequality property of the distance metric, the pairwise distances for the entire image is no longer required utilizing a bisection search approach [66]. The sensitivity of the detector is controlled by the size and magnitude of the subset indices. Employing a large set of indices would make the algorithm more robust to small deviations, and using higher magnitude indices would offer a larger guard against clustered anomalies.

2.4.2 Blocked Adaptive Computationally Efficient Outlier Nominators (BACON).

The Blocked Adaptive Computationally Efficient Outlier Nominators (BACON) method is based on the assumption that homogeneous multivariate data should display common variance characteristics [17]. However, in the presence of anomalies, it can be difficult to find the common characteristics needed to judge the data. As a remedy, the BACON algorithm iteratively judges the “outlyingness” of points based on their Mahalanobis distance to the estimated Gaussian distribution from the data. Starting with an initial basic subset assumed to belong to the background class, points are nominated to join the background class based on their statistical distance. The process iterates until the size of the background class no longer changes. The BACON algorithm is also a useful tool for robust regression [17].

2.4.3 RX and its Extensions.

One of the principal baseline anomaly detectors within HSI is the Reed–Xiaoli (RX) detector [81]. Developed as one of the field’s earliest anomaly detectors for multispectral images, the RX detector is a local detector that uses either the radiance values of the bands or the principal component scores of a subspace [86] to perform statistical analysis. Locality is provided by sweeping windows centered around each point within the image to estimate the mean, $\hat{\mu}$, and covariance, $\hat{\Sigma}$, excluding the Pixel Under Test (PUT). Control of bias towards regions of high variance is performed by selecting an appropriate size for the applicable window. Properties of the window are used to generate an RX score for the PUT, x :

$$\text{RX}(x) = (x - \hat{\mu})^T \left[\left(\frac{N}{N+1} \right) \hat{\Sigma} + \left(\frac{1}{N+1} \right) (x - \hat{\mu})(x - \hat{\mu})^T \right]^{-1} (x - \hat{\mu}) \quad (2.5)$$

where $\hat{\mu}$ and $\hat{\Sigma}$ are the mean and covariance, respectively. Then, with the assumption that all pixels are independent and Gaussian, scores are compared to a chi-squared (χ^2) statistic through pixel application of the Generalized Likelihood Ratio Test (GLRT) [81]. This assumption is necessary since the determination of the probability density function (pdf) of true background pixels within an image is not estimable. When the distribution assumption is reasonably met, the RX family of detectors have what is known as a Constant False Alarm Rate (CFAR) property [81]. This property implies that detectors are built in a probabilistic manner such that during operation the probability of a false alarm is constant and adjustable [70]. This allows for a detector to be adjusted according to a desirable level of performance.

As RX was originally developed for multispectral images (2-25 spectral bands for image), it faces computational issues in high spectral resolution HSI. As a result, a number of adaptations have been suggested to improve the performance of the RX detector. Application of the detector on subspaces determined from dimensionality reduction techniques have shown promise [86]. One of the simplest adaptations is to

delete the first few principal components before using the detector. This should reduce false positive declarations in what is known as Subspace RX (SSRX) [86]. However, elimination of specific components creates another anomaly detection problem, since components that differentiate between background and anomalies are then exploitable for discrimination [87].

While RX is generally accepted as a decent detector, it can have difficulty with anomalies that are spatially large (i.e. anomalies that cover a number of contiguous pixels). Since the covariance matrix calculation relies on neighboring pixels, spatially large anomalies may overwhelm the calculation and cause false negative declarations. One suggested remedy is to use a guard window that excludes the nearest pixels from the covariance and mean calculations [54]. Another adaptation is that of Linear RX (LRX), which uses sweeping lines instead of windows to calculate the background statistics [104]. This may also be overcome through iterative exclusion of potential anomalies from the statistics [96]; however, this comes at a computational cost by repeating the entire RX algorithm. Examples of these different windowing techniques are seen in Figure 2.3.

Another adaptation to RX that seeks to remove the effects of anomalies on the estimation of background statistics is the ensemble approach of the Topological Anomaly Detector-RX (TAD-RX) detector [7]. TAD-RX eliminates pixels from the background statistics calculation based on the points' topological characteristics. A topological characteristic is calculated using the Euclidean distance from one point to the another under consideration by using TAD. This is similar to the BACON approach, but with the Euclidean distance instead of the Mahalanobis distance. After excluding potential anomalies from the windows, the RX algorithm is processed on the image for determining background statistics and declaring anomalies. There are also RX modifications aimed at statistically removing the effects of the anomalies on the statistical estimates. One such

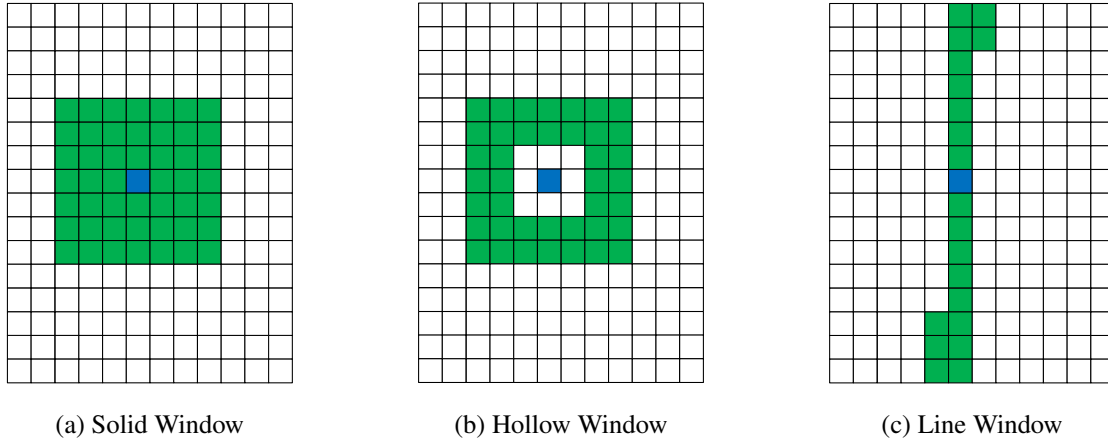


Figure 2.3: Examples of Windowing Techniques for the RX detector. The solid window technique in (a) surrounds the PUT with the window. The hollow window technique of (b) excludes the portion of pixels immediately around the PUT. The “lineing” technique of (c) uses a vertical line that wraps around the edges.

approach replaces the theoretical χ^2 background distribution with a mixture of gammas [33] at a increased computational cost.

The above approach of isolating potential anomalies from the windows of other pixels is extended in a detector known as Iterative RX (IRX) [96]. However, instead of using different detectors, IRX iteratively computes the RX scores of the pixels by excluding potential anomalies for each iteration. This process is allowed to run until either stabilization in anomaly identification or a maximum amount of iterations is reached.

2.4.4 Support Vector Data Description.

An anomaly detector based on a support vector approach [18] is known as Support Vector Data Description (SVDD) [5]. Using a possibly non-linear mapping function, $\Phi(x)$, this approach determines the set of support vectors needed to generate a measure that can be used to replace Mahalanobis distance of RX. Support vectors are determined

using the optimization problem using the data $x_i \in X$:

$$\begin{aligned} \min_{R,a} R^2 \\ \text{subject to } \|\Phi(x_i) - a\|^2 \leq R^2 \end{aligned} \quad (2.6)$$

where a is the center of the space, $\|\cdot\|$ is a norm (usually the 2-norm), $R > 0$ is some radius, and $\Phi(x)$ is some mapping from the input space to another space [5]. The complementary slackness condition of the Karush-Kuhn-Tucker (KKT) conditions, $(\|x_i - a\|^2 - R^2)\alpha_i = 0$, supplies the set of support vectors with non-zero α_i . After optimization, only the support vectors x_i with a corresponding non-zero α_i values are used in the function

$$\begin{aligned} \text{SVDD}(y) &= \|\Phi(y) - c\|^2 \\ &= \langle \Phi(y), \Phi(y) \rangle - 2 \sum_i (\alpha_i \langle \Phi(y), \Phi(x_i) \rangle) + \sum_{i,j} \alpha_i \alpha_j \langle \Phi(x_i), \Phi(x_j) \rangle. \end{aligned} \quad (2.7)$$

The original method relies on training a hyperparameter for the kernel function on labeled data. However, modifications have been proposed to create an unsupervised adaptation [74]. In this approach, BACON is used to identify the subset of data most likely of being background.

2.4.5 Methods using Component Analysis.

Another approach for anomaly detection lies in finding a set of linear subspaces that best separates the background class from the anomalies. When properly identified, projection of the data into a lower-dimensional subspace allows for a straight-forward approach to anomaly detection. However, determining which subspaces separate background and anomalies is akin to the original problem [87]. A number of methods have been developed that essentially follow the same multivariate analysis process to reach anomaly declarations within HSI: Autonomous Global Anomaly Detector (AutoGAD) [49], Multiple PCA Anomaly Detector (MultiPCA) [46], and Global Factor Analysis-based Anomaly Detector (GFAAD) [74].

Early work that used component analysis focused on PCA starting color image processing of Jackson and Morris, which used the residuals of the subspace projection [48]. This residual-based approach to anomaly detection has been extended to HSI [47, 80].

2.4.5.1 *AutoGAD*.

One such component analysis based anomaly detector, AutoGAD is rooted in the identification of independent components that separate the anomalies [49]. The algorithm has four phases as shown in Figure 2.4.

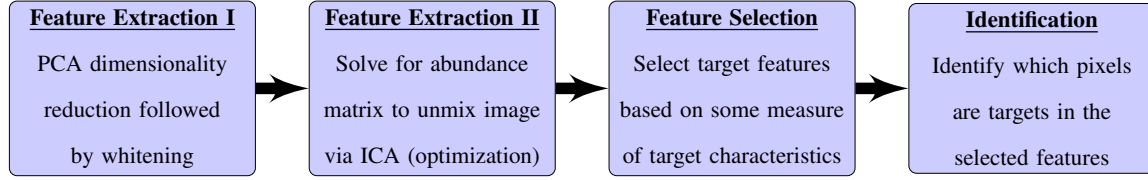


Figure 2.4: AutoGAD four-phased approach [49].

Once the algorithm has computed the independent components, feature selection is performed based on both an assessment of the maximum absolute pixel score in the component and the potential anomaly signal to noise ratio (PA-SNR) [49]. This PA-SNR is dependent upon the thresholding of the components to identify potential anomalies from the background and the variance of the scores in the two sets using

$$\text{PA-SNR} = 10 \cdot \log_{10} \left(\frac{\text{var}(\text{potential anomaly scores})}{\text{var}(\text{background})} \right). \quad (2.8)$$

Anomaly thresholding in phase 3 and the final identification in phase 4 both rely on the zero-bin histogram heuristic of Chiang, Chang, and Ginsberg [24]. This heuristic generates a histogram of the anomaly scores based on a predetermined bin width and sets the anomaly threshold at the level of the first empty histogram bin with a positive value. It relies on a reasonably sized gap between the extreme background scores and the lower

scores of the anomalous pixels. When concerned with performance stability across many images, setting the bin width parameter can be problematic [74].

2.4.5.2 Multiple PCA.

Developed sequentially from AutoGAD, the MultiPCA algorithm of Jablonski *et al.* [47] removes the ICA portions of AutoGAD and replaces them with an ensemble of measures calculated from the reconstruction error and principal component scores using a subset of principal components to describe the data. Along with decreasing the average processing time, this removes most of the stochastic elements of the AutoGAD algorithm. The ensemble measures of the MultiPCA algorithm take advantage of the characteristics of the principal components. Since the principal components define the optimal subspace in the sense of reconstructive error, the ensemble measures should be minimal for pixels that are truly from the background class, whereas anomalies should display higher reconstructive error.

2.4.5.3 GFAAD.

Subsequent to the development of AutoGAD and MultiPCA, the GFAAD family of algorithms exploits the characteristics of a VARIMAX factor rotation-based model of the background [75]. This detector builds on previous work replacing the ICA components of AutoGAD with a factor analysis approach. Using the VARIMAX rotation method, the detector is attempting to find the factors in the image that commonly gives rise to the majority of the variance in the spectral signatures. Then, given a pixel that does not relate to the factors of the image, anomalies can then be identified. Additionally, to eliminate potential anomalies from impacting the statistical estimates, GFAAD operates in an iterative manner similar to IRX.

2.4.6 Anomaly Detection in HSI using Random Fields.

A very popular approach to spectral-spatial anomaly detection is random field modeling. Random fields, or a two-dimensional spatial stochastic processes, is a method

to probabilistically describe a set of data with known correlation characteristics. For example, in remote sensing applications, neighboring pixels are more likely to be built from a similar material than pixels with greater separation. This gives rise to methods that use a probabilistic model of the random field to make inferences about the status of neighboring pixels. Although most of the research using random fields in HSI focuses on the segmentation and classification of scenes [51, 55, 100], there have been a number of random field approaches for anomaly detection. One of the earlier anomaly detectors developed for HSI anomaly detection was the Gauss–Markov Random Field (GMRF) detector from the work of Schweizer [88]. Using a one-pixel window for each pixel, the method uses an approximate maximum likelihood approximation for the Markov parameter estimates [89]. Then, using GLRT, declares anomalies based on the Markov parameters covariance structure. However, Banerjee [5] notes “the model assumes that the background is locally homogeneous, which can lead to a poor performance when the targets are located along the clutter boundaries.”

2.4.7 Spatially Dependent Methods.

Many of the anomaly detection methods for HSI focus heavily on the spectral portion of the image. In applications where the spatial resolution of the sensor is relatively high compared to the size of the anomalies present, this may be disregarding useful information for the detector. However, inclusion of spatial information in an anomaly detector is complicated by the decision on the size of the neighborhood and the accounting method for the information [90]. A number of approaches in HSI have been proposed to account for the spatial information of images, with the majority relating to the segmentation and classification of scenes [35, 107]. Two notable approaches to scene classification use a composite spectral-spatial kernel method [34] and the classification of spatial clusters with the majority vote of a spectral support vector machine (SVM) [97]. One common step in which to incorporate spatial information is the denoising of the image or components of

the image. For example, this is the approach of AutoGAD [49]. One of the most common denoising algorithms is the 2-dimensional Wiener filter [102]. It is popular both because of its computational simplicity and its adaptive nature.

A Wiener filter, also known as the Wiener-Kolomogorov in honor of similar work by Kolomogorov [58, p.342], is adaptive in the sense that, in regions of large variance, it performs less signal smoothing than in regions of small variance [20]. A two-dimensional application of a Wiener filter uses an annulus around each pixel to determine local behavior. For example, suppose apply filtering to an image with values $x(t)$, where t denotes the individual pixels in the image. The size of the window with pixels in the set η determines the number of pixels, N , used in calculations. Local behavior is characterized by the mean, μ , and the variance, σ^2 , of the pixels in the PUT window as defined as

$$\mu = \frac{1}{N} \sum_{t \in \eta} x(t) \quad (2.9)$$

and

$$\sigma^2 = \frac{1}{N-1} \sum_{t \in \eta} x(t)^2 - \mu^2 \quad (2.10)$$

respectively.

Filtering is applied to the signal $x(t)$ using

$$y(t) = \mu + [x(t) - \mu] \cdot \max \left\{ 0, \frac{\sigma^2 - \nu^2}{\sigma^2} \right\} \quad (2.11)$$

with a new resulting value of $y(t)$. The noise variance, ν^2 , if unknown, is usually estimated as the average of all values of σ^2 for the image. The ratio of the difference from the noise variance, $\sigma^2 - \nu^2$, and the variance observed at a pixel, σ^2 , provides the filtering method with an adaptive aspect. In regions of high noise, particularly those above the mean of the observed σ^2 values, this method performs little to no smoothing of the signals. Conversely, in regions displaying little noise the method filters more heavily [20].

Another approach that uses spatial information is the incremental exclusion of potential neighboring anomalies from the statistical calculations. This is the approach

seen in IRX [96]. During each iteration, calculation of the background covariance matrix for the PUT is performed only with pixels not previously marked as potential anomalies. This effectively marginalizes anomalous contributions to background statistical calculations by using the spatial information of the PUT and masking potential anomalies from being considered background.

There are also methods developed on the principal of zero crossing count analysis similar to those of time-series data. The method presented by Yuan, Wang, and Zhu is based on determining the difference of crossing counts of high-order differences of a PUT from its neighboring pixels [106]. The high-order differences are calculated from the spatially-weighted (w_{ij}) first-order difference of a pixel i , from the pixels j in the neighborhood around i , $j \in \{N_i\}$, using

$$L_i^{(1)} = \sum_{j \in \{N_i\}} \|h_i - h_j\|^2 w_{ij}. \quad (2.12)$$

High-order differences are then recursively determined using

$$L_i^{(k)} = \sum_{j \in \{N_i\}} \|L_i^{(k-1)} - L_j^{(k-1)}\|^2 w_{ij}. \quad (2.13)$$

These differences are passed through a change detector and a count of changes for a pixel compared to its neighborhood is determined. An anomaly should display a higher change count than that of a background pixel.

Another common placement of a spatially aware component to an HSI algorithm is as a post-processor. This stems from treating the anomaly detector as a coherent block distinct from pre-processing steps, such as absorption band elimination and noise reduction, and post-processing steps, such as score normalization. A very common technique used in segmentation and classification algorithms is done by using mathematical morphology through watershed segmentation study using dilation and erosion [16]. By carrying out the dilation and erosion of segments, the image is coalesced into a stable number of segments which are then treated as homogeneous clusters. Other

methods use the classification of surrounding pixels to possibly correct classifications pixel-by-pixel [41] or to exclude regions of anomalies based on their spatial attributes [67].

2.5 Quantifying Performance

The performance of an anomaly detector is based on numerous factors, so to compare the performance of anomaly detectors, one first needs to define some ways to quantify performance. The terminology commonly used labels anomalies as the positive class and the background as the negative class. As anomaly detectors are unsupervised algorithms, the condition of a data point being anomalous or not is not provided to the detectors and, when known, is used only to determine their performance.

Now to define a few terms: a True Positive (TP) result is when an anomaly is correctly declared by the algorithm, whereas a True Negative (TN) result is when the algorithm correctly declares a background point as a non-anomalous point. Border pixels—those pixels that are of uncertain condition—are excluded from the measure calculations. Similarly, a False Negative (FN) result is an anomaly that the algorithm failed to detect, and a False Positive (FP) result is an incorrect declaration of a point as anomalous.

Table 2.1: Confusion Matrix Layout.

		Detector Output	
		P	N
Truth	P	#TP	#FN
	N	#FP	#TN

The most basic measures for the performance of an anomaly detector compare the number of anomalies and background declared in an image. The True Positive

Fraction (TPF), also known as sensitivity, recall, or hit rate, is defined as

$$\text{TPF} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (2.14)$$

while False Positive Fraction (FPF), also known as specificity, is

$$\text{FPF} = \frac{\text{FP}}{\text{FP} + \text{TN}} . \quad (2.15)$$

Two other rates derived directly from the confusion matrix is Label Accuracy (LA), also known as precision and Positive Predictive Value (PPV),

$$\text{LA} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2.16)$$

and Negative Predictive Value (NPV)

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}} . \quad (2.17)$$

While it may obfuscate some performance issues, it is often advantageous to state performance into a single measure that balances both positive and negative classes. For this approach there are a number of approaches such as Matthew's Correlation Coefficient (MCC) [65], Youden's J-index [105] and the F1 Score. The F1 Score is the harmonic mean of TPF and LA:

$$\text{F1 Score} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} . \quad (2.18)$$

Youden's J index, also known as Informedness, is a common measure for the performance of diagnostic tests defined by:

$$J = \text{TPF} + (1 - \text{FPF}) - 1. \quad (2.19)$$

MCC, also commonly referred to as the correlation coefficient, includes all values from the confusion matrix:

$$\text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} . \quad (2.20)$$

In particular, Youden's J index and MCC are structured for anomaly detection applications since they include information from both classes in the ratios which guards against the impact of class imbalance [12].

Another common measure of performance is known as Area Under the Curve (AUC) [36]. Unlike the other measures mentioned above, this measure does not require thresholding of the output to declare positive and negatives. Instead, a Response Operating Characteristic (ROC) curve is formed using all possible thresholds. An example of a ROC curve can be seen in Figure 2.5. On the ROC curve, moving from left to right along the FPF axis as the threshold increases more anomalies are identified than background, increasing TPF. However, this transitions when the curve flattens out and begins to identify mostly background data. The equal chance curve is representative of an uninformed detector. If TPF and FPF are defined as functions of a single threshold value t , $F_{TPF}(t)$ and $F_{FPF}(t)$, respectively, then AUC is defined as [36]

$$AUC = \int_0^1 F_{TPF}(t) dF_{FPF}(t) = \int_{-\infty}^{\infty} F_{TPF}(t) f_{FPF}(t) dt. \quad (2.21)$$

Performance evaluation between algorithms based on AUC assumes that the algorithm has the ability to determine an anomaly threshold according to the misclassification cost of the application.

2.5.1 Thresholding Methods.

A critical step in autonomous anomaly detection is the thresholding of the anomaly scores for the declaration of anomalies. There are four approaches to selecting an anomalous threshold [6]:

1. A fixed threshold;
2. Thresholds based on the mean and variance of the scores;
3. Threshold based on parametric statistical assumptions on the background;

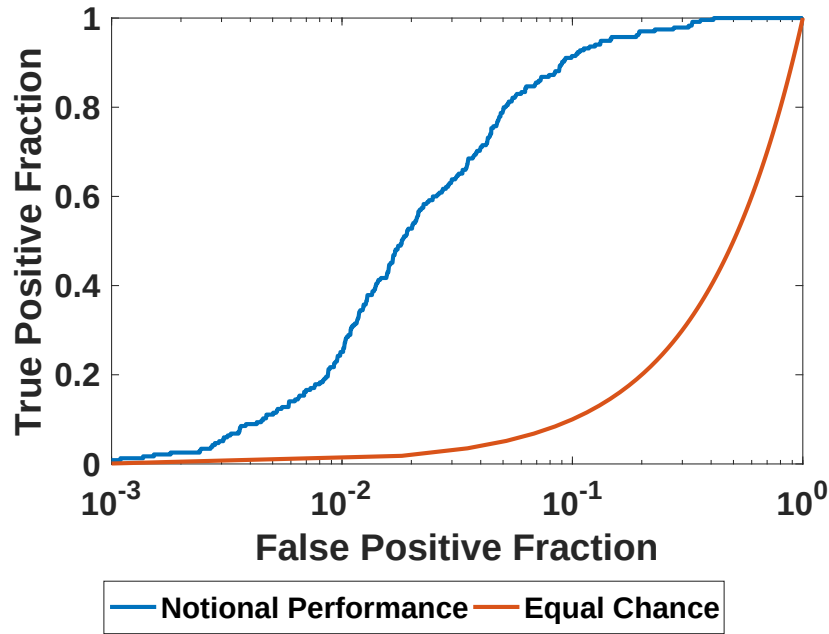


Figure 2.5: Example of a ROC curve with a notional performance curve and an equal chance curve.

4. Thresholds with non-parametric statistical assumptions of the background.

Ideally, anomaly thresholding is performed in a manner that identifies the maximum number of true positive (TP) declarations with a minimum number of false positive (FP) declarations. In different applications, increasing the number of TP declarations may warrant an imbalanced optimality causing larger rates of FP declarations. Thus, it is desirable to have some known capability of an anomaly detector to control the false positive rate. Different thresholding methods are normally used with detectors with varying impacts on performance. The fixed threshold and mean/variance approaches are uncommon in anomaly detectors due to the complexity of the processes. An example of parametric-based thresholding is in RX. It assumes the background follows a multivariate

Gaussian distribution around the local mean and is traditionally thresholded with a critical value from a Chi-squared distribution.

An example of a non-parametric approach to thresholding is using the empirical output from the detector. Assuming the anomalies are more likely to produce high anomaly scores, selecting a high empirical quantile (e.g. top 10%) provides a computationally simple method of thresholding, however, it provides no adaptability based on the inferred presence of anomalies. Under this method a detector will always nominate a constant amount of pixels to be reviewed by a human analyst. For real-time HSI applications, this may overwhelm the capacity of analysts to review and result in missed identification of anomalies. This is the approach that is implemented by the SVDD detector. As noted in 2.4.5, AutoGAD relies on an empirical heuristic based on finding the first zero-count bin in the histogram of the anomaly scores.

An often desirable thresholding method characteristic is known as CFAR [70]. This presents a class of thresholding methods that are adaptive to the background data such that the False Alarm Rate (FAR) is controllable to a desired level. Controlling the FAR in an anomaly detector is often critical in the acquisition and deployment of detectors.

2.6 Collected Images

The HSI images for this document were captured using the Hyperspectral Digital Imagery Collection Equipment (HYDICE) image sensor [82]. The sensor is a push-broom type sensor operated from an aircraft returning 210 spectral bands between 397 nm and 2,500 nm. Spatial resolution varies due to altitude variation of the sensor at time of collection.

Table 2.2: HYDICE Image Properties.

	Height $\times$ Width	Number of Anomaly Clusters	Proportion of Anomalies	Mean Mahalanobis Distance of Anomalies	Setting
1D	291x199	6	0.004	15.8	Desert
1F	191x160	11	0.035	49.0	Forest
2D	215x104	46	0.026	33.5	Desert
2F	312x152	30	0.007	30.2	Forest
3D	156x156	4	0.018	51.0	Desert
3F	266x136	19	0.005	33.5	Forest
4	460x78	15	0.026	44.2	Desert
4F	205x80	29	0.007	32.7	Forest
5	355x150	15	0.011	39.0	Desert
5F	470x155	45	0.015	43.3	Forest

III. Method of Sensitivity Analysis in Anomaly Detection Algorithms for Hyperspectral Images

3.1 Introduction

Hyperspectral Imaging (HSI) has seen a rapid expansion of application areas over the last few decades. From the monitoring of agriculture crops for ripeness to remote sensing of military targets, HSI has proven itself a valuable tool [62]. HSI is employed in both supervised and unsupervised settings. In a supervised setting, the analyst knows the spectral signature of the material they are searching for in an image and uses algorithms for spectral matching. However, there are applications where the spectral signatures that are of interest are unknown, or perhaps not present at all. In this setting, algorithms for the detection of anomalies within the spectral signatures are the most useful. In this unsupervised setting, the problem set is now much larger as we no longer have a defined set of objects to identify. This arises from the fact that in anomaly detection the characteristics of the set of all anomalies is undefined [98]. Traditionally, anomaly detection algorithms in HSI are compared using a limited set of hand labeled images to provide what is arguably anecdotal evidence of performance. Such a procedure runs the risk of forming algorithms that are overly sensitive to characteristics of the limited number of observed anomalies present in the images. The chapter presents an extension to the evaluation of algorithms on limited data to provide a deeper understanding of an algorithms performance characteristics.

3.2 Background

At its core, anomaly detection is meant to detect and identify patterns, objects, or observations that do not belong within the greater majority of the data set [22]. Very similar to outlier detection and novelty detection, anomaly detection has applications in a

number of areas such as cyberspace intrusion detection, financial fraud detection, and medical monitoring [22]. There have been a number of excellent survey papers on anomaly, outlier, and/or novelty detection [22, 79]. Each application area differs in the characteristics that make up an anomaly. For example, in cyberspace intrusion detection, repeated requests for access to a network may be a sign of an intrusion or just a very active user. Whereas, in medical monitoring, the absence of a regular signal may be cause for concern. Within HSI, the anomalies are pixels, or collections of pixels, that display spectral signatures different than the other signatures in the image. Herein, anomaly detection is concerned with understanding the statistical nature of the background data in order to be able to discern the anomalies.

3.2.1 Anomaly Detection Methods for HSI.

As discussed in Section 2.4, there are many proposed algorithms for the detection of anomalies in HSI. One of the most prominent is Reed–Xiaoli (RX) [81]. Developed as one of the earliest anomaly detectors in the field for multispectral images, RX is a local detector that uses either the radiance values or the principal component scores of a subspace [96] of the image to perform statistical analysis. It is largely treated as a benchmark for anomaly detection in HSI. This chapter examines RX and a number of its derivative algorithms. Additionally, results from the Support Vector Data Description (SVDD) detector [5] and the Topological Anomaly Detector (TAD) [8], along with the component based detectors of Autonomous Global Anomaly Detector (AutoGAD) [49], Global Factor Analysis-based Anomaly Detector (GFAAD) [74], and Multiple PCA Anomaly Detector (MultiPCA) [47] to show how different anomaly detection approaches respond during sensitivity analysis. For an overview of these algorithms, see Chapter 2.4.

3.2.2 Comparing Anomaly Detectors.

A number of HSI datasets exist for benchmark analysis of algorithms. For our investigations here, the results are from five images of the Forest I and five images from the Desert II collection events of the AVIRIS platform utilizing the Hyperspectral Digital Imagery Collection Equipment (HYDICE) sensor. Furthermore, as these were the images used for the training of the initial AutoGAD algorithm's parameters [49], the added benefit of our approach to understanding an algorithms performance characteristics is highlighted.

3.2.2.1 Other Sensitivity Analysis Approaches.

As previously stated, the majority of techniques for the evaluation of anomaly detectors on hyperspectral images have focused on a handful of labeled images. This approach to evaluation operates under the assumption that the images being evaluated are an acceptable representation of the expected images to be observed during real-world application. There have been some efforts to examine the sensitivity of algorithms to changes in the characteristics of the anomalies. One such method for use in the presence of subpixel anomalies (i.e. spatial size of the anomalies is less than that of the spatial resolution of the sensor) is known as the target implant method [10]. The method varies a convex combination of a background and anomaly signature replacing the observed anomaly signature. This is meant to approximate a pixel that is proportionally filled with an anomalous signature (i.e. 30% of the pixel was anomalous material). Decreasing the proportion of anomalous signature increases the difficulty in finding the anomalous pixel.

The sensitivity analysis on an algorithm is also closely related to more complex methods aimed at determining the performance of an algorithm. One such method uses Monte Carlo methods to evaluate the volume of the background distribution that the detector uses for declaring anomalies [99]. In doing so, it seeks to gain a proxy measurement of a missed detection rate for a given level of performance. The ideal detector would return the smallest volume while a poor detector, or one highly sensitive to

the anomalies present, would return large volumes at the same level of false positive rate. This assumes that anomalies are uniformly distributed across the entire space and a detector with a smaller volume for a given false positive rate would return a smaller number of false negative declarations.

3.2.3 Mahalanobis Distance.

Developed on the assumption of multivariate normal data, the Mahalanobis Distance (MD) is a generalization of the univariate measure of standard deviations [60]. The distance is a function of a column vector x as well as the spectral mean vector, μ_B , and covariance matrix, Σ_B , of the background pixels. The MD is determined by

$$MD(\mathbf{x}) = \sqrt{(\mathbf{x} - \mu_B)' \Sigma_B^{-1} (\mathbf{x} - \mu_B)}. \quad (3.1)$$

Our method is based on the Mahalanobis Distance of the anomalies to the background. This measure is chosen over other distance measures due to the nature of the collection techniques for HSI. Since the sensor is recording the reflectance of the materials in the scene, the value are dependent both on the materials themselves and the illumination provided by the light source (typically the sun). It may be the case that pixels under different illumination conditions produce similar reflectance values resulting in a low Euclidean Distance (ED), whereas with MD they are well separated regardless of illumination. The comparison of an anomalous signature to its nearest neighbor (NN) using Euclidean and Mahalanobis distances can be seen in Figure 3.1 using the the anomalous signature in pixel number 11111 from ARES 3F. Examining the NN of the blue signature with regards to the background of the image and using ED identifies the black signature of pixel 14648, which belongs to the background class. A more relevant relationship for anomaly detection is shown with the red signature of pixel 8369 which is the MD NN and belongs to the anomalous class. Even though the absolute reflectance of the two signatures is very different, the actual variation of the signatures with regards to

the background variance is less. It is this difference with regards to the background of the image that is the most critical information.

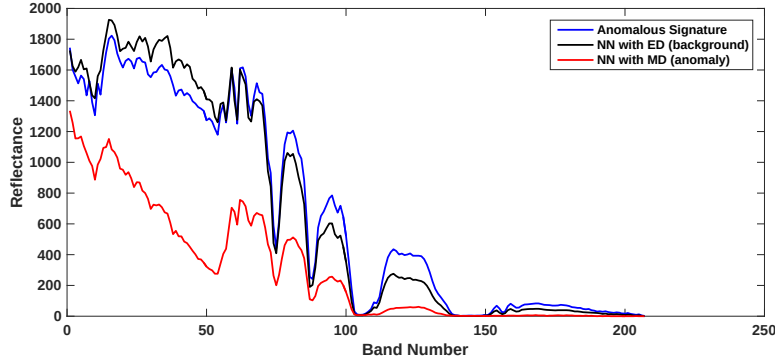


Figure 3.1: Nearest neighbors to an anomalous signature (pixel 11111 in black) in terms of ED (pixel 14648 in blue) and MD (pixel 8369 in red). Interestingly, the ED nearest neighbor is a background signature while the MD returns an anomalous signature. Example is from ARES3F image.

3.3 Methodology

This section proposes an iterative method to investigate the sensitivity of an algorithm to slight changes in the characteristics of the anomalies. During this process the anomaly detector being evaluated is repeatedly tasked to find the anomalies in progressively more difficult images. These progressively more difficult images are created by iteratively “masking” the characteristics of the anomalies. This masking is done with the respect to the background accomplished by slightly decreasing the MD of each anomaly with respect to the background. Using the labeled background mean, μ_B , and covariance, Σ_B , each anomaly’s observed spectral radiance vector x_0 changes from iteration i , x_i , to the next, x_{i+1} , using

$$\mathbf{x}_{i+1} = \mathbf{x}_i - \alpha \Sigma_B^{-1} (\mathbf{x}_i - \mu_B) \quad (3.2)$$

with a variable step size α . This is similar to methods for gradient-based learning within learning vector quantization [52] with our anomalies treated as the vectors and the gradient provided by the covariance of the background. The value of α should be small to ensure that highly anomalous behavior in a spectral signature does not dominate the movement. The overall effect creates anomalies more difficult to detect with each successive iteration since they move closer and closer to the mean of the background. This method of sensitivity analysis is called Mahalanobis Masking.

As can be seen in Figure 3.2, Mahalanobis Masking causes the anomalous signals to not only move towards the mean vector of the background, but also change their shape to more resemble the background with respect to the spectral covariance. In contrast, when the Euclidean distance is used for masking the of the anomalies such as with the target implant method, the general shape path of the signature during smoothing is restricted to moving towards the mean without changing the overall shape characteristics of the signature.

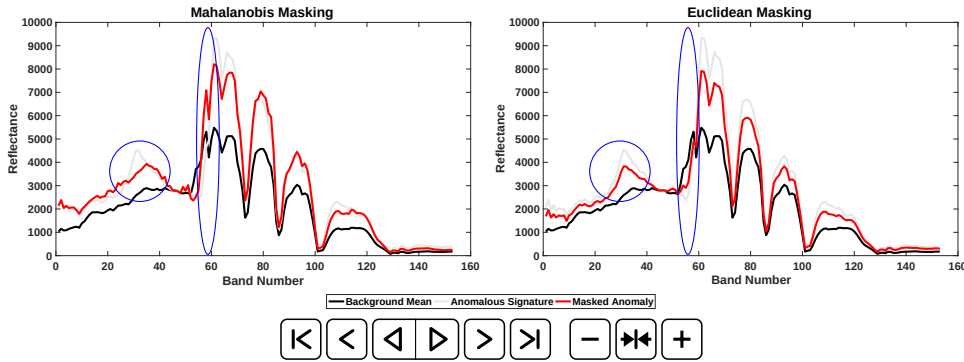


Figure 3.2: Impact of Mahalanobis Masking on an anomalous pixel's spectral signature.

This approach to evaluating the performance of algorithms is valuable in that it ensures the algorithms are not overly adapted to the characteristics of the handful of anomalous signatures present in a few images. Ideally, an algorithm would present a

stable, gently decreasing performance during iterations of Mahalanobis Masking. Poor results would display inconsistent performance during Mahalanobis Masking with large jumps in performance between iterations or a quick degradation in performance compared to other detectors.

3.3.1 *Evaluation of Sensitivity.*

The method for sensitivity analysis presented above is focused on presenting increasingly difficult to detect anomalies in order to test the algorithms. Qualitatively, evaluation of competing algorithms could be performed through the observation of the performance metrics of choice over a fixed number of masking iterations. This is done through graphical inspection of the results over many iterations of masking. Samples of these graphs are presented below in Figure 3.3.

While the qualitative approach to sensitivity analysis is useful as a diagnostic approach, the overall evaluation of an algorithm is difficult over many images and comparing numerous of competing detectors. For a deeper evaluation of algorithm performance we need more quantitative method of evaluation. Our recommended approach assumes that the detectors are functioning in a manner that is controlling the False Alarm Rate (FAR). Since many detectors are built in a manner that prevents high FAR through thresholding methods, this assumption is often met. Assuming the FAR is controlled, we can focus on the true positive detection capabilities of the detectors. There are two principal performance characteristics that we want to understand from the Mahalanobis Masking process: detector ability and detector stability. Detector ability is focused on the overall performance of the algorithm's True Positive Fraction (TPF) across the iterations of masking. Let $D_{TPF}(I_i)$ be the TPF of detector D on image I with i iterations of masking. We define the mean measure for the ability of a detector across N iterations of masking as

$$\mu_{D,N} = \sum_{i=0}^N \frac{D_{TPF}(I_i)}{N}. \quad (3.3)$$

A high mean TPF across iterations indicates that the algorithm is able to detect anomalies across the range of difficulty tested by masking. However, it does not directly indicate the stability of the detector through the iterations of masking. The stability measure of the algorithm is concerned with the consistency of the detector's ability to detect anomalies. We define the stability measure of a detector as

$$\sigma_{D,N}^2 = \sum_{i=0}^N \frac{(D_{TPF}(I_i) - \mu_{D,N})^2}{N}. \quad (3.4)$$

These measures need not be in competition with each other nor are they directly related. A detector can have relatively high overall performance with high variability or any combination of the two. While the performance of an algorithm during masking is not stationary, the two measures provide a gross measure of the performance of the algorithm on all variations of the anomalies. In viewing these measures as a mean-like measure in $\mu_{D,N}$ and a variance-like measure in $\sigma_{D,N}^2$, they can further be combined into a single measure using an approach native to the Robust Parameter Design (RPD) field [56] for combining the mean and variance performance into a single Mean-Squared Error (MSE)-like measure defined by

$$\hat{\text{MSE}}_{D,N} = (\mu_{D,N} - 1)^2 + \sigma_{D,N}^2. \quad (3.5)$$

This measures the deviation from the optimal of the mean performance and the variance of the performance over the masking iterations.

3.4 Results

In order to highlight the usefulness of the sensitivity analysis method using the qualitative approach, we examine the performance of the analysis on three anomaly detectors presented above: RX, SVDD, and AutoGAD. For the quantitative approach we report the results from all nine detectors tested. The overall process begins by setting the initial starting performance of each detector. The settings of the parameters at this initial operating condition will be the same settings of the algorithm's parameters that it uses

throughout all iterations of Mahalanobis Masking and across all images to simulate a detector used with selected settings when deployed in the field. In order to have a fair comparison, we used the same method to select parameters for each algorithm. This is done using the initial performance on the unmasked images and then minimizing the 1-norm (Manhattan distance) from the ideal performance point across the images. This ideal performance point is set at 1 for the TPF and 0.05 for the FPF, to replicate perfect detection and at a CFAR of 5%. This resulted in the settings for the RX algorithm with a window size of 25×25 pixels, the first ten PCA scores, and a $\alpha = 0.07$ for the threshold critical value. All versions using the Linear RX (LRX) approach used a line length of 625 pixels. The α value for the threshold was 0.007 for LRX, 0.0007 for High-Low FAR LRX (HLRX), and 0.00007 for Iterative Linear RX (ILRX). TAD was thresholded using the 95% percentile. For AutoGAD, we used the parameters recommended by Williams *et al.* [103] except for the bin width for identification of anomalous components as (2100/Number of Pixels) and a constant bin width for the potential anomaly signal to noise ratio (PA-SNR) determination as 0.04. The parameters for GFAAD are provided by Paciencia in [74] and similarly the recommended settings of Jablonski et al. are used for MultiPCA [47]. For SVDD, we used BACON [17] with a c value of 1000 and a α of 0.1 to initially screen for anomalies and for SVDD we used a Gaussian kernel with a σ value of 0.05. The performance at the initial settings is summarized in Tables 3.1 and 3.2.

Table 3.1: Initial TPF Performance of Algorithms.

	ARES1D	ARES1F	ARES2D	ARES2F	ARES3D	ARES3F	ARES4	ARES4F	ARES5	ARES5F	Mean
RX	0.860	0.208	0.589	0.840	0.226	0.903	0.316	0.624	0.581	0.323	0.547
HLRX	0.991	0.680	0.994	0.863	0.297	0.952	0.760	0.927	0.957	0.900	0.832
LRX	0.817	0.547	0.901	0.655	0.276	0.848	0.577	0.798	0.836	0.847	0.710
ILRX	1.000	0.656	0.998	0.951	0.263	0.952	0.966	0.936	0.968	0.918	0.861
TAD	0.753	0.327	0.989	0.951	0.221	0.883	0.476	0.679	0.651	0.399	0.633
GFAAD	0.898	0.963	0.945	0.964	0.874	0.855	0.797	0.835	0.885	0.728	0.874
AutoGAD	0.847	0.980	0.996	0.971	0.973	0.841	0.918	0.927	0.962	0.811	0.923
MultiPCA	0.791	1.000	0.927	0.984	0.993	0.841	0.797	0.569	0.882	0.721	0.851
SVDD	0.694	0.697	0.967	0.772	0.575	0.862	0.617	0.661	0.843	0.599	0.729

Table 3.2: Initial FPF Performance of Algorithms.

	ARES1D	ARES1F	ARES2D	ARES2F	ARES3D	ARES3F	ARES4	ARES4F	ARES5	ARES5F	Mean
RX	0.061	0.043	0.016	0.046	0.060	0.049	0.035	0.027	0.046	0.049	0.043
HLRX	0.074	0.017	0.023	0.076	0.079	0.053	0.074	0.068	0.036	0.068	0.057
LRX	0.054	0.024	0.022	0.057	0.061	0.051	0.050	0.054	0.039	0.050	0.046
ILRX	0.091	0.008	0.014	0.087	0.074	0.029	0.110	0.060	0.020	0.070	0.056
TAD	0.047	0.038	0.024	0.043	0.046	0.045	0.036	0.044	0.042	0.042	0.041
GFAAD	0.023	0.012	0.004	0.016	0.008	0.069	0.021	0.028	0.033	0.019	0.023
AutoGAD	0.047	0.019	0.036	0.062	0.074	0.050	0.093	0.077	0.065	0.090	0.061
MultiPCA	0.023	0.018	0.004	0.056	0.026	0.021	0.026	0.026	0.014	0.017	0.023
SVDD	0.053	0.031	0.031	0.051	0.045	0.051	0.041	0.050	0.046	0.044	0.044

3.4.1 Qualitative Results.

The results of the sensitivity analysis show a number of intriguing results depicting the usefulness of the method. Two forms of the results, one qualitative and the other quantitative, provide insight. The first form is a graphical depiction of the performance characteristics of the algorithms at each step of the analysis. This can be seen in Figure 3.3 for a small subset of the algorithms tested. The entirety of the charts are located in Appendix B.

The results of the RX detector in Figure 3.3 are the most stable stemming from the algorithm's reliance on MD, albeit a local MD, as the anomaly measure. Furthermore, since the RX detector uses a probabilistic threshold method, changes in the performance metrics are stable relative to the changes to the anomalies. The RX detector displays the most ideal behavior of the algorithms tested with an overall smooth downward slope to its performance during the Mahalanobis Masking operations. However, the algorithm also returns the lowest performance averages compared to the other detectors; thus, the stability comes at a cost of overall performance.

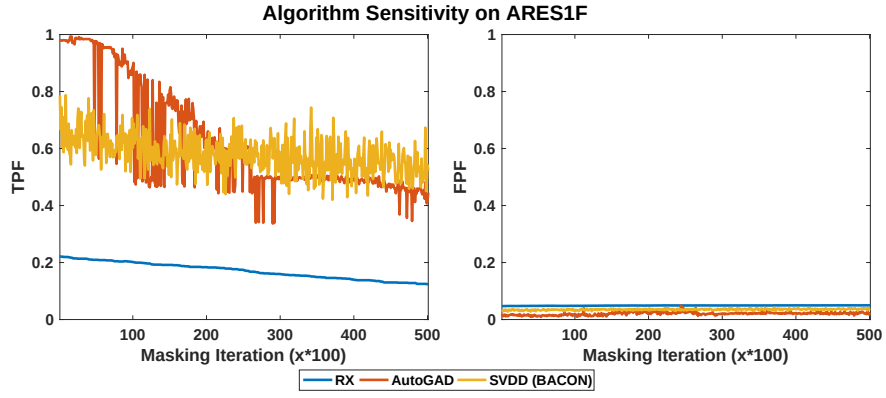
The performance of SVDD is shown to be relatively stable with moderate variability around a declining average performance similar to that of the RX detector. However, given the relative stability of the detector over a number of Mahalanobis Masking iterations, it seems that the detector is generally robust to the characteristics of anomalies present in the image. The variance appears to be related to the selection of the subset that SVDD initially selects using the BACON algorithm. Since the algorithm displays the general stability of the RX algorithm the unsupervised adaptation of SVDD is the best overall detector of the three analyzed.

The most interesting results stem from the performance of AutoGAD. While the algorithm starts the sensitivity analysis with the highest mean performance on all images, this advantage quickly degrades on most images. This indicates that the AutoGAD

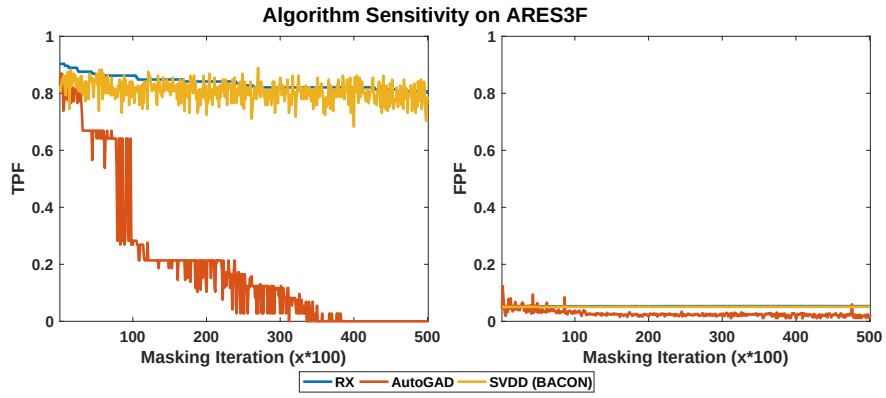
algorithm quickly loses the ability to identify anomalies as they are masked into the background. It seems that this is a result of the process that AutoGAD uses to select components for the identification of anomalies and the component score thresholding method that AutoGAD implements. Recall that AutoGAD uses the zero-bin histogram algorithm for setting a declaration threshold for both the identification of anomalies and as a precursor to the calculation of the PA-SNR. This thresholding technique seems the most sensitive to small changes in the characteristics of the image. As the scores of the anomalies gradually decrease during masking, they can pass through the histogram bin that was previously empty potentially causing a large change to the threshold value. This appears to be affecting the stability of the algorithm itself. Additionally, the algorithm displays large drops in performance when the algorithm fails to identify useful additional dimensions in the subspace. The combination of these factors results in the algorithm being the least stable to changes in the anomalies of the algorithms tested. Thus, even though the algorithm is capable of being tuned using its parameters to high performance characteristics, the stability of the algorithm calls into question the overall performance capabilities of the algorithm.

3.4.2 Quantitative Results.

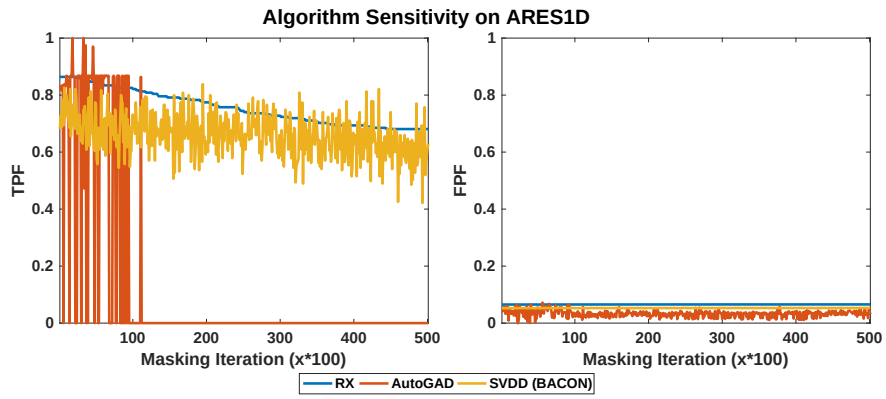
As the qualitative analysis of the algorithms is difficult to compare in absolute terms across many algorithms, this section extends the sensitivity analysis to all nine algorithms tested. First, to verify that the assumption that each algorithm has False Positive Fraction (FPF) in an acceptable range, the mean FPF is examined across the images. The only algorithm that seems to have difficulty constraining the FPF is AutoGAD. The charts from the qualitative evaluation also depict evidence of this. However, even though the algorithm returns FPF far above some of the other algorithms, it is retained as it is the top performing algorithm in the initial baseline.



(a)



(b)



(c)

Figure 3.3: Notional Performance of Algorithms after Iterative Mahalanobis Masking. Each subfigure is the performance of the algorithm's on a different image.

Table 3.3: Mean TPF of Algorithms across 500 iterations of Masking.

	ARES1D	ARES1F	ARES2D	ARES2F	ARES3D	ARES3F	ARES4	ARES4F	ARES5	ARES5F	Mean
RX	0.738	0.161	0.557	0.796	0.194	0.828	0.213	0.560	0.549	0.309	0.491
HLRX	0.965	0.572	0.959	0.813	0.258	0.869	0.579	0.820	0.942	0.884	0.766
LRX	0.748	0.429	0.830	0.612	0.225	0.742	0.430	0.680	0.803	0.822	0.632
ILRX	0.998	0.368	0.962	0.902	0.242	0.870	0.704	0.818	0.956	0.902	0.772
TAD	0.709	0.266	0.963	0.921	0.193	0.850	0.467	0.612	0.622	0.388	0.599
GFAAD	0.325	0.446	0.776	0.771	0.891	0.437	0.478	0.545	0.482	0.597	0.575
AutoGAD	0.095	0.621	0.946	0.837	0.822	0.203	0.621	0.447	0.524	0.731	0.585
MultiPCA	0.002	0.022	0.073	0.242	0.188	0.013	0.212	0.028	0.046	0.145	0.097
SVDD	0.666	0.583	0.885	0.675	0.492	0.808	0.562	0.539	0.731	0.186	0.613

Table 3.4: Mean FPF of Algorithms across 500 iterations of Masking.

	ARES1D	ARES1F	ARES2D	ARES2F	ARES3D	ARES3F	ARES4	ARES4F	ARES5	ARES5F	Mean
RX	0.061	0.044	0.021	0.047	0.059	0.049	0.038	0.028	0.047	0.049	0.045
HLRX	0.073	0.016	0.022	0.077	0.077	0.053	0.066	0.067	0.036	0.067	0.055
LRX	0.054	0.026	0.022	0.058	0.060	0.052	0.048	0.055	0.039	0.049	0.046
ILRX	0.089	0.007	0.012	0.083	0.074	0.030	0.091	0.060	0.020	0.068	0.053
TAD	0.047	0.040	0.027	0.044	0.047	0.045	0.039	0.045	0.042	0.042	0.042
GFAAD	0.022	0.012	0.004	0.024	0.010	0.027	0.022	0.037	0.034	0.032	0.022
AutoGAD	0.032	0.020	0.023	0.099	0.059	0.026	0.072	0.055	0.053	0.105	0.055
MultiPCA	0.015	0.001	0.001	0.048	0.035	0.018	0.064	0.048	0.001	0.013	0.024
SVDD	0.053	0.035	0.035	0.052	0.046	0.051	0.043	0.051	0.047	0.116	0.053

Table 3.5: Variance TPF of Algorithm across 500 iterations of Masking.

	ARES1D	ARES1F	ARES2D	ARES2F	ARES3D	ARES3F	ARES4	ARES4F	ARES5	ARES5F	Mean
RX	3.5E-03	7.9E-04	1.3E-04	1.7E-04	1.7E-04	6.3E-04	1.5E-03	2.3E-03	2.6E-04	2.1E-05	0.001
HLRX	2.9E-04	3.1E-03	1.4E-04	2.9E-04	2.2E-04	7.5E-04	5.6E-03	3.9E-03	2.2E-05	1.3E-05	0.001
LRX	1.7E-03	2.7E-03	1.1E-03	1.5E-04	4.6E-04	1.2E-03	3.3E-03	2.9E-03	1.5E-04	1.7E-05	0.001
ILRX	4.5E-06	1.2E-02	1.3E-04	1.3E-04	1.2E-04	7.1E-04	1.5E-02	3.6E-03	5.3E-06	2.2E-05	0.003
TAD	3.2E-04	3.5E-04	6.5E-05	5.8E-05	6.4E-05	2.4E-04	3.6E-06	6.0E-04	1.0E-04	1.3E-05	0.000
GFAAD	7.3E-02	6.7E-02	1.2E-03	1.1E-02	8.0E-05	1.4E-02	6.1E-03	3.2E-02	3.7E-02	1.3E-03	0.024
AutoGAD	7.2E-02	3.6E-02	8.2E-04	1.0E-02	8.9E-02	6.0E-02	2.4E-02	3.5E-02	6.8E-02	3.3E-03	0.040
MultiPCA	1.3E-03	1.2E-02	2.8E-02	9.9E-02	1.2E-01	6.0E-03	1.8E-02	8.0E-03	2.5E-02	4.5E-02	0.036
SVDD	5.4E-03	4.0E-03	8.0E-04	2.8E-03	1.2E-02	1.3E-03	2.2E-03	1.3E-03	4.2E-03	3.4E-02	0.007

Table 3.6: Mean Squared Error of TPF of Algorithm across 500 iterations of Masking.

	ARES1D	ARES1F	ARES2D	ARES2F	ARES3D	ARES3F	ARES4	ARES4F	ARES5	ARES5F	Mean
RX	0.072	0.704	0.196	0.042	0.650	0.030	0.620	0.196	0.203	0.477	0.319
HLRX	0.002	0.186	0.002	0.035	0.551	0.018	0.183	0.036	0.003	0.013	0.103
LRX	0.065	0.329	0.030	0.151	0.602	0.068	0.328	0.105	0.039	0.032	0.175
ILRX	0.000	0.412	0.002	0.010	0.574	0.017	0.103	0.037	0.002	0.010	0.117
TAD	0.085	0.539	0.001	0.006	0.652	0.023	0.284	0.151	0.143	0.375	0.226
GFAAD	0.529	0.374	0.051	0.063	0.012	0.331	0.279	0.239	0.305	0.164	0.235
AutoGAD	0.890	0.180	0.004	0.037	0.121	0.695	0.168	0.341	0.295	0.075	0.280
MultiPCA	0.997	0.968	0.888	0.673	0.782	0.981	0.639	0.953	0.934	0.777	0.859
SVDD	0.117	0.178	0.014	0.109	0.270	0.038	0.194	0.214	0.077	0.696	0.191

Table 3.3 shows that there is some major changes in the performance that is seen in the initial images (Table 3.1). Whereas the component-based AutoGAD and GFAAD are amongst the top performing algorithms, once masking is applied to the images this performance quickly degrades. Evaluating the values of $\mu_{D,500}$ for AutoGAD, GFAAD, and MultiPCA indicate a larger drop from the initial TPF than with the other algorithms. The ILRX returns the highest ability score of the algorithms tested.

Continuing the quantitative analysis of the algorithms by examining the $\sigma_{D,500}^2$ values sets a common theme to that of the $\mu_{D,500}$ values. The component based methods return mean $\sigma_{D,N}^2$ values an order of magnitude larger than the other algorithms. The most dominant algorithm in terms of stability is the TAD algorithm.

The final measure of performance in our analysis is the $M\hat{S}E_{D,500}$ of the ideal performance point for the algorithms. This measure aims to show which algorithm performs the best in terms of both ability and stability. The HLRX algorithm shows the best performance in terms of MSE followed closely by the ILRX algorithm. This is expected as the algorithms are very closely related. However, as the performance of the algorithms are close, other performance characteristics such as computationally complexity can be weighed to determine the dominant algorithm. In this case, as the maximum number of iterations for ILRX can often be ten times the number required for HLRX, the HLRX is determined to be the best performing algorithm.

The overall value of the Mahalanobis Masking approach is the highlighted most clearly with the performance of the MultiPCA algorithm. During the initial performance the algorithm performed as one of the best algorithms on the images with a mean TPF of 0.851. However, during masking the algorithm is heavily degraded resulting in a mean performance value of 0.097. This indicates that the algorithm is overly sensitive to the characteristics of the anomalies present. Similar performance degradations are evident in the other component-based algorithms of AutoGAD and GFAAD. This commonality in

instability under masking to the component-based algorithms may indicate a difficulty in isolating the anomalies from the estimation of the global covariance matrix, a common process across all three of the algorithms.

3.5 Conclusion

This chapter presents a method to perform sensitivity analysis on the performance of anomaly detection algorithms for hyperspectral images. By incrementally decreasing the Mahalanobis distance of the anomalies present in the image with respect to the background, the detection of the anomalies is progressively increased in difficulty. This procedure can highlight issues with algorithms that were previously unknown as is the case with the component based algorithms of AutoGAD and GFAAD. Further research into this algorithms is required to determine the source of the performance degradation due to Mahalanobis Masking. The sensitivity of these algorithms to Mahalanobis Masking may indicate that the algorithms are not robust in their parameter settings. The following chapter aims to show that this knowledge can be injected into a RPD framework to yield performance increases.

IV. Identification of Robust Parameters using Mahalanobis Masking

4.1 Introduction

The field of anomaly detection is principally focused on the identification of unexpected subset present within a larger dataset. As with other application areas, anomaly detection in Hyperspectral Imaging (HSI) faces the challenge of using large amounts of data to make decisions regarding whether a data point is an anomaly or simply part of the background. The sheer amount of data that must be analyzed has spurred the development of autonomous algorithms, with a vast collection them populating the literature for the detection of anomalies in HSI [63]. The autonomous nature of these algorithms necessitates the selection of parameter values that control performance characteristics of the algorithm.

One difficulty that arises in the development of many algorithms is the identification of parameters which provide superior performance characteristics during operation. These parameters control how the algorithms operate but require *a priori* knowledge of the operational environment to be properly set. The determination of parameter values is often done using only a limited amount of truth-labeled data. When an algorithm has numerous parameters, it may be difficult to truly understand the performance characteristics from the amount of data available. One systematic approach for the determination of parameter values is Robust Parameter Design (RPD). Using a model of the system (i.e. the detection algorithm), RPD determines parameter settings by finding the optimal values balancing performance and variability [94]. When these models are determined from scarce data, as is often the case in anomaly detection, they can become unstable and biased from the true system response. There are a number of proposed approaches to remedy this, such as the use of synthetic imagery through the combination of two images [15] or physics-based simulations [30]. This chapter proposes a separate approach, which focuses on

incrementally modifying the anomalies in real-world collected imagery to make them either more or less difficult to detect. Termed *Mahalanobis Masking* in Chapter 3, this process changes the spectral signatures of the anomalies according to their Mahalanobis distance from the background [68]. The aim is to show that when combined with RPD, Mahalanobis Masking can improve quality of the resulting robust parameters for anomaly detection algorithms when limited truth-labeled data is available.

This chapter is organized as follows: First, the relevant background on the RPD technique along with the Multivariate Delta Method (MDM) approach for building and optimizing the model of the system is presented. Next, Mahalanobis Masking and the modifications necessary for its use within a noise framework for RPD is discussed. Finally, the chapter ends examining the new process methodology and empirical results on the parameter determination of an example anomaly detector, Autonomous Global Anomaly Detector (AutoGAD) [49].

4.2 Robust Parameter Design

The field of RPD is focused on the selection of control factors in a system resulting in the best performance of the system in terms of desired performance and variability. Genichi Taguchi is credited with the introduction of RPD to the United States in the 1980s [94, 95]. The basis for RPD is built on the assumption that most of the variation resulting from a system is due to noise factors that are uncontrollable in operation but may be controlled or measured during laboratory conditions. Through experiments wherein these noise factors are controlled and with known levels, the controllable factors are varied to return a level of system performance that balances the desired performance with the variance of the output. A standard RPD model is built using the vector of control factors, $\mathbf{x}$, and a vector of noise factors, $\mathbf{z}$ and takes the form

$$y = \beta_0 + \mathbf{x}'\boldsymbol{\beta} + \mathbf{x}'\mathbf{B}\mathbf{x} + \mathbf{z}'\boldsymbol{\gamma} + \mathbf{x}'\boldsymbol{\Delta}\mathbf{z} + \epsilon \quad (4.1)$$

where $\mathbf{x}$ is an $m \times 1$ vector, $\mathbf{z}$ is a $n \times 1$ random vector, β_0 is an intercept term, β is an $m \times 1$ vector of first-order control factor coefficients, $\mathbf{B}$ is an $m \times m$ matrix of second-order control factor coefficients, γ is an $n \times 1$ vector of first-order noise factor coefficients, Δ is a $m \times n$ matrix of second-order noise factor coefficients, and ϵ is the scalar random variable that models the error term assumed to be normally distributed with a mean of 0 and a variance of σ^2 . When uncontrolled, the noise factors are assumed to be normally distributed, such that $E[\mathbf{z}] = \mathbf{0}$ and $\text{Var}[\mathbf{z}] = \sigma_z^2 I$ where I is an $n \times n$ identity matrix. As proposed, this model assumes all terms higher than second-order control interactions or first-order control by noise interactions are insignificant.

The original approach of RPD calls for the optimization of derived Signal-to-Noise Ratio (SNR) values for the model. This approach is controversial, but Taguchi's contribution in forming the RPD framework into an optimization formalism that accounts for the mean and variance of the response is widely accepted [78]. This method derives both the Mean Model, $\hat{\mu}$, and Variance Model, $\hat{\sigma}_y^2$, from the Standard Model. Optimization is conducted by applying techniques from multi-objective optimization to these models. For example, from the standard RPD model, the Mean Model and Variance Model are, respectively,

$$\hat{\mu} = E[y] = \beta_0 + \mathbf{x}'\beta + \mathbf{x}'\mathbf{B}\mathbf{x} \quad (4.2)$$

$$\hat{\sigma}_y^2 = \text{Var}[y] = (\gamma' + \mathbf{x}'\Delta)' \Sigma_z (\gamma' + \mathbf{x}'\Delta) + \sigma_z^2. \quad (4.3)$$

One of the most common techniques for the optimization of the Mean and Variance Models is the bivariate Mean-Squared Error (MSE) approach of Lin and Tu [56]. Depending on the desired level of mean response being maximized, minimized, or having a desired value (T), this approach uses one of three objective functions, respectively,

$$f_{min} = \hat{\mu}^2 + \hat{\sigma}_y^2, \quad (4.4)$$

$$f_{max} = -\hat{\mu}^2 + \hat{\sigma}_y^2, \quad (4.5)$$

$$f_{target} = (\hat{\mu} - T)^2 + \hat{\sigma}_y^2. \quad (4.6)$$

Later this bivariate technique is applied to a bounded performance metric that utilizes the target-is-best MSE model with a target value of 1.

4.2.1 Multivariate Delta Method.

The standard approach of RPD uses low-order polynomials to approximate the system. This is also known as the Response Surface Methodology (RSM) approach [72]. This approach may be overly cumbersome within computer simulations since the large number of factors to test, the time consuming nature of the evaluations, and the possible non-linear relationships in the model result in the application of more complex test designs and modeling techniques than that of the RPD approach [11]. By exchanging the method used to model the system, the RPD framework can be effectively applied to a wider range of systems. One such extension is the combined-array Multivariate Delta Method approach of Bellucci and Bauer [14]. This approach uses an approximation of the system from an initial experimental design to perform the optimization. This approximation is built using a Radial Basis Function Neural Network (RBFNN) fit of the system response using the control and noise factors as inputs. Optimization is performed on the second-order Taylor series approximation of the MSE model as there is no closed form of the Mean and Variance Model equations. The MDM approach to RPD allows the model to be more adaptive to the system under test than other approaches [14].

The traditional RSM approach to RPD and the MDM approach are similar yet differ in a few key aspects. The MDM approach begins with the same assumptions on the model and noise variables as the traditional RSM approach to RPD [13]. But the MDM approach deviates from the RSM approach in the method it uses to approximate the model of the system. Let $\hat{y}(\mathbf{v})$ be a RBFNN approximation of the model system for input vector $\mathbf{v} = \begin{bmatrix} \mathbf{x} \\ \mathbf{z} \end{bmatrix}$

defined by

$$\hat{y}(\mathbf{v}) = - \sum_{n=1}^N w_n \exp \left[-\frac{1}{2\sigma_n^2} \sum_{m=1}^K (v_m - \mu_m^{(n)})^2 \right] \quad (4.7)$$

where w_n is the top layer weights, σ_n^2 is the $n^{text{th}}$ the spread value for the Radial Basis Function (RBF) neurons, and $\mu_m^{(n)}$ is $n^{text{th}}$ center vector for the $m^{text{th}}$ feature. The second-order Taylor series approximation, $T_2(\hat{y}(\mathbf{v}))$ of the network about the point

$\mu_{\mathbf{v}} = \begin{bmatrix} \mathbf{x} \\ \mu_{\mathbf{z}} \end{bmatrix}$, is derived by way of the Multivariate Delta Method [21]:

$$T_2(\hat{y}(\mathbf{v})) = \hat{y}(\mu_{\mathbf{v}}) + \nabla \hat{y}(\mu_{\mathbf{v}})'(\mathbf{v} - \mu_{\mathbf{v}}) + \frac{1}{2}(\mathbf{v} - \mu_{\mathbf{v}})'(\hat{\mathbf{H}}(\mu_{\mathbf{v}}))(\mathbf{v} - \mu_{\mathbf{v}}) + \epsilon. \quad (4.8)$$

Noting that the $\mathbf{x}$ partition of $\mathbf{v}$ are constants, the mean and variance models of the approximation reduce to

$$E[T_2(\hat{y}(\mathbf{v}))] = \hat{y}(\mu_{\mathbf{v}}) + \frac{1}{2}\text{tr}(\hat{\mathbf{y}}_{\mathbf{zz}}(\mu_{\mathbf{v}})\mathbf{\Sigma}_{\mathbf{z}}) \quad (4.9)$$

and

$$\text{Var}[T_2(\hat{y}(\mathbf{v}))] = \hat{\mathbf{y}}_{\mathbf{z}}(\mu_{\mathbf{v}})' \mathbf{\Sigma}_{\mathbf{z}} \hat{\mathbf{y}}_{\mathbf{z}}(\mu_{\mathbf{v}}) + \frac{1}{2}\text{tr}(\hat{\mathbf{y}}_{\mathbf{zz}}(\mu_{\mathbf{v}})\mathbf{\Sigma}_{\mathbf{z}}\hat{\mathbf{y}}_{\mathbf{zz}}(\mu_{\mathbf{v}})\mathbf{\Sigma}_{\mathbf{z}}) + \sigma_{\epsilon}^2 \quad (4.10)$$

respectively [14]. In the equations (4.9) and (4.10), $\hat{\mathbf{y}}_{\mathbf{z}}(\mu_{\mathbf{v}})$ is the vector of the first-order partial derivatives of $\hat{y}(\mu_{\mathbf{v}})$, $\hat{\mathbf{y}}_{\mathbf{zz}}(\mu_{\mathbf{v}})$ is the matrix of the second-order partial derivatives of $\hat{y}(\mu_{\mathbf{v}})$, $\mathbf{\Sigma}_{\mathbf{z}}$ is the covariance matrix of the noise factors, and σ_{ϵ}^2 is the variance of the random error term ϵ .

The hyperparameters of the RBFNN, namely the spread and number of neurons, are critical to the model's quality. When either of these parameters is not properly selected, the RBFNN is less likely to approximate the true system response. For the selection of RBFNN hyperparameters, Bellucci and Bauer suggest a grid search with a k-fold cross-validation procedure to find the hyperparameters that minimize the MSE across hold out samples. The networks replicate the finding of Bellucci and Bauer that the RBFNN training process routinely selects only 20-50% of the training data for training vectors in the final network [14]. In order to provide for bounds of the search space of the spread

parameter of (1, 10), the control factors are coded to the $[-1, 1]$ range, while the noise factors are scaled to zero sample mean and unit sample variance.

Once the RBFNN is adequately trained to the system, a Taylor-series approximation is used to determine the robust parameters. This requires calculations of both the gradient vector and the Hessian matrix for a given control and noise factor vector $\mathbf{v}$. The k^{th} element of the gradient vector for input vector, v_k , is

$$\frac{\partial \hat{y}}{\partial v_k} = - \sum_{n=1}^N \frac{w_n (v_k - \mu_k^{(n)})}{\sigma_n^2} \exp \left[-\frac{1}{2\sigma_n^2} \sum_{m=1}^K (v_m - \mu_m^{(n)})^2 \right] \text{ for } k = 1, \dots, K \quad (4.11)$$

while the i, j^{th} element (for $i, j = 1, \dots, K$) of the Hessian matrix for the RBFNN output is

$$\frac{\partial^2 \hat{y}}{\partial v_k^2} = \begin{cases} \sum_{n=1}^N \frac{w_n (v_i - \mu_i^{(n)})^2 - \sigma_n^2}{\sigma_n^4} \exp \left[-\frac{1}{2\sigma_n^2} \sum_{m=1}^K (v_m - \mu_m^{(n)})^2 \right] & \text{for } i = j \\ \sum_{n=1}^N \frac{w_n (v_i - \mu_i^{(n)})' (v_j - \mu_j^{(n)})}{\sigma_n^2} \exp \left[-\frac{1}{2\sigma_n^2} \sum_{m=1}^K (v_m - \mu_m^{(n)})^2 \right] & \text{for } i \neq j \end{cases} \quad (4.12)$$

The MSE of the RBFNN for the training data is used for the approximation of σ_ϵ^2 in (4.10).

4.2.2 Noise Framework for RPD in Anomaly Detection.

A key aspect of RPD is the control of the noise factors during data acquisition. This allows for interactions between the control and noise factors to inform the optimization of the control factors. Thus, the application of RPD in HSI anomaly detection requires a definition of noise factors. The expense and complication of additional data acquisition complicates a pure application of RPD. For example, it is not feasible to collect images with every combination of illumination, weather conditions, and environmental setting. Furthermore, it would be infeasible to observe all possible anomalies. However, one can assume noise as stems from two primary sources: the background and the anomalies. Noise provided by the background is normally characterized as clutter and manifests itself statistically as non-homogeneous data. This complicates the estimation of the background statistics. The anomalies, characterized as statistical outliers, also must be isolated from the estimation of the background statistics.

The methodology in this chapter builds on the work of Mindrup *et al.* for the development of a noise framework for the application of RPD to HSI [69]. Their work uses three noise factors to describe the noise in each image: the Fisher Ratio, the Percent Target, and the Number of Clusters. The first two factors describe the noise from an anomaly standpoint, while the third is a characterization of the background. The first of the noise factors, z_1 , is based on Fisher's Ratio as a measure of class separability [31]. Since Fisher's Ratio is a univariate measure, Mindrup *et al.* proposes using the average of the Fisher Ratio from each of the $k = 1, \dots, K$ bands in the i^{th} image, resulting in the metric

$$z_{1i} = \frac{1}{K} \sum_{k=1}^K \left(\frac{(\mu_{a_k} - \mu_{b_k})^2}{\sigma_{a_k}^2 + \sigma_{b_k}^2} \right) \quad (4.13)$$

where $(\mu_{a_k}, \sigma_{a_k}^2)$ and $(\mu_{b_k}, \sigma_{b_k}^2)$ are the mean and variance of the anomalous and background spectral signatures, respectively. The second noise factor described by Mindrup *et al.* is the proportion of the image labeled as an anomaly. If we have the number of anomalies labeled in the i^{th} image, $n_{a,i}$, and the corresponding number of background pixels, $n_{b,i}$, the proportion of anomalies is simply their ratio

$$z_{2i} = \frac{n_{a,i}}{n_{b,i}}. \quad (4.14)$$

The third noise factor included by Mindrup *et al.* is the number of clusters determined via the X-Means algorithm of Pelleg and Moore [76]. The X-means algorithm is an extension of the k-means algorithm which uses the Bayesian Information Criterion (BIC) to determine the optimal number of clusters present in a dataset. Mindrup *et al.* acknowledge that these three noise factors are only a subset of the many available noise characterizations of an image.

It is important to note that the benefit of any noise framework in RPD is the identification of noise factors that result in significant interactions with the control factors. This enables the optimization phase of RPD to identify parameters that are robust to the noise factors. Without these interactions RPD can only identify the variance as a result of

the overall noise present, but the noise factors have no impact on the identification of robust parameters.

4.3 Mahalanobis Masking

An inherent complication in the field of anomaly detection is the undefinable nature of the anomalies themselves [98]. In a laboratory setting, the datasets studied contain only a small portion of the set of anomalies that may be observed in real-world applications. During algorithm development, parameters are identified to produce the best algorithm performance on this small set of labeled datasets with the expectation that the results will also extend to real-world operating conditions. While it might be possible to collect additional samples of the background expected in a real-world scenarios, we will always be limited to a subset of the anomalies we can examine. Thus, this approach to algorithm development is susceptible to producing parameters that may be robust to changes in the background, yet not robust to the complete set of anomalies that could be encountered. To remedy this, proposed is a method for exposing the algorithm to a much larger collection of anomalies during development and possibly identifying more robust parameter settings. Recall from Chapter 3, Mahalanobis Masking is a technique used to adapt the spectral characteristics of anomalies in HSI to more resemble the background spectral characteristics. This approach moves the spectral signature of an anomaly along the gradient defined by the covariance structure of the background signatures. This gradient defines the path of greatest descent with respect to the Mahalanobis distance of the anomaly from the background. The background is assumed to be Multivariate Gaussian represented by a mean, μ_B , and covariance matrix, Σ_B ; both estimated from the labeled background data. With an observed anomalous vector $\mathbf{x}_0$, a new masked version $\mathbf{x}_{i+1}^-$ after $i + 1$ iterations is created using

$$\mathbf{x}_{i+1}^- = \mathbf{x}_i^- - \alpha \Sigma_B^{-1}(\mathbf{x}_i^- - \mu_B). \quad (4.15)$$

The value of α is a step size parameter that controls the distance between the spectra of $\mathbf{x}_{i+1}^-$ and $\mathbf{x}_i$. The process creates anomalies which are more difficult to detect yet are still related to the original, real-world experimental anomalies.

In order to generate the reverse effect, a negative α can be used to create anomalies that are easier to detect than those originally present in the image. However, there is a slight modification that must be made to (4.15) to ensure anomalies that are highly anomalous move at a slower pace than the less anomalous signatures. Otherwise, highly anomalous signatures become overly distorted before appreciable changes occur in less anomalous signatures. This is accomplished by scaling the step size by the inverse Mahalanobis distance of the pixel:

$$\mathbf{x}_{i+1}^+ = \mathbf{x}_i^+ - \frac{\alpha}{\sqrt{(\mathbf{x}_i^+ - \mu_B)\Sigma_B^{-1}(\mathbf{x}_i^+ - \mu_B)'}}\Sigma_B^{-1}(\mathbf{x}_i^+ - \mu_B). \quad (4.16)$$

Using both masking directions, creates a new dataset of images that contain anomalies from the original observations ($\mathbf{x}$), anomalies that are slightly more difficult to detect ($\mathbf{x}^-$) due to their proximity to the background, and anomalies that are easier to detect ($\mathbf{x}^+$). A simple example of this is seen in Figure 4.1. This figure has a simple two-dimensional probability density function (pdf) of the background distribution represented in the contour lines with 5 anomalies represented by the dots. The masked version of these anomalies are either closer, as is the case with the x^- points, or further, x^+ , from the pdf contours of the background pdf. We see that the more anomalous a point is, represented by being further from the contours of the distribution, the larger the deviation the masked points are from the original anomaly. Together, these new anomalous data synthetically triples the number of anomalies the algorithm is exposed to during development without needing the direct acquisition of new data. This negates the costs and complexity associated with acquiring additional data.

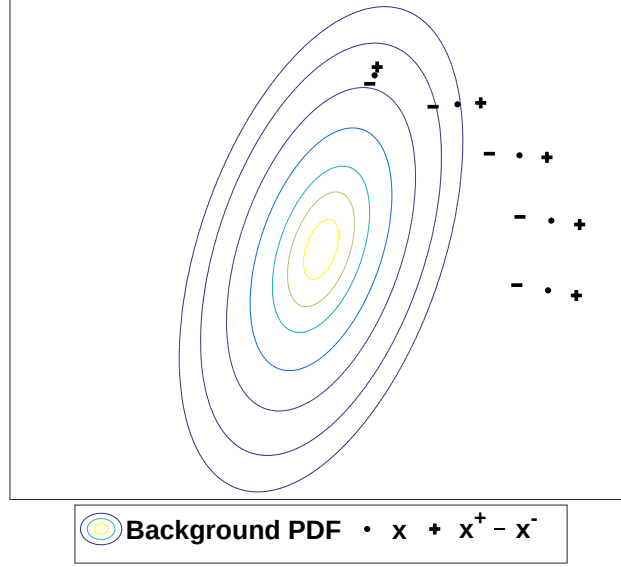


Figure 4.1: Mahalanobis Masking creating variation of the original anomalous signals (x) both more (x^-) and less (x^+) difficult.

4.4 Noise Framework Incorporating Mahalanobis Masking

In light of the observations presented in Chapter 3, some additional changes to the noise framework are proposed. Recall, the use of Euclidean distance fails to account for the characteristics of spectral signatures whereas this difference is better viewed through Mahalanobis distance. The proposed Fisher Ratio for HSI of Mindrup *et al.* [69] is essentially a weighted Euclidean distance between the mean vectors of the background and anomalies. The primary purpose of this noise factor is to measure the difficulty of detection for the image. Thus, instead of using the average of the Fisher Ratio for each of the K bands in an image, using the Area Under the Curve (AUC) score resulting from applying the Reed–Xiaoli (RX) algorithm [81] provides a more direct measure of detection difficulty. This is because averaging the Fisher scores over each spectral band may hide characteristics that are actually indicative of easily detectable anomalies. The AUC score is a better measure of the anomaly detection difficulty within an image. This

RX AUC score is used as noise variable, z_1 , in lieu of the Fisher Ratio. Because the AUC is impacted by the effects of Mahalanobis Masking, each level of masking results in different levels of z_1 .

A further proposed change to the noise framework incorporates the use of Mahalanobis distance in the cluster determination. The approach of Mindrup *et al.* uses the X-means algorithm [76] which determines the number of clusters based on the minimum BIC. As the X-means algorithm uses K-means to cluster the data, it is based off a Euclidean distance metric. Instead, using Gaussian Mixture Models to replace the K-means clustering algorithm provides better cluster determination. This allows for the determination cluster number based on the covariance structure of the spectral signatures, which are more representative of material properties, instead of the reflectance values, which are more representative of illumination. The number of Gaussian Mixture model clusters which returns the minimum BIC in the background of the image is used as z_3 . The noise factor z_2 is left unchanged. Note that since z_2 and z_3 are not impacted by the masking of anomalies, these scores are constant for the levels of masking within an image. Each of the noise variables z_1 , z_2 , and z_3 are normalized to zero mean and unit variance across all image/masking pairs before inclusion into the model.

The direct implementation of Mahalanobis Masking into the noise framework is accomplished with an additional noise variable z_4 . The structure of this variable is slightly different from the other noise variables. Instead of normalizing to zero mean and unit variance, the levels of masking are coded with values 0, -1 , 1 for $\mathbf{x}$, $\mathbf{x}^-$, and $\mathbf{x}^+$, respectively. Coding of the non-masked anomalies at 0 is critical to ensure that the mean model is built on the unmasked anomalies. Recall that the Mean Model applies the expectation operator on the noise variables; thus, with the unmasked images coded at the mean of noise variable z_4 , ensuring that the optimization of the mean is most representative of observed anomalies. The impact of Mahalanobis Masking occurs

primarily on the Variance Model of the detector. Regions where performance is degraded on either the more or less difficult masked anomalies will cause an increase in variance estimation due to noise. This is the desired effect of Mahalanobis Masking inside the RPD framework. Since changes to the anomalies through Mahalanobis Masking are slight, we desire the detector to be robust to the numerous other anomalous characteristics that may be present in extended operation. When included in the bivariate optimization approach, this aids in the identification of robust parameters that both provide high performance on the original images, as well as robustness to the presence of anomalies not directly observed in the original data.

4.4.1 Mahalanobis Masking as Adversarial Behavior.

Mahalanobis Masking as an approach to RPD can also be viewed as an adversary trying to avoid detection. Where ground truth-labeled images provide a simulated real-world collection scenario with known anomalous spectral signatures, an adversary may want to conceal anomalous signatures within the background. While there are many methods of concealment, the most efficient would take advantage of variation in background spectral signatures (e.g. appear more similar to organic material). Since there are numerous anomaly detection methods that take advantage of covariance relationships in spectral signatures to identify potential anomalies, an adversary aiming to conceal anomalies would try to decrease the dissimilarity between anomalies and background signatures. A method that traverses the gradient of the distance measure used by a particular anomaly detector would be the most efficient means of concealing an anomaly. Thus, Mahalanobis Masking could also be viewed as characterizing the adversarial behavior an algorithm should be robust against. Extending this viewpoint into the inclusion of Mahalanobis Masking into the noise framework of RPD drives a robustness into the algorithm that combats this adversarial behavior.

4.5 Measuring Performance

There are numerous measures to determine performance of anomaly detectors, such as those outlined in Section 2.5. Typically, the RPD approach requires a single performance measure for optimization. In this chapter, we use the measure of performance of Matthew’s Correlation Coefficient (MCC) [65]. Recall, that MCC uses the formula

$$\text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}. \quad (4.17)$$

This measure is chosen for the balance it places on the False Positive (FP) and False Negative (FN) levels [3]. As we often have class imbalance in anomaly detection applications, the mis-identification of both background and anomalies must be balanced. However, by no means is this the only measure of algorithm performance and, in practice, multiple metrics may need to be balanced in order to truly characterize the overall performance of an algorithm. Techniques from multi-objective optimization can provide approaches when required by the application.

4.6 Images

The data used in this chapter is from Hyperspectral Digital Imagery Collection Equipment (HYDICE) sensor, which is a push-broom, aircraft-based sensor [82]. The images come from the Forest Radiance I and Desert Radiance II collection events. Ten images in total are used in the following section, each of which have been split in half to create twenty image halves. These image halves are then assigned equally to a training and test set as seen in Table 4.1. This table also presents the original noise variables of Mindrup *et al.* as a comparison to our new, proposed noise variables.

4.7 RPD Experiments using AutoGAD

This section, using the improved noise framework discussed above and the MDM approach for RPD, extends the work Mindrup *et al.* For the purpose of comparison, their

Table 4.1: Calculated noise variables of images from Mindrup *et al.* calculated on the training and test image sets [69].

	Image	Half	Fisher Ratio (z_1)	Percent Target (z_2)	Number of Clusters (z_3)
Training Set	1D	Top	1.7797	0.0043	3
	1F	Top	0.4335	0.0028	5
	2D	Top	0.0957	0.0225	4
	2F	Top	0.9633	0.0247	7
	3D	Bottom	1.4299	0.0288	3
	3F	Top	0.265	0.0084	8
	3F	Bottom	0.2153	0.0085	5
	4	Bottom	2.6382	0.0034	4
	5	Top	0.2658	0.0053	6
	5F	Top	0.1991	0.0078	10
Test Set	1D	Bottom	1.6265	0.0156	3
	1F	Bottom	0.3148	0.0275	5
	2D	Bottom	0.1762	0.0063	3
	2F	Bottom	0.9311	0.0094	7
	3D	Top	0.1695	0.0109	3
	4F	Top	0.0826	0.0392	7
	4F	Bottom	0.0779	0.0033	8
	4	Top	1.4093	0.0046	6
	5	Bottom	1.8451	0.0078	4
	5F	Bottom	0.7412	0.0052	7

image training and test structure is replicated, as well as the use of the detector AutoGAD.

An outline of the AutoGAD algorithm is found in Section 2.4.5.1. Recall that AutoGAD is primarily controlled through eight parameters described in Table 4.3. For this chapter experimentation focused on seven of the eight parameters, excluding “Dimensionality Adjustment” since it is a discrete variable and the MDM process requires continuous variables. The value remains at the previously prescribed level of -2 [103].

To show the advantage of Mahalanobis Masking within RPD, the RPD process is applied using two noise frameworks. The first is similar to that of Mindrup *et al.* but uses

Table 4.2: Training images with new calculated noise variables.

Image	Half	RX AUC (z_1)			Percent Target (z_2)	Number of Clusters (z_3)
1D	Top	0.973	0.966	0.972	0.0043	6
1F	Top	0.759	0.693	0.759	0.0028	3
2D	Top	0.957	0.945	0.960	0.0225	4
2F	Top	0.884	0.865	0.883	0.0247	7
3D	Bottom	0.703	0.682	0.709	0.0288	5
3F	Top	0.984	0.967	0.984	0.0084	5
3F	Bottom	0.913	0.877	0.913	0.0085	3
4	Bottom	0.939	0.879	0.939	0.0034	3
5	Top	0.968	0.962	0.968	0.0053	7
5F	Top	0.842	0.827	0.844	0.0078	9
	Anomaly States (z_4)	None	More Difficult (-1)	Less Difficult (1)		

only the new definitions for the noise factors z_1 , z_2 , and z_3 . The second framework uses the additional information gathered from the implementation of Mahalanobis Masking in factor z_4 . The new variable values for the training image halves can be found in Table 4.2.

Using the ten training image halves, a 200-point Space Filling Design in the bounded parameter space (Table 4.3) is created using a combined-array design with noise factors from the ten test image halves with positive and negative levels of masking. The negative level of masking is set to 10,000 iterations with an $\alpha = 0.5$. Using the mean amount of change to the anomalies' Mahalanobis distances, the positive level of masking corresponded to iterations that result in the opposite level of distance to the negative masking. This results in a total of 2,000 data points corresponding to the unmasked images and an additional 4,000 data points with masking. Since AutoGAD has some stochasticity in the algorithm due to FastICA [44], five replications at each test point are collected and the mean of the replications at each test point is kept as the response for

Table 4.3: AutoGAD parameters and bounds.

Parameter	Description	Initial Setting	Lower Bound	Upper Bound
Dimensionality Adjustment	Adjustment to the dimensionality determination from the MDSL method	-2	-2	-2
Max Score Threshold	Threshold on maximum observed score in a component to be kept as a anomalous component	6	6	14
Bin Width (SNR)	Bin width (x/number of pixels) for histogram formation for potential anomaly component selection	2100	1000	4,000
PA SNR Threshold	PA SNR Threshold for Declaring a Potential Anomaly Component	2	0	5
Bin Width (Identify)	Bin width (x/number of pixels) for histogram formation	2100	1000	4,000
Smoothing Iterations (High)	Number of IAN Smoothing Iteration for Components with High PA SNR	100	50	120
Smoothing Iteration (low)	Number of IAN Smoothing Iteration for Components with Low PA SNR	20	0	50
PA SNR Smoothing Threshold	PA SNR Threshold for Deciding between IAN Iterations	5	0	7

each parameter setting. This is analogous to the process displayed by previous approaches for RPD in HSI [69, 103]. After the optimal parameters are determined from the model, the settings are then verified on the test image halves that are not used in the parameter determination process. These test images are left unmasked in order to most accurately display the performance in real-world scenarios. However, further analysis of the algorithm's stability can be examined through the process outlined in Chapter 3.

4.7.1 Neural Network Structure.

After the collection of all data points, the spread parameter and the number of neurons allowed in the RBFNN are determined using a grid search with a cross-validation procedure. The optimal parameters for the network with the masking images included

results in a spread parameter for the network of 3 and 250 neurons. Whereas the best spread parameter without the masking images is 6 and 200 neurons. It should be noted that the neighborhood near these optimal values was relatively stable in the cross-validation MSE, and thus, the exact values are not critical, especially with regards to the number of neurons used.

4.7.2 *Optimal Settings.*

After training the RBFNN with the determined hyperparameters, optimization of the parameters is carried out using MATLAB's built-in *patternsearch* function with implementation of Greedy Randomized Adaptive Search Procedures (GRASP) to decrease the chance of finding a local optima [103]. The GRASP heuristic is used to guard against the identification of a local optima since the surface of the RBFNN can be non-linear. The results of this optimization are found in Table 4.4.

Table 4.4: Recommended robust operating points for AutoGAD as determined without masking and with Mahalanobis Masking.

Parameters	No Masking	Mahalanobis Masking
Dimensionality Adjustment (x_1)	-2	-2
Max Score Threshold (x_2)	7.216	10.286
Bin Width SNR (x_3)	3,970	1,010
PA SNR Threshold (x_4)	4.90	0.05
Bin Width Identify (x_5)	2,619	3,904
IAN Iterations High (x_6)	51	50
IAN Iterations Low (x_7)	50	50
Low PA SNR (x_8)	5	0

Using the two noise frameworks, the optimal solutions found significantly differ on four parameters: Max Score Threshold (x_2), Bin Width SNR (x_3), PA SNR Threshold (x_4), and Bin Width Identify (x_5). The impact of these parameter differences appears to make the algorithm more susceptible to nominating components for anomaly consideration, as indicated by the smaller Bin Width SNR setting with compensation from the increased

Max Score Threshold. This seems to be a result of the more difficult anomalies being on components that are not nominated under the no masking parameter values. Both approaches agree that the IAN Iterations should be at 50, regardless of the PA SNR level, and thus, the value of the x_8 parameter has no impact on the algorithm.

Figures 4.2–4.5 present a univariate view of each of the four variables for which the processes recommended different settings. In each Figure, the other seven variables are set to a prescribed operating point as in Table 4.4, while the variable under study is allowed to vary over its coded values from -1 to 1. The values of the MSE, Mean, and Variance Model are presented to show the impact of the Masking framework. In each Figure, a very different Variance Model for each variable between the two frameworks, with the Mean Model also being significantly impacted for Bin Width SNR (x_3). Such changes to the estimated models result in different operating points for each of the variables. Examining the plots in Figures 4.2 and 4.3, we see that increased noise due to masking significantly changes the optimal point with respect to the Mean Model. Whereas in Figures 4.4 and 4.5, the optimal point is indicate more impact in the the Variance Model. For the other three variables, x_6 , x_7 and x_8 , the impact on model estimations does not change the overall prescribed robust operating point.

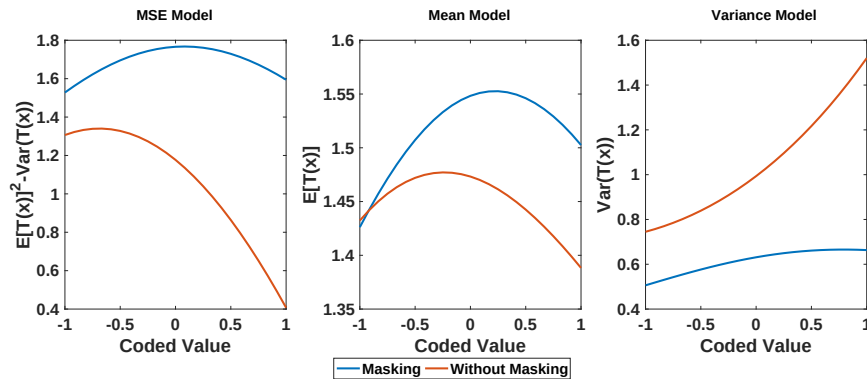


Figure 4.2: Impact of Control Variable x_2 at robust parameter setting.

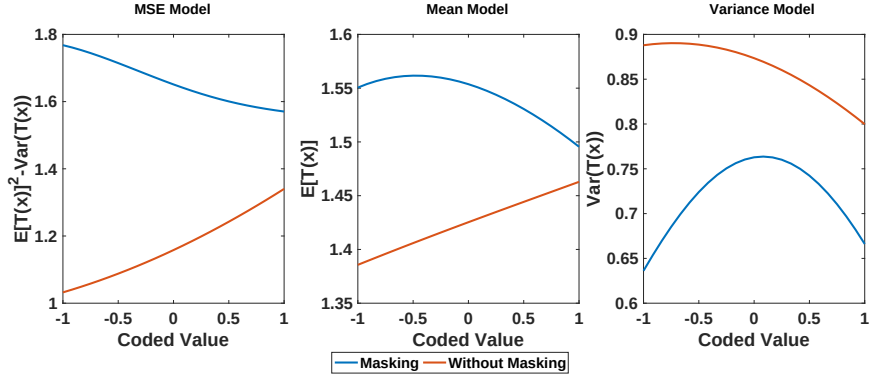


Figure 4.3: Impact of Control Variable x_3 at robust parameter setting.

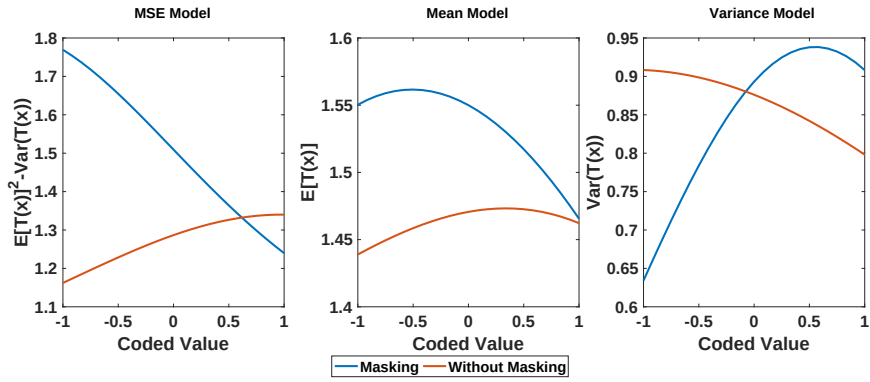


Figure 4.4: Impact of Control Variable x_4 at robust parameter setting.

4.7.3 Test Image Performance.

The 10 hold out test images are used to determine the performance of the parameters. The calculated noise variables for the image halves can be seen in Table 4.5. Using the prescribed parameter settings and five replications of the algorithm on each image, the resulting performance measures are found in Table 4.6. The results are indicative of an added tendency to nominate additional components for anomaly thresholding with a

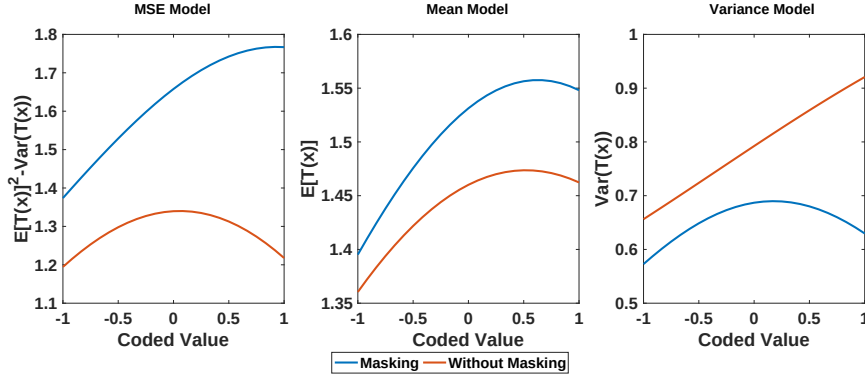


Figure 4.5: Impact of Control Variable x_5 at robust parameter setting.

decrease in overall MSE of 0.22 with Mahalanobis Masking over the previously prescribed parameter values. This primarily results from a much higher mean response for the MCC values on testing images with a marginal increase to the variance of the response. This is expected since the Mahalanobis Masking allows for improved estimation of the noise in the algorithm responses due to the characteristics of the anomalies that may be present.

Table 4.5: Test images with new calculated noise variables.

Image	Half	RX AUC (z_1)	Percent Target (z_2)	Number of Clusters (z_3)
1D	Bottom	0.962	0.0156	7
1F	Bottom	0.931	0.0275	5
2D	Bottom	0.965	0.0063	5
2F	Bottom	0.913	0.0094	3
3D	Top	0.924	0.0109	4
4F	Top	0.919	0.0392	4
4F	Bottom	0.869	0.0033	4
4	Top	0.933	0.0046	5
5	Bottom	0.924	0.0078	7
5F	Bottom	0.913	0.0052	8

Table 4.6: Results of robust operating points on test images.

	Initial Settings	No Masking	Mahalanobis Masking
Mean (MCC, Target =1)	0.25	0.35	0.59
Var (MCC)	0.0266	0.0595	0.0635
MSE	0.5240	0.3637	0.1071

4.8 Conclusions

This chapter shows how using Mahalanobis Masking inside a RPD framework can result in improved parameter selection in anomaly detectors. This is accomplished using Mahalanobis Masking to create new anomalies that are either easier or more difficult to detect than the anomalies in directly observed images. In doing so, the improved RPD process is then able to expose the algorithm to a wider range of the set of anomalies that may be seen in extended operation. As such, the RPD approach can identify parameters that increase performance on hold-out testing images.

V. Truncated Extreme Value Anomaly Thresholding

5.1 Introduction

The field of anomaly detection is concerned with the identification of a subset of data that does not truly belong within the greater dataset [22]. These could be signals, objects, or data that are unexpectedly mixed with a larger set. The unexpected nature of the anomalies depends on the application at hand; however, the objective of an anomaly detector is to separate the expected, or background data, from the unexpected, or anomalous data. This is carried out in a number of ways in applications such as intrusion detection in computer networks, fraud detection in financial transactions, or malignant tumors in an MRI image [22].

Anomaly detectors typically provide a score for each data point. These scores are used to assess how likely a point is to be an anomaly. When a detector is operating properly, we expect detector scores of anomalies to be stochastically larger than that of the non-anomalous background. Thus, observations assigned higher detector scores are more suspect of being anomalous. A key piece in many anomaly detectors used in an autonomous manner is determining an anomaly threshold for the scores. The anomaly threshold forms the decision boundary for the scores [79]. In application, there are a number of methods for the determination of an anomaly threshold based on output from a detector. The most common and computationally simple method is selecting a top percentile (e.g. top 1%) of the detector's output to label as anomalies. While simple, this method is not adaptive to situations where anomalies populate a proportion of the dataset at a level near or above the percentile selected. This inability to adapt to the presence of anomalies can cause large deviations from the desired performance characteristics. For example, in a dataset with a large proportion of anomalies, perhaps more than the percentile selected, the threshold will be too high. Another common anomaly thresholding

method is based on building a histogram of the scores. The anomaly threshold becomes the value attached to the first bin above the median with a frequency of zero [24]. While still relatively computationally simple, this zero-bin histogram method is highly dependent on the individual scores and on the pre-selected bin width [74]. Beyond these threshold approaches, some detectors are built under statistical assumptions to provide parametric characterization of the background distribution, such as Reed–Xiaoli (RX) in the field of hyperspectral imaging anomaly detection that assumes Gaussian noise [81]. When the assumptions are met, parametric thresholding is very effective. However, when the required assumptions are violated, these methods can return inconsistent results [104].

There are a number of desirable characteristics for anomaly thresholding approaches such as robustness to the presence of anomalies and adaptability to different background behaviors. Another characteristic is commonly known as Constant False Alarm Rate (CFAR). A false alarm, or false positive, occurs when an anomaly detector incorrectly identifies a non-anomalous data point as anomalous. This occurs when the anomaly score of the data is above the anomaly threshold. While it is true that even truly non-anomalous data can display extreme behavior, it is critical in some applications to minimize and control the occurrence of false alarms. A threshold that possesses CFAR returns false alarms, or incorrectly declared data, at a constant proportion of the data being tested. A controllable CFAR threshold allows the output to be tuned to an accepted level of false alarms. The false alarm rate depends on a number of factors such as the nature of the dataset and the quality of the anomaly detector, but one of the most critical aspects is determining the anomaly threshold. In order to estimate the CFAR threshold, an anomaly threshold method must first characterize the background in some effective manner. This background characterization depends on the the application and the methods used, but the quality of the characterization is closely related to overall detector performance [64]. There are many detectors that are theoretically CFAR when assumptions about the data

are met [70], including RX [81]. However, issues with these detectors arise when the assumption that grant CFAR are not met [1, 104].

Another approach to gaining CFAR performance in a detector involves characterizing only the extreme values of the anomaly scores, instead of the entirety of the background [61]. An extreme value is one that is far from the median of the data (i.e. located in the tail of the distribution). The statistical field of Extreme Value Theory (EVT) provides proof under fairly general conditions that the extreme values of the background converges in distribution to the Generalized Pareto Distribution (GPD) [28]. A number of approaches based on this principle are present in the literature. Manolakis *et al.* [61] present one such method to introduce CFAR into anomaly detection algorithms previously without the CFAR property. However, their method fails to account for the anomalies contaminating the distribution of the tail. Broadwater *et al.* [19] improve on this method by utilizing a Goodness-of-Fit (GoF) test to eliminate possible anomalies during the fitting of the GPD. They show that this method, known as Adaptive Threshold Estimation via Extreme Value Theory (ATEEVT), is useful in Hyperspectral Imaging (HSI) anomaly detection. However, as we will discuss later, ATEEVT relies on the assumption that the scores of the anomalies are at least partially separable from background scores. This may not always be the case in applications where the detector has difficulty separating the anomalies from the background and can impact the anomaly threshold estimation.

In this chapter, we present a new method for providing an asymptotically CFAR thresholding method for detectors that lack such capability natively. This method extends the work discussed above using EVT to characterize the extreme values of the background data. However, we propose the use of more powerful GOF tests, the inclusion of the truncation point in the parameter estimation process, and an iterative elimination of potential anomalies. The remainder of the chapter is organized as follows. Section 5.2 reviews EVT, truncated distributions, and GoF tests. Section 5.3 develops the new method

in its entirety and, Section 5.4 presents results on simulated and applied detector output. Section 5.5 closes the chapter with brief conclusions.

5.2 Background

The following is a brief overview of the contributing fields used in this chapter. We begin with EVT, which is the theory motivating the examination of the extremes of the detector scores. We follow with a discussion of the use of truncated distributions. Finally, GoF tests for empirical data are discussed.

5.2.1 *Extreme Value Theory.*

EVT is concerned with the behavior of the tails of a distribution, or the extremes [28]. Often the distribution under study is unknown with only sample observations available. The primary area of research in EVT focuses on the maximum of disjoint sequences, or epochs, of the data. An example of an epoch may be a month for daily temperature; then the maximum observed monthly temperature is the extreme for the each epoch. Fisher and Tippett [37] and also, independently, Gnedenko [39] derived the conditions that must be present for the extremes of a distribution to converge in distribution to one of the three Extreme Value Distributions: *Gumbel*, *Fréchet*, and *Weibull* [28, p. 6]. These three distributions can be consolidated to form the *Generalized Extreme Value (GEV)* distribution. A key aspect to EVT is the definition of the domain of attraction of the GEV. A distribution, F , is said to be in the domain of attraction of the GEV if the extremes converge in distribution to a GEV with a valid shape parameter. For example, when the underlining distribution is normal, the tail of the distribution converges to the *Gumbel* distribution [84]. This EVT method is used in a number of anomaly detection methods, e.g. [83, 93]. However, while the above EVT approach is useful in some applications, if no epoch definition is readily apparent in the data, the selection of epoch size may at times be arbitrary and problematic [40].

In light of the possibly problematic epoch requirement, a subsequent development in EVT stems from the work of Balkema and de Haan [4] and Pickands [77] with a method to characterize the tail of a distribution using *points over threshold* (POT) techniques. This alleviates the need for arbitrarily defining epochs within the observed data. The Pickands-Balkema-De Haan (PBD) Theorem results in a known GPD for the data in the tail of a distribution with only the requirement that this distribution be continuous and in the general class of distributions in the domain of attraction of the GEV [28].

Pickands-Balkema-De Haan Theorem [28]. *Assume random variable X is independently sampled from a continuous, yet unknown, distribution $F_X(x)$ in the domain of attraction for the GEV Distribution. Then, for some high threshold u ,*

$$P(X > x - u | X \geq u) = 1 - G(x)$$

where $G(x)$ is a GPD.

The Cumulative Distribution Function (CDF) of the GPD is

$$G_{GPD}(x; k, \sigma, u) = \begin{cases} 1 - \left(1 - \frac{kx}{\sigma}\right)^{1/k} & \text{if } k \neq 0, \sigma > 0 \\ 1 - e^{-x/\sigma} & \text{if } k = 0, \sigma > 0 \end{cases} \quad (5.1)$$

which is equivalent to a Pareto Distribution when $k > 0$ and the Exponential distribution when $k = 0$. The POT EVT approach has been shown to characterize the tail behavior of a number of distributions [32] and in application areas, such as radar thresholds [73], weather phenomena [43], and various empirical datasets [26].

It should be noted that given empirical data, the determination of the threshold u for the PBD Theorem can be difficult unless known *a priori* and is critical to this type of analysis [85]. However, some of the issues with threshold selection are alleviated with the threshold stability property of POT data. Threshold stability is the property that if for $u > 0$ the distribution $F(x|x > u)$ is GPD, then for any $u' > u$ the distribution $F(x|x > u')$ is

also distributed GPD [85]. This property allows some imprecision in the determination of the threshold if the amount of data available allows for it [26]. Thus, a common approach to setting the value of u is to use a top empirical percentile (e.g. top 5%) of the data. This allows the fitting of the GPD to data that is expected to display extreme behavior; however, the percentile for selecting the appropriate u value is distribution dependent. There are numerous other approaches to determining u empirically [85]. However, these approaches are computationally complex and may not be compatible with the approach used to detect the presence of anomalies above u . Thus, this chapter focuses on selecting u as a deterministic percentile of the data in order to provide stability when detecting the presence of anomalies. A future research concern is to determine the compatibility of the truncation methods presented here with more advanced u selection methods.

Regardless of the approach used to determine the POT, the parameters of the GPD are typically also determined from the data. Presented with empirical data from an unknown parent distribution, the fitting of the parameters of the GPD to the extreme data requires one of numerous approaches each with its own properties of bias, variance, and sensitivity to threshold choice. These methods include Maximum Likelihood Estimation (MLE), weighted MLE, Maximum Goodness-of-Fit, Methods of Moments, and Methods of L-Moments [59].

5.2.2 Truncated Distributions.

In some applications, the data from a distribution is known to only be observed over a portion of the distribution's support. One example of this is the effect of deductibles on the distribution of insurance claims. Losses below the deductible are unobserved resulting in a truncation of in observed insurance claims. This results in what is called "left-truncation". When values above some level are unobserved, we label that "right-truncation". If the untruncated distribution and truncation point, v , are known, generating a right-truncated probability density function (pdf), $f_v(x)$, is found by scaling the pdf of the distribution,

$f(x)$, by the probability of being above v using the CDF of the distribution, $F(v)$:

$$f_v(x) = \begin{cases} \frac{f(x)}{1-F(v)} & \text{if } x < v \\ 0 & \text{if } x \geq v \end{cases}. \quad (5.2)$$

However, often the pdf of either the truncated or untruncated distribution is unknown and must be fit to the data. The truncation of data can have a profound effect on the fitting of a distribution. The estimation of the proportion of unaccounted for data is typically included in the estimation process. As such, the truncation point and the truncated proportion of data additional parameters are added to parameter estimation. An example of this can be seen in Figure 5.1. The data are taken from the extremes of a simulated GPD distribution with right truncation occurring at a value of 1. The red line is fitted without accounting for the truncation of the data and results in a fit that severely underestimates the probability near and above the truncation point. Conversely, the yellow line is from fitting while estimating the density that has been truncated allows for better approximation of the true distribution (blue line), in this case a GPD with a shape and scale parameter of 1.

Later in the chapter, a right-truncated Generalized Pareto Distribution (tGPD) is used to account for the truncation of data suspected to be contaminated with anomalies. The CDF of the tGPD with the truncation point of v is of the form

$$G_{tGPD}(x; k, \sigma, u, v) = \begin{cases} \frac{1 - (1 - \frac{kx}{\sigma})^{1/k}}{1 - (\frac{kv}{\sigma})^{1/k}} & \text{if } k \neq 0, x < v \\ \frac{1 - e^{-x/\sigma}}{e^{-v/\sigma}} & \text{if } k = 0, x < v \\ 1 & \text{if } x \geq v \end{cases} \quad (5.3)$$

for $\sigma > 0$ and $x \geq 0$. This allows for the estimation of the tail of the background distribution based on the characteristics of the extreme background scores below the truncation point while preventing the anomalies above the truncation point from impacting the distribution estimation. This is critical in applications with poor separability between the classes because background scores above the truncation point can have a significant

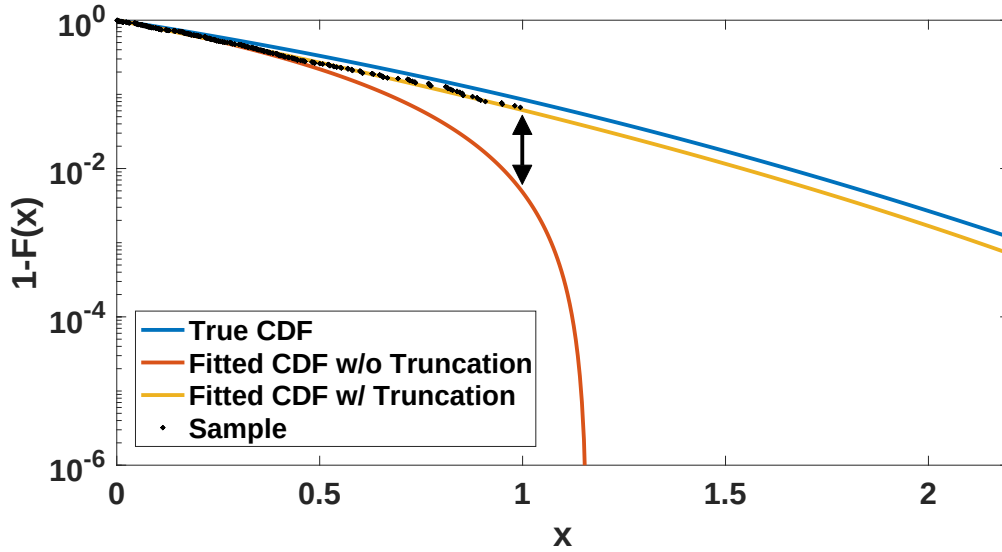


Figure 5.1: Example of Impact on Distribution Fit with and without accounting for truncation.

effect on the estimation of the distribution's untruncated parameters. An example of this is from the data in Figure 5.1. Without estimated the truncation point, the distribution underestimates the probability of being above the threshold by approximately 0.07.

5.2.3 Goodness-of-Fit Tests.

A key aspect to the methodology in this chapter is the characterization of extreme value behavior in the data through an application of a GoF test of the GPD to the extremes. This examination is used to reveal anomalous scores present with in the POT.

There are a number of methods for determining the GoF of a distribution to empirical data. One of the most prominent family of methods is based on the empirical distribution function (EDF) including supremum statistics based tests (e.g. Kolmogorov-Smirnov) and quadratic statistics based tests (e.g. Anderson-Darling) [27, p.100]. These tests all compare the theorized CDF ($\hat{F}(x)$) and EDF ($F_n(x)$) of the data to build test statistics

using the ordered statistics of the data sample, $x_{(i)}$. The EDF of the sample is given by

$$F_n(x_{(i)}) = \frac{1}{n} \sum_{i=1}^n I_{[-\infty, x]}(x_{(i)}) \quad (5.4)$$

where I is the indicator function. The supremum-based tests use the deviations of the CDF and EDF. The most common of this type is the Kolmogorov-Smirnov (KS) test that uses the maximum absolute deviation between the curves defined by

$$D = \sup_i |\hat{F}(x_{(i)}) - F_n(x_{(i)})|. \quad (5.5)$$

A related statistic, attributed to Kuiper [53], traditionally labeled V , which is the sum of the maximum upper and lower deviations, are

$$D^+ = \sup_i (\hat{F}(x_{(i)}) - F_n(x_{(i)})) \quad (5.6)$$

and

$$D^- = \sup_i (F_n(x_{(i)}) - \hat{F}(x_{(i)})) \quad (5.7)$$

respectively.

The quadratic-based Cramér-von Mises (CvM) family of tests are based on a measure of discrepancy [27] given by

$$Q = n \int_{-\infty}^{\infty} \{F_n(x) - \hat{F}(x)\}^2 \psi(x) dF(x). \quad (5.8)$$

The measure of discrepancy accounts for deviation from the theoretical CDF along the entire range of the sample, instead of the few points used by the supremum statistics. The CvM test statistic, W^2 , is the result from $\psi(x) = 1$ while the quadratic Anderson-Darling, A^2 , is a result from $\psi(x) = \{\hat{F}(x)[1 - \hat{F}(x)]\}^{-1}$ [27]. The weighting function of A^2 test results in more weight placed on the tails of the distribution. In some situations, such as financial data [23], it may be more critical to fit one of the tails to the data. In this case, derivatives of the Anderson-Darling test for the upper and lower tails are used [91]. The upper tail test, which we label as A_{up}^2 , uses $\psi(x) = [1 - \hat{F}(x)]^{-1}$, and the lower tailed test,

Table 5.1: GOF statistics for samples $\{x\}$ using the EDF, F_n , theoretical CDF, $\hat{F}(x)$, for samples from a data set of size n . $z_i \equiv \hat{F}(x_i)$ [23].

Name	Label	Description	Computing Formula
Kolmogorov-Smirnov (KS)	D	$\sqrt{n} \sup_x F_n(x) - \hat{F}(x) $	$\sqrt{n} \max\{\max_i [\frac{i}{n} - z_{(i)}], \max_i [z_{(i)} - (i-1)/n]\}$
Kuiper	V	$\sqrt{n} \left(\sup_x \{F_n(x) - \hat{F}(x)\} + \sup_x \{\hat{F}(x) - F_n(x)\} \right)$	$\sqrt{n} \max[i/n - z_{(i)}] + \max_i [z_{(i)} - (i-1)/n]$
Cramér-von Mises (CvM)	W^2	$n \int_{-\infty}^{\infty} (F_n(x) - \hat{F}(x))^2 d\hat{F}(x)$	$\frac{1}{12n} + \sum_{j=1}^n \left(z_{(j)} - \frac{(2j-1)}{2n} \right)^2$
Anderson-Darling	A^2	$n \int_{-\infty}^{\infty} \frac{(F_n(x) - \hat{F}(x))^2}{\hat{F}(x)(1 - \hat{F}(x))} d\hat{F}(x)$	$-n + \sum_{j=1}^n \left[\frac{(2j-1)}{n} (\log z_{(j)} + \log(1 - z_{(j)})) \right]$
Anderson-Darling Upper	A_{up}^2	$n \int_{-\infty}^{\infty} \frac{(F_n(x) - \hat{F}(x))^2}{(1 - \hat{F}(x))^2} d\hat{F}(x)$	$\frac{n}{2} - \sum_{j=1}^n \left[2z_{(j)} + \left(2 - \frac{2j-1}{n} \right) \log(1 - z_{(j)}) \right]$
Anderson-Darling Lower	A_{low}^2	$n \int_{-\infty}^{\infty} \frac{(F_n(x) - \hat{F}(x))^2}{\hat{F}(x)^2} d\hat{F}(x)$	$\frac{-3n}{2} + \sum_{j=1}^n \left[2z_{(j)} - \frac{2j-1}{n} \log(z_{(j)}) \right]$

A_{low}^2 , uses $\psi(x) = \{\hat{F}(x)\}^{-1}$ [2]. A summary of the GOF statistics, as well as their formulas, are presented in Table 5.1.

The determination of GoF with truncated data requires the adaptation of the test statistics. This is depicted in the charts in Figure 5.2 showing the relationships between the known CDF of the simulated data and the empirical CDF. In Figure 5.2a the data is untruncated and the empirical CDF closely follows the true CDF. However, in the presence of truncation, the empirical distribution deviates from the theoretical. This is shown in Figure 5.2b where the data is truncated and the empirical CDF, without correction, departs from the true CDF. The remedy to this deviation is the scaling of the EDF by the proportion of data that is unobserved above the truncation point, as seen in Figure 5.2c.

Chernobai *et al.* present the general approach of adapting GoF tests for left-truncation, which we adapt to the right-truncation scenario [23]. Our approach scales the empirical distribution by the probability of being above the truncation point, ν , providing a closer approximation of the theoretical CDF. Without a known truncation point, the estimation of the value of ν must be done empirically, which we discuss later in the chapter. Normally, when truncation occurs due to the way the data is observed, the maximum observed value can serve as an estimate for ν , but in our case the truncation is a result of data contamination so the maximum value may be anomalous. Thus, ν must be estimated where we have reasonable expectation that the contamination is minimal. Table 5.2 presents the computing formulas of the GoF tests with the truncation of the distribution at ν .

Later in the chapter, these right-truncated GoF tests also determine a threshold where departure from the distribution is suspected. It should be noted that when anomalies are present in the dataset used to fit the distribution, the truncated test statistics can give misleading indications. The method of Broadwater *et al.* relies on a stochastic separation between the background and anomaly distributions [19]. Unfortunately, when the

anomalies are not easily distinguishable by the detector from the background, the fit of the GPD can be biased away from the background. The resulting GoF scores for the background are inflated, which results in a truncation of background data. To remedy this, a small portion of the points with high test statistics can be rejected in an iterative fashion as this prevents large portions of background data from being truncated. With this approach there is a trade-off between the speed of convergence and the ability to ensure convergence. When large portions of extremes are allowed to be truncated during each iteration the variance of subsequent parameter estimations can increase due to small sample sizes.

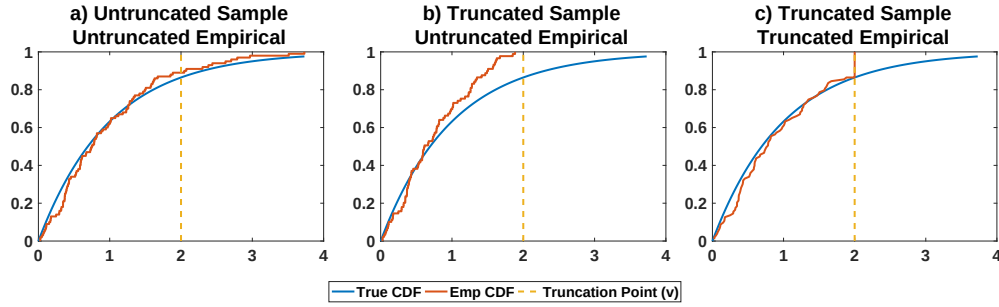


Figure 5.2: The three charts show the effect of ignoring the truncation of data when doing GoF testing.

Each of the GoF test statistics are compared to critical values to determine whether to accept or reject the null hypothesis that the distribution fits with known parameters. However, as in our case, when the shape and scale parameters of the distribution are not specified, the critical values of the test must be modified [42]; further, the presence of anomalies precludes the actual application of a true p-value to the test [19]. The result is that the critical value used to judge test statistics is more of sensitivity parameter instead of a true probabilistic value.

Table 5.2: GOF statistics for samples $\{x\}$ with right truncation at v using the EDF, F_n , theoretical CDF, $\hat{F}(x)$, for samples from a data set of size n . $z_{(i)} \equiv \hat{F}(x_{(i)})$ and $z_v \equiv \hat{F}(v)$

Label	Computing Formula
D_v	$\frac{\sqrt{n}}{z_v} \max \left\{ \sup_j \left[z_{(j)} - \frac{j-1}{n}(z_v) \right], \sup_j \left[\frac{j}{n}z_v - z_{(j)} \right] \right\}$
V_v	$\frac{\sqrt{n}}{z_v} \left(\sup_j \left[z_{(j)} - \frac{j-1}{n}(z_v) \right] + \sup_j \left[\frac{j}{n}z_v - z_{(j)} \right] \right)$
W_v^2	$\sum_{j=1}^n \left[\frac{z_{(j)}}{z_v} - \sum_{j=1}^n \frac{(2j-1)}{(2n)} \right]^2 + \frac{1}{12n}$
A_v^2	$-n + 2n \log(z_v) + \sum_{j=1}^n \left[\frac{(1-2j)}{n} \log(z_{(j)}) - \frac{(1+2(n-j))}{n} \log(z_v - z_{(j)}) \right]$
$A_{v,up}^2$	$\frac{nz_v}{2} + z_v n \log z_v - \sum_{j=1}^n 2z_j - \sum_{j=1}^n \log(z_v - z_{(j)}) (2z_v - \frac{z_v(2j-1)}{n})$
$A_{v,low}^2$	$-\frac{3nz_v}{2} + nz_v \log z_v + \sum_{j=1}^n \log(z_{(j)}) (\frac{(1-2j)z_v}{n}) + \sum_{j=1}^n 2z_{(j)}$

5.3 Methodology

This section presents a methodology for determining a CFAR anomaly threshold by isolating the extreme behavior of the detector's scores. It is assumed that the detector scores are distributed in the general class of the domain of attraction for the GEV distribution. There are three values that the algorithm requires and their relative placement are illustrated in Figure 5.3. The first of these values, u , is the threshold for the POT technique. This value demarcates the extreme scores from the non-extremes. The next value, v , is the right-truncation point where the effect of the anomalies begins to dominate the extreme behavior. The scores above v , both background and anomalies, are excluded during distribution parameter estimation. The region between u and v provide the scores to fit a tGPD. The final value is the CFAR threshold, the objective of the algorithm, based on the background distribution. The CFAR threshold is determined using the fitted tGPD inverse CDF and the desired false alarm rate.

5.3.1 Algorithm Overview.

A flow chart of the algorithm, called Truncated Extreme Value Anomaly Thresholding (TEVAT), is represented in Figure 5.4. TEVAT begins with output of

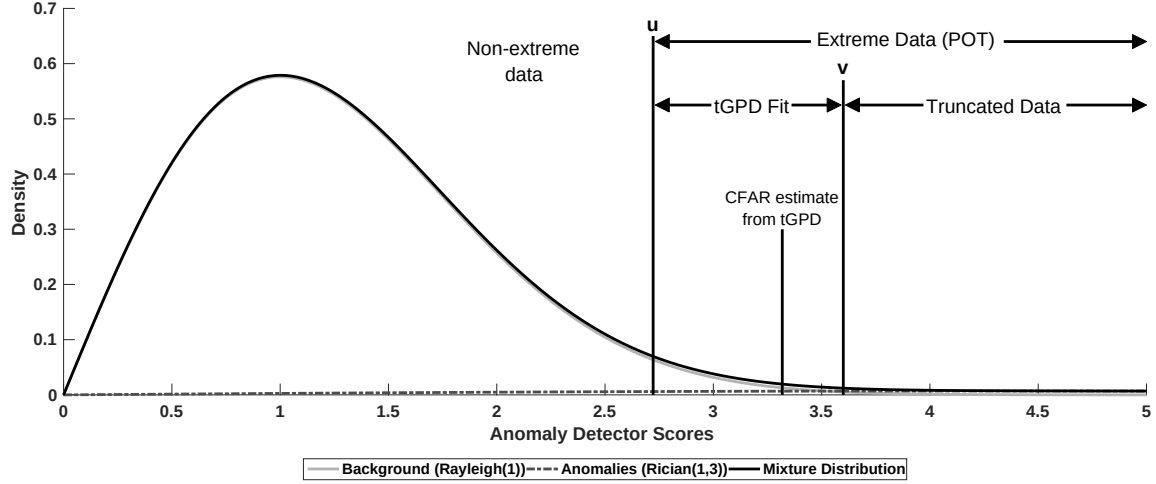


Figure 5.3: Example Relative Placement of u , v , and CFAR estimates.

anomaly scores from a detector (Fig. 5.4: AD Output) after which the selection of a threshold u provides the subset of extreme values (Fig. 5.4: Subset POT) to allow the application of EVT. This provides the POT for in which to find extreme behavior while excluding the influence of likely anomalies. The value of the threshold depends on the application and the assumed proportion of anomalies. A pre-determined empirical percentile is used to select the value of u which will be stated before each application.

After selecting a value for u , the next step in the method is to fit a GPD to the POT (Fig. 5.4: Fit GPD). There are a number of methods for fitting the GPD to the extremes [59]. This research uses the Maximum GoF (MGF) techniques as the range of GPD parameters may be outside the range suitable for Maximum Likelihood Estimation [57] when anomalies are present in the sample.

After the estimation of the GPD parameters, similar to the work of Broadwater *et al.* [19], a GoF test serves to determine the presence of anomalies in the scores (Fig. 5.4:

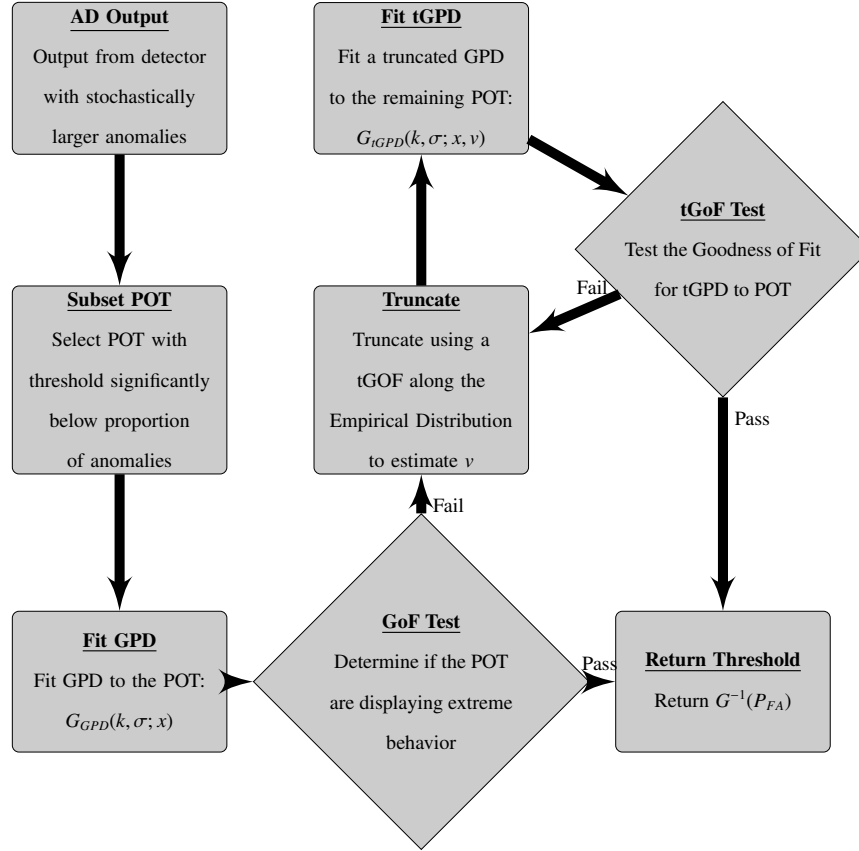


Figure 5.4: Flowchart depiction of Thresholding Algorithm.

GoF Test). Absence the influence of anomalies, EVT compels the extremes of the anomaly scores to converge to the GPD. Thus, when the GoF is determined to be poor, the assumption is this stems from the anomalies skewing the GPD.

If a poor GoF test is indicated, the introduction of a truncation point ν could mitigate the impact of the anomalies on the parameter estimate of the GPD. If ν is set too low, the parameters of the GPD can become highly variable. Our proposed method uses a truncated GoF test statistic calculated as the extreme points are eliminated, point by point, moving from the most extreme point, to estimate the point of departure from expected extreme behavior (Fig. 5.4: Truncate). For some test statistic t_ν from Table 5.2, the value

for v is selected such that the value of t_v is below the critical value for the test statistic. Data points greater than the value of v are subsequently eliminated from the estimation of the tGPD. In order to prevent the exclusion of background scores with large truncated GoF (tGoF) statistics due to the biasing of GPD parameters from the anomalies, a limited proportion of scores can be truncated at each truncate/fit/test iteration (i.e. rejecting up to 10% of the data).

After a new value for the truncation point v is set, the remaining extreme scores are fit to a tGPD (Fig. 5.4: Fit tGPD). After the parameters of the tGPD are estimated, the quality of this fit is determined through a tGoF test (Fig. 5.4: tGoF Test). If the tGoF test passes for the critical value, the tGPD is determined to be a representation of the background distribution and the CFAR threshold is determined using the inverse CDF of the tGPD or GPD depending if truncation has occurred (Fig. 5.4: Return Threshold). If the tGoF test fails, additional truncate/fit/test iterations are performed until the fit passes, indicated by the cycle in Figure 5.4.

This work builds on the ATEEVt thresholding approach, but differs in three critical aspects. The first is the use of the truncated Generalized Pareto distribution (tGPD) to fit data that is known to be right truncated. This allows for estimation of the proportion of background extremes that are above the truncation point. The second improvement is the use of iterations to converge to the best fit tGPD to the extreme data. This compensates for bias in the parameter estimates for the GPD when anomalies are present in the data. The third is the use of a more powerful GoF test for the presence of anomalies in the extremes. The power of a number of GoF tests are examined below.

A possible modification to the algorithm modifies the initial distribution in Fig. 5.4: Fit GPD. In applications where there is an absence of anomalies, this initial GPD is untruncated. However, if the presence of anomalies is highly likely in the output, it is possible to screen the largest scores from the initial fit of the GPD parameters and fit a

tGPD instead. This can serve to reduce the number of truncate/fit/test iterations later in the algorithm as the initial screening prevents the biasing of the initial fit. However, too low of an initial truncation can screen a large portion of the background extremes also biasing the fit of the tGPD. This is mitigated later in the algorithm with the use of the tGPD instead of a GPD which allows the estimation the probability of background scores being above the initial threshold, subsequent thresholds can be higher and thus include the previous screened background data.

Below is an investigation of the power of a number GoF tests to detect the presence anomalies in simulated detector output. The most powerful GoF test in this respect would be able to detect anomalies in the extremes in lower proportions and with a lower Signal-to-Noise Ratio (SNR) (as defined below).

5.3.2 *Power of GOF Tests for Detecting Anomalies.*

The determination that anomalies are present in a sample is one of the most critical steps in the process. As such, we desire a GoF test with the greatest power to detect the anomalies in the extremes of the scores. To this end, results are presented from a simulation of detector output which allows for the varying of the proportion and SNR of the anomalies present. This test is the Swerling type 0 experiment of [19] in which the background data, (H_0) , is a Rayleigh distribution with pdf:

$$f(x|H_0) = x \exp \left\{ -\frac{x^2}{2} \right\}. \quad (5.9)$$

The anomalies, (H_1) , are from a Rician distribution with a selected constant amplitude A with a pdf given by

$$f(x|H_1) = x \exp \left\{ -\frac{x^2 + A^2}{2} \right\} I_0(xA) \quad (5.10)$$

where I_0 is the Modified Bessel function of zeroth order. Under this scenario, we have a controllable Signal-to-noise ratio (SNR) for the anomalies present in the background. This SNR is a function of A : $SNR = A^2/2$. As the value of A is increased, the detectability of

the anomalies increases as well, shown by the proportionality of the SNR. At significant SNR levels, the anomalies are stochastically separated from the background such that their detection is straight-forward by all tests. The POT were selected as the top 5% of the 10,000 point sample, and the proportion of anomalies and the SNR were varied. In this experiment, at each SNR level, we want to know the statistical power of the GoF tests $(1 - \beta)$ from Table 5.1 to detect departure from the GPD. Thus, for different proportions of anomalies, 0.005, 0.01, and 0.02, power curves for the tests at P_{FA} of 0.05 are found in Figure 5.5. The results show that the A_{up}^2 is the most powerful GoF tests for detecting the presence of anomalies. This is expected since this test applies more weight to the right-tail of the GPD, which is where a majority of anomalies reside. The additional weighting of the tail would drive the test statistic to depart from nominal with less deviation from fitted GPD than the tests that place less weighting on the right-tail. Note that the KS test (D) is the least powerful of the GoF tests at the detection of anomalies in this simulation at most SNR levels.

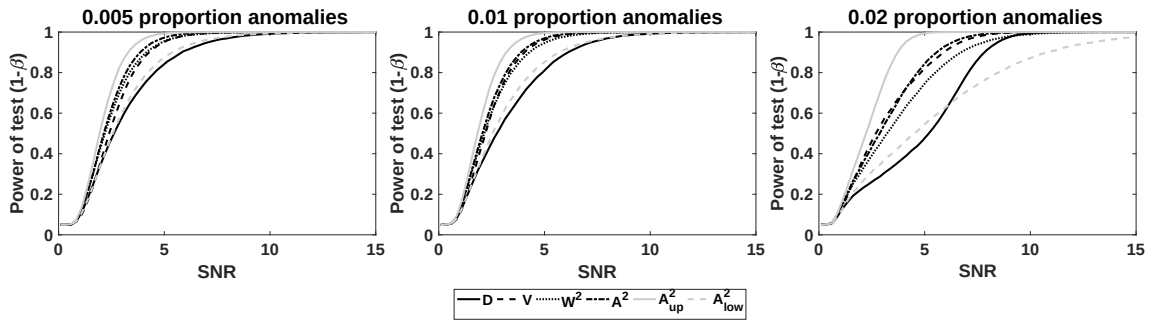


Figure 5.5: Power of GoF tests on simulated Swerling type 0 data.

5.4 Results

This section presents the results of the TEVAT thresholding method with two examples. The first example uses simulated univariate distributions for the background

and the anomalies. This allows for the deliberate control of the characteristics of the anomalies relative to the background in order to gain an understanding of the performance characteristics of TEVAT compared to other thresholding methods. Subsequently, results from a common anomaly detector for Hyperspectral Images which is built to return CFAR performance provide further evidence for the value of the TEVAT algorithm.

5.4.1 Simulated Output.

The main purpose of this experiment is to compare the characteristics of a number of thresholding techniques. To this end, the thresholding techniques of empirical percentile, zero-bin histogram [24], ATEEVT [19], and TEVAT are provided. As the desired impact of implementing the tGPD is to return a more accurate estimate of the CFAR threshold, a comparison is shown for each thresholding approach to the known CFAR threshold in terms of mean square error (MSE). In this experiment, our aim is to estimate the CFAR threshold for a $P_{FA} = 0.01$. For the empirical percentile method, the 1% percentile provides the threshold. The zero-bin histogram method is implemented with a bin width of 0.05. As the zero-bin histogram is not designed for CFAR thresholding, this bin width provides the most comparable performance in the range of SNR tested. Both the ATEEVT and TEVAT use the top 10% of data as the POT and an $\alpha \approx 0.01$ for the GoF test. Using the same experimental setup used above with a Swerling type 0 background, 10,000 replications are conducted at each SNR point to estimate the MSE of the threshold estimate. The results of these experiments can be seen in Figure 5.6, for proportion of anomalies simulated at 0.005, 0.01, and 0.02.

Examining the empirical percentile method, it performs very well, as expected, with low proportion of anomalies. However, the bias caused by the presence of anomalies becomes pronounced as the proportion increases. Yet, as SNR increases and causes a large amount of stochastic separation in our distributions, the MSE of the estimate begins to level off as the bias stabilizes. The high bias overcomes the relatively low variance of the

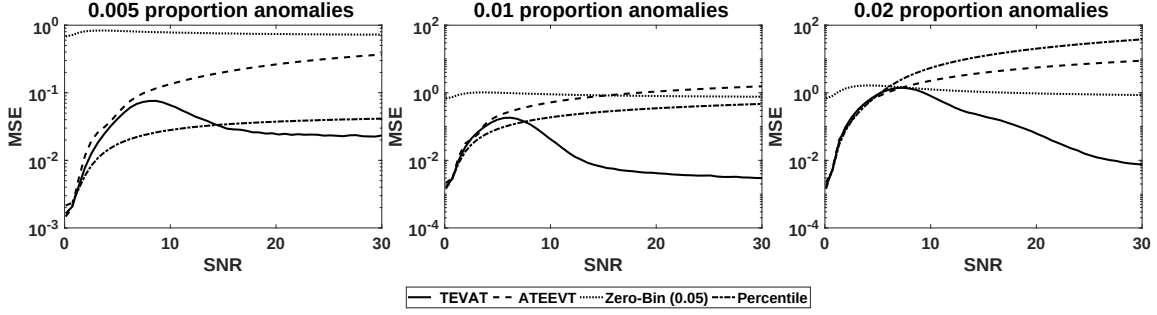


Figure 5.6: MSE of CFAR threshold estimate on simulated Swerling type 0.

method to increase the MSE, especially in the 0.02 proportion case. As such, the empirical percentile method should only be applied when very few anomalies are suspected in the data.

Next, we see that the zero-bin histogram returns high MSE across all the tested proportions of anomalies. As this is a distance based metric not necessarily estimating a CFAR threshold, it is the most dependent on each data sample and returns high variance. This is evident in the high MSE across all ranges. Additionally, as the method essentially relies on the distance between data points, at low SNRs the method tends to also display high bias.

Examining the behavior of the ATEEV method, we find it is limited by its inability to estimate the truncation probability and perform additional iterations to screen anomalies. Recall, the method of ATEEV uses the KS test and a GPD to estimate the threshold. the GPD estimation can be affected when significant overlap exists between the anomalies and the background distributions as the KS test allows for the screening of anomalies from affecting the threshold estimate. As SNR increases, the estimated GPD begins to be biased by the presence of the anomalies until, at moderate SNR levels, the KS test is able to detect the presence of anomalies. However, at moderate SNRs no stochastic separation between the background and anomaly distributions and the estimation of the

GPD is influenced by the truncation of background data. At higher SNRs, ATEEV is unable to exclude all of the anomalies based on the initial GPD fit and KS-test, thus the bias of the estimate continues to increase with SNR. These effects together result in poor estimation of the CFAR threshold at high SNR.

Lastly, we see the improvements provided by TEVAT display the most consistent behavior across the range of proportion of anomalies and SNR levels. The tGPD provides the ability to estimate the proportion of background data that should lie above the truncation point and the more powerful GoF test is more capable of detecting anomalies. Additionally, the iterative approach allows for the exclusion of anomalies when little to no stochastic separation exists. Overall, TEVAT displays consistent estimation of the desired CFAR threshold for the simulated data. The bias of the CFAR threshold by the presence of high SNR anomalies in the dataset is better controlled compared to the other methods. This can be seen by the decrease and stabilization of the MSE for TEVAT across each proportion level as SNR increases.

5.4.2 *Detector Output from Hyperspectral Images.*

This section demonstrates the ability of the TEVAT method on anomaly detector output in the field of Hyperspectral Imaging (HSI). Results are from one of the most common anomaly detectors in HSI, known as the RX [81]. RX utilizes a moving window approach in which a current *pixel under test* (PUT) is compared to the pixels in a spatial neighborhood around it. The neighboring pixels provide estimates for the background mean vector, μ , and covariance matrix, Σ . Together with the number of pixels in the window, N , each spectral vector, x , is assigned an RX score by

$$\text{RX}(x) = (x - \mu)^T \left[\left(\frac{N}{N+1} \right) \Sigma + \left(\frac{1}{N+1} \right) (x - \mu)(x - \mu)^T \right]^{-1} (x - \mu) \quad (5.11)$$

Note that as the size of the window grows, (5.11) converges to the Mahalanobis distance of the pixel's spectral vector to the mean vector. Therefore, for adequately sized windows, the RX score is compared to a $\chi^2_{P_{FA}, (N-1)}$ critical value for a desired false alarm rate, P_{FA} .

When all assumptions for RX are met, this critical value should present a CFAR threshold. However, RX relies on the estimation of the covariance matrix the Gaussian distributed background signatures. This estimation requires independence of observations which is often violated due to the spatial correlation of neighboring pixels. One approach to mitigate this effect is to use moving linear windows instead of the rectangle of traditional RX [104]. This adaptation, known as Linear RX, drives up the average distance between the PUT and the pixels inside the window, thus presumably reducing the spatially correlation effects. However, the effects of spatial correlation still remain and can cause RX scores to deviate from the theoretical distribution. Using RX scores from six hyperspectral images aims to show the value of TEVAT applied to real-world data. Each image is from the Hyperspectral Digital Imagery Collection Equipment (HYDICE) sensor [82] acquired in the Forest Radiance I and Desert Radiance II collection events. The anomaly thresholds are determined separately on each image using the parametric $\chi^2_{P_{FA},(N-1)}$ critical value, the empirical percentile, the zero-bin, ATEEV, and TEVAT methods. As above, the empirical percentile method uses the top P_{FA} percent of data to determine the anomaly threshold, while the zero-bin histogram method is implemented with a bin width of 0.15. This bin width is selected for its performance at $P_{FA} = 0.01$, as there is no fixed *a priori* relationship between FAR and bin width. As the anomaly and background condition of each pixel is known, the actual threshold for a false alarm rate is used to calculate the MSE of each method's estimate for the range of false alarm rates. The results of this experiment are shown in Figure 5.7. As is easily seen, our proposed method returns the most consistent estimate across all desired false alarm rates. Of particular note, TEVAT shows a decrease in MSE method by over an order of magnitude compared to the parametric estimator.

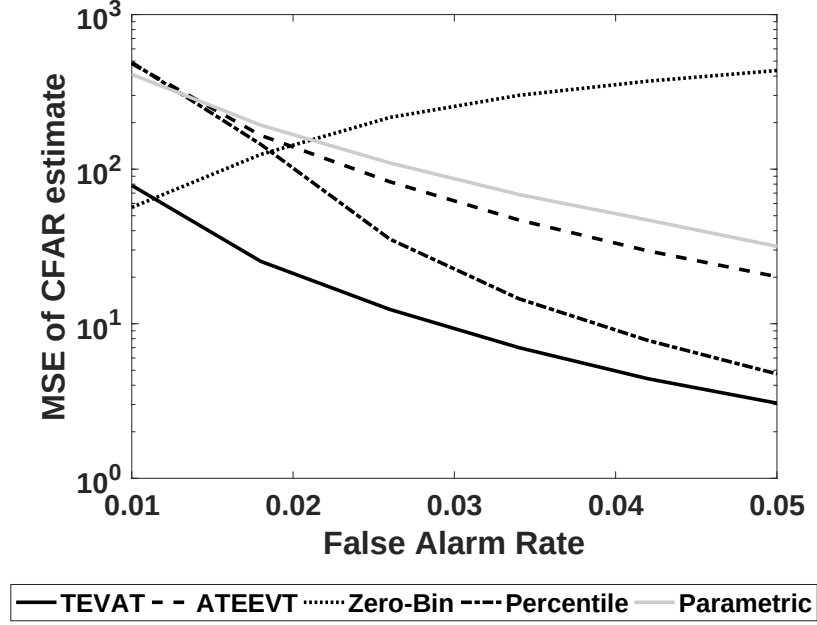


Figure 5.7: MSE of threshold estimate on RX scores from HYDICE images.

5.5 Conclusions

This chapter presents an improved method for the estimation of a CFAR threshold for anomaly detectors either lacking the CFAR ability, or in violation of required assumptions. The method applies to a rather general class of background distributions as it relies only on the extremes of the anomaly scores through application of EVT. It detects anomalies based on the GoF of a tGPD to a truncated portion of the extreme values of an anomaly detector's output scores. Results show that it better estimates, in terms of MSE, a CFAR threshold on simulated data, as well as output from a common class of anomaly detectors for hyperspectral images.

VI. Extreme Value Thresholding as Mitigation to Correlation Effects in Anomaly Detection

6.1 Introduction

As discussed in section 2.4, the common benchmark algorithm for anomaly detection in Hyperspectral Imaging (HSI) is that of Reed–Xiaoli (RX). Built as a Constant False Alarm Rate (CFAR) detector, RX relies on the assumption of Gaussian noise in the background to form the RX score of each pixel. With a window of size N used to estimate of the background mean vector, μ , and covariance matrix, Σ , the RX Score of a spectral vector, $\mathbf{x}$ is calculated as

$$\text{RX}(x) = (x - \mu)^T \left[\left(\frac{N}{N+1} \right) \Sigma + \left(\frac{1}{N+1} \right) (x - \mu)(x - \mu)^T \right]^{-1} (x - \mu) \quad (6.1)$$

As the size of the windows grows, this RX score converges to a χ^2 distribution when the background is homogeneous and stationary. Therefore, it is expected that the algorithm is CFAR capable with a threshold from a χ^2 distribution for some probability of false alarm, P_{FA} . But investigation of the results in Chapter 3 tells a different story about the CFAR capability of RX. The False Positive Fraction (FPF) performance of the RX-type detectors exhibits greater variance than the other algorithms presented and it would seem that the CFAR characteristics of the detectors is deficient. For example, on the sample of ten images in Chapter 3, the range on the FPF is from 0.016 – 0.061 for a desired FPF of 0.05. There are a number of factors that may prohibit this convergence such as large spatial neighborhoods of anomalies and non-homogeneous, non-stationary behavior in the background. Numerous adaptations to the general algorithm of RX have been proposed as discussed in section 2.4.3. The most pertinent to this chapter are that of Linear RX (LRX) and Iterative Linear RX (ILRX) as they are used as base algorithms for embellishments later in the chapter.

Chapter 5 discussed the development of an Extreme Value Theory (EVT) based anomaly thresholding technique with CFAR properties. This technique, called Truncated Extreme Value Anomaly Thresholding (TEVAT), is based on the extreme values of the anomaly detector output and therefore is applicable to a general set of parent distributions [28]. This chapter shows that RX-type detectors using extreme value theory based thresholding techniques can improve the False Alarm Rate (FAR) stability of the detectors.

6.2 Algorithms

As previously described above in Section 2.4.3, the RX-family of anomaly detectors use the generalized likelihood ratio test to identify anomalies. Numerous adaptations to the general algorithm of RX have been proposed. When the assumption of multivariate normality is met for the spectral signature, this results in the RX scores from eq. 5.11 being disturbed according to the χ^2 distribution. However, since departures from multivariate normality is highly likely in HSI, these detectors can often return inconsistent false positive rates [104]. Further difficulties are caused by the effects of spatial correlation and non-homogeneity in the estimation of the background statistics. In applications where the FAR must be controlled, this inconsistency can render the algorithm unsuitable for deployment. While some of the RX-type detectors attempt to mitigate this effect through window shape [104], iterative covariance estimation [96], and the use of mixture models for the distribution [33]; the underlying problem remains.

In addition to applying TEVAT to regain CFAR capability, a slightly modified version of the ILRX algorithm is developed. This is done in the interest of reducing the computational complexity of the iterative algorithm. The iterative nature of ILRX can cause the algorithm be much more computationally complex when compared to other RX algorithms. For example, on many images the ILRX can conduct upwards of 30 iterations of LRX before convergence. This results in a significant increase in computation time.

Thus, a slight modification to the ILRX framework is proposed named High-Low FAR LRX (HLRX). HLRX is similar to the approach of ILRX but is limited to two iterations. During the first iteration, a high P_{FA} threshold is used to screen potential anomalies from the subsequent background statistic estimation. Once this iteration is complete, the second iteration is conducted with the desired P_{FA} threshold to declare anomalies. This approach fulfills the spirit of the ILRX approach without the immense computational burden of conducting upwards of 30 iterations of LRX. Additionally, as HLRX controls the number of iterations that are conducted, it also has more control over the compounding of type-I errors.

6.2.1 *Extreme (Ext) RX.*

The Extreme Threshold, or Ext-type, RX detectors are a set of new embellished RX detectors. These detectors use TEVAT within the context of an RX-type detector replacing the RX detectors' typical $\chi^2_{P_{FA}, N-1}$ threshold. This typical threshold is highly susceptible to the departure of the empirical distribution of the data from the χ^2 distribution. Using TEVAT can overcome this difficulty caused by the effects of spatial correlation and non-homogeneity on the estimation of the background statistics. Recall, TEVAT uses the empirical extremes of the anomaly detection scores to fit a Generalized Pareto Distribution (GPD), truncates possible influential outliers, and returns an estimate for the CFAR threshold for a given P_{FA} . By using TEVAT, four different embellishments on RX-type algorithms are created: Extreme RX (ExtRX), Extreme LRX (ExtLRX), Extreme ILRX (ExtILRX), and Extreme HLRX (ExtHLRX). Note that TEVAT can be directly applied on any anomaly detector with a single anomaly score output.

The approach to all four Ext-type detectors is essentially the same. The baseline algorithm determines the anomaly scores of each pixel and use TEVAT as the means of determining the threshold, replacing the χ^2 critical value as the anomaly threshold. This

can return CFAR performance to the algorithms. This is demonstrated in the results section of this chapter.

In the iterative algorithms TEVAT is used for each iteration. For ExtHLRX, TEVAT thresholding is used for both the high and low iterations. The high P_{FA} value is set as twice the desired P_{FA} .

An example of the effect that correlation in an image can have on the RX scores is presented as motivation for these changes. As can be seen in Figure 6.1, the RX scores display a generally χ^2 shape but the parametric distribution, severely underestimates the very low and very high scores. This is the effect of correlation on the estimation of the local covariance matrix. Conversely, the probability density function (pdf) built from the distribution fit with TEVAT returns a close approximation of the background extremes. Note that this pdf only starts at the 85% percentile of the data since that value is used to determine the *points over threshold* (POT).

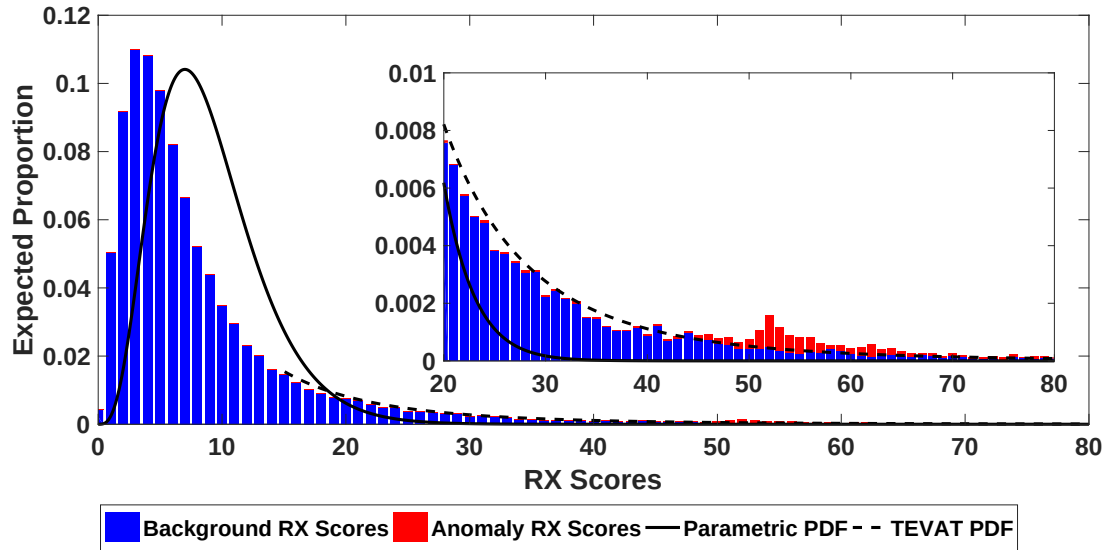


Figure 6.1: Sample of RX scores from LRX on ARES5F.

6.3 Methodology

To show the CFAR capabilities improvement of the Ext-type detectors, each anomaly detector's performance is compared on the same collection images from the Forest Radiance I and Desert Radiance II collection events from the Hyperspectral Digital Imagery Collection Equipment (HYDICE) sensor that is used in Chapter 4 [82]. The sensor is a push-broom type sensor operated from an aircraft returning 210 spectral bands between 397 nm and 2,500 nm. Atmospheric absorption bands are selected using the process of Paciencia [74] which eliminates spectral bands with low loadings in a Factor Analysis model. These removed bands are listed by image in Table 6.1.

Table 6.1: Spectral bands removed due to indication of atmospheric absorption.

Image	Bands Removed	Number Bands Retained
ARES1D	104:110; 138:153; 157:158; 187:210	161
ARES1F	10:11; 103:109; 138:152; 156:157; 180:210	153
ARES2D	2:11; 104:108; 139:152; 201; 204; 206; 208:210	175
ARES2F	1:8; 104:108; 139:151; 204; 206:210	178
ARES3D	2:9; 104:108; 139:152; 208:210	180
ARES3F	103:107; 139:151; 204:210	185
ARES4	104:109; 139:152; 157; 192:210	170
ARES4F	1; 103:108; 138:151; 200; 202:210	179
ARES5	10:11; 103:108; 138:152; 195:210	171
ARES5F	2:4; 104:108; 139:151; 204; 206:210	183

As this chapter is focused on using algorithms previously shown to have desirable performance capabilities [104], the performance of the algorithms is shown without a training and validation split of images. As with any empirical display of performance on a small set of data, this approach is limited in the demonstration of the power of the anomaly detector. However, as this method is only meant to return CFAR capability detectors which are assumed to have desirable performance attributes already, the limited

amount of sample images is not a significant detriment to the approach. Below is a comparison of the four embellishments to their baseline methods.

6.4 Results

Thresholding of the anomaly scores of a detector contributes only a small part to the overall performance of the algorithm. However, as Minkler and Minkler described, the fundamental problem in automatic detection is selecting a detection strategy that provides some “optimal” balance between the True Positive Fraction (TPF) and FPF [70]. For the same set of anomaly scores, this balance results in trade-offs between the TPF and the FPF. For pairs of base and Ext-type embellishments, a lower FPF will result in a lower TPF and, vice versa, higher FPF results in higher TPF since the scores being thresholded are the same. The consistency and ability to provide the optimal balance of a CFAR algorithm across a number of test images is of greater concern. While the increased TPF may seem desirable, the inconsistency of the results may cause difficulty for subsequent analysis efforts.

Note that while the set of images tested is small, the results still present compelling and succinct evidence to the increased CFAR characteristics of the Ext-type algorithms over the base RX detectors. Of particular note is the drastic reduction in the bias and variance of the observed FPF across the images of the Ext-type detectors over the baseline algorithms.

Examining ILRX, there is evidence of a strong bias towards high FAR from repeated testing of the scores for a constant threshold. The iterative nature of the algorithm results in compounding of the FAR across all iterations. Direct control of the FAR during iterations when using the χ^2 threshold would require assumptions on the proportion of the image that is anomalous before running the algorithm. However, as TEVAT uses the

Table 6.2: FPF Performance on Images at Desired FAR of 1%.

		RX	LRX	ILRX	HLRX
Base	Mean FPF	0.016	0.050	0.188	0.104
	Var FPF	4.9E-05	3.3E-04	6.8E-03	1.6E-03
Ext-type	Mean FPF	0.015	0.011	0.016	0.019
	Var FPF	4.2E-04	2.5E-04	2.9E-04	4.2E-04

Table 6.3: TPF Performance on Images at Desired FAR of 1%.

		RX	LRX	ILRX	HLRX
Base	Mean TPF	0.523	0.779	0.989	0.936
	Var TPF	0.070	0.025	0.000	0.008
Ext-type	Mean TPF	0.538	0.690	0.838	0.809
	Var TPF	0.067	0.044	0.066	0.061

collection of anomaly scores from the entire image, the risk of compounding type-I error is reduced.

Table 6.4: FPF Performance on Images at Desired FAR of 5%.

		RX	LRX	ILRX	HLRX
Base	Mean FPF	0.035	0.087	0.334	0.163
	Var FPF	1.4E-04	5.8E-04	6.0E-03	1.9E-03
Ext-type	Mean FPF	0.050	0.045	0.049	0.051
	Var FPF	2.2E-04	1.4E-04	2.3E-04	4.1E-04

Table 6.5: TPF Performance on Images at Desired FAR of 5%.

		RX	LRX	ILRX	HLRX
Base	Mean TPF	0.523	0.779	0.989	0.936
	Var TPF	0.070	0.025	0.000	0.008
Ext-type	Mean TPF	0.538	0.690	0.838	0.809
	Var TPF	0.067	0.044	0.066	0.061

At all levels for desired FAR, the Ext-type detectors return FPF performance far closer to desired than the base detectors. This is the desired performance effect of

applying the TEVAT algorithm. Examining the RX algorithm more closely at the desired FAR of 1%, the base algorithm actually nearly matched the performance of ExtRX. However, this performance is lost at the higher levels of desired FAR where the RX algorithm begins to produce a lower FPF at a detriment to TPF performance.

Table 6.6: FPF Performance on Images at Desired FAR of 10%.

		RX	LRX	ILRX	HLRX
Base	Mean FPF	0.054	0.118	0.489	0.228
	Var FPF	2.4E-04	7.6E-04	6.1E-03	1.6E-03
Ext-type	Mean FPF	0.122	0.128	0.123	0.125
	Var FPF	3.1E-04	7.0E-05	1.0E-04	1.1E-04

Table 6.7: TPF Performance on Images at Desired FAR of 10%.

		RX	LRX	ILRX	HLRX
Base	Mean TPF	0.560	0.820	0.994	0.964
	Var TPF	0.068	0.020	0.000	0.003
Ext-type	Mean TPF	0.630	0.832	0.905	0.910
	Var TPF	0.055	0.021	0.028	0.023

The benefit of achieving CFAR for the Ext-type algorithms is highly pronounced for all detectors, especially the iterative ones. This stems from the compounding of the type-I error using the parametric threshold. During each iteration, a background pixel has a chance to be declared anomalous at the P_{FA} and after 30 iterations this begins to compound. Combined with the biasing of the covariance matrix estimate, the algorithm begins to include a large amount of background pixels in the declared anomalies during later iterations. When using TEVAT this is minimized since the threshold is determined based on the entirety of the scores resulting from the image providing a more accurate estimation of the threshold. The high desired FAR the iterative algorithms can miss the mark on the FPF upwards of 20-40%, as seen in the FPF Tables.

Another key aspect to CFAR performance is the consistency of the FAR. The variance of the 10 image FPF can be seen in Table 6.8 at a desired FAR of 0.05. For each detector, the Ext-type results in a reduction in the observed variance with statistical significance in the reduction for LRX, ILRX, and HLRX when tested with an F-test.

Table 6.8: Variance of FPF at Desired FAR of 10% (Bolded value represent statistical significance at a p-value of 0.05).

		RX	LRX	ILRX	HLRX
σ^2	Ext-type	9.34E-05	6.06E-05	2.35E-05	1.64E-04
	Base	1.58E-04	5.13E-04	1.73E-04	1.08E-03
$\sigma_{Ext}^2/\sigma_{Base}^2$		0.59	0.12	0.14	0.15

6.5 Conclusions

This chapter presented new embellishments on several RX-type detectors using Extreme Value Thresholding. In particular, the use of the TEVAT algorithm to replace the parametric anomaly thresholding technique traditionally used in RX-type detectors is shown to return the CFAR ability of the algorithms. This new family of Ext-type algorithms retain most of the desirable performance characteristics of the base algorithms, but provide increased consistency in the FPF across the images tested. Specifically, additional performance characteristics of the novel algorithm HLRX, as originally presented in Chapter 3, was displayed.

VII. Conclusion

The research presented in this dissertation focused on the techniques and approaches taken to develop quality anomaly detection algorithms. In doing so, the objective of the research is to aid in the development of anomaly detectors with limited data, select parameters that are robust possible anomalies, attain desirable false alarm rate characteristics, and account for spatial correlation problems. The subsequent chapters presented specifics on the new methods.

7.1 Original Contributions

Chapter 3 introduced a method for the sensitivity analysis of multivariate anomaly detection algorithms specifically anomaly detection in Hyperspectral Imaging (HSI), known as Mahalanobis Masking. This method is based on incrementally decreasing the Mahalanobis Distance of the anomalies from the background. In doing so, the anomalies become more difficult to detect and the performance of the detector drops. However, when we witness drastic drops in performance during masking there may be stability issues with the algorithm. The chapter also presented quantitative measures of ability, stability, and overall performance for the direct comparison of algorithms.

Chapter 4 extended previous work on the definition of a noise framework for Robust Parameter Design (RPD) of HSI. This framework uses Mahalanobis Masking to increase the variation in anomalies that are tested. As a major challenge in the development and testing of anomaly detection algorithms is the limited availability of labeled datasets, this is a valuable approach. Since the level of masking is controllable, we are able to vary the level of difficulty embedded into the RPD process. Additionally, this work extends the approach of Bellucci [13] to a larger number of control and noise factors than previously shown.

Chapter 5 presents an improved approach towards the estimation of a Constant False Alarm Rate (CFAR) threshold for anomaly detection. This approach builds on previous work using Extreme Value Theory (EVT) to characterize the extreme values of the anomaly scores. The improvements of the method are based on the application of more powerful GoF test for the indication of anomalies present in the extreme values, an iterative approach to counteract the bias of the distribution parameters towards anomalies, and the use of a truncated distribution to account for the proportion of background data that may be excluded during truncation. Overall, this method presents threshold estimation on both simulated and real-world data with better properties than other approaches.

Chapter 6 developed a new family of detectors derived from the RX detection approach. These detectors, named the Ext-type detectors, replace the parametric threshold of the RX detectors with a threshold estimated using the Truncated Extreme Value Anomaly Thresholding (TEVAT) algorithm. In doing so, they return a degree of CFAR capability to the algorithms that is lost in application due to spatial correlation effects. Additionally, an adaptation of the Iterative Reed–Xiaoli (RX) (IRX) algorithm, known as High-Low False Alarm Rate (FAR) Linear RX (LRX) (HLRX), which produces comparable performance results at much lower computational cost is examined.

7.2 Suggested Future Work

- Extend the Mahalanobis Masking approach to other multivariate anomaly detection applications such as Computer Network Intrusion Detection.
- Determination of sources of variability in the component based methods during Mahalanobis Masking.
- Integration of TEVAT with estimation methods for the POT.

Appendix A: Derivation of Formulas

A.1 Derivation of Mahalanobis Masking Iterative Formula

In this section we develop the vector updating formula for Mahalanobis Masking:

$$\mathbf{x}_{i+1} = \mathbf{x}_i - \alpha \Sigma^{-1}(\mathbf{x}_i - \mu) \quad (\text{A.1})$$

We start by defining the distance cost function of the initial vector, $\mathbf{x}_0$. For Mahalanobis Masking, we use the Mahalanobis distance calculated using the mean vector, μ , and covariance matrix, Σ , of the background data.

$$D^M(\mathbf{x}_i, \mu, \Sigma) = \sqrt{(\mathbf{x}_i - \mu)' \Sigma^{-1} (\mathbf{x}_i - \mu)} \quad (\text{A.2})$$

With this cost function, we set up Mahalanobis Masking as a gradient descent using the cost function gradient with a step size α :

$$\mathbf{x}_{i+1} = \mathbf{x}_i - \alpha \nabla D^M(\mathbf{x}_i, \mu, \Sigma) \quad (\text{A.3})$$

The cost gradient of the distance function is, using the symmetry of Σ

$$\alpha \nabla D^M(\mathbf{x}_i, \mu, \Sigma) = \alpha \frac{\partial D^M(\mathbf{x}_i, \mu, \Sigma)}{\partial x} \quad (\text{A.4})$$

$$= \frac{1}{2} \alpha \left((\mathbf{x}_i - \mu)' \Sigma^{-1} (\mathbf{x}_i - \mu) \right)^{-1/2} (2 \Sigma^{-1} \mathbf{x}_i - 2 \Sigma^{-1} \mu) \quad (\text{A.5})$$

$$= \alpha \Sigma^{-1} (\mathbf{x}_i - \mu) \quad (\text{A.6})$$

where in all scalars in (A.6) are subsumed in the step size, α .

A.2 Derivation of Right-truncated Goodness-of-Fit Statistics

Below we derive the goodness-of-fit test statistics for various GoF test in the presence of right-truncated data. Accounting for truncation in the testing of fit is critical since the empirical distribution function (EDF) already displays the effects of truncation. Therefore, the corresponding theoretical cumulative distribution function must also be adjusted. Each

of the derivation follows similarly to the left-truncation test statistics derived by Cherbonai et al. [23]. In the derivations below we will use the following notation:

z_v	$\hat{F}(v)$
$x_{(j)}$	The j -th ordered statistics of set $\{X\}$
$z_{(j)} \equiv \hat{F}(x_{(j)})$	Theoretical Cumulative Distribution Function
F_n	Empirical Distribution Function

A.2.1 Derivation of Right-Truncated Kolmogorov-Smirnov and Kuiper Goodness of Fit Test statistics.

In the presence of right-truncation, the modification of the statistics D_n^+ and D_n^- is straight forward. By scaling the EDF by the truncated probability, the statistics become

$$D_v^+ = \frac{\sqrt{n}}{z_v} \left(\sup_j \left[z_{(j)} - \frac{j-1}{n}(z_v) \right] \right)$$

and

$$D_v^- = \frac{\sqrt{n}}{z_v} \left(\sup_j \left[\frac{j}{n}z_v - z_{(j)} \right] \right).$$

The right-truncation does not change the remaining form of the two statistics. Thus, the KS statistics becomes $D_v = \max[D_v^+, D_v^-]$ and the Kuiper statistic remains

$V_v = \max[D_v^+, D_v^-]$ This is the general approach in the derivation of the remaining GoF statistics.

A.2.2 Derivation of the Right-Truncated Cramér-von Mises Goodness of Fit Test statistic.

Following the method of Chernobai et al. [23] we use the probability integral transform to change to the uniformly distributed z values for the distribution:

$$W_v^2 = n \int_{-\infty}^v (F_n(x) - \hat{F}(x))^2 d\hat{F}(x) = \frac{n}{z_v} \int_0^{z_v} \frac{(F_n(z^*)(z_v) - z)^2}{z_v^2} dz$$

From here, we can use the observations of the data set $x_j, j = 1, 2, \dots, n$ to define F_n and re-write the above as

$$\frac{z_v^3 W_v^{2*}}{n} = \int_0^{z_1} z^2 dz + \sum_{j=1}^{n-1} \int_{z_{(j)}}^{z_{(j+1)}} \left(\frac{j}{n} z_v - z \right)^2 dz + \int_{z_{(n)}}^{z_v} (z_v - z)^2 dz$$

Then, through the application of the anti-derivatives and distributing the summations:

$$\begin{aligned} \frac{z_v^3 W_v^2}{n} &= \frac{z_{(1)}^3}{3} + \sum_{j=1}^{n-1} \int_{z_{(j)}}^{z_{(j+1)}} \left(\frac{j}{n} z_v \right)^2 - 2 \frac{j}{n} z_v z + z^2 dz + \int_{z_{(n)}}^{z_v} (z_v^2 - 2 z_v z + z^2) dz \\ \frac{z_v^3 W_v^2}{n} &= \frac{z_{(1)}^3}{3} + \sum_{j=1}^{n-1} \left[\left(\frac{j}{n} z_v \right)^2 z - \frac{j}{n} z_v z^2 + \frac{z^3}{3} \right]_{z_{(j)}}^{z_{(j+1)}} + \left[z_v^2 z - z_v z^2 + \frac{z^3}{3} \right]_{z_{(n)}}^{z_v} \\ \frac{z_v^3 W_v^2}{n} &= \frac{z_{(1)}^3}{3} + \sum_{j=1}^{n-1} \left[\left(\frac{j}{n} z_v \right)^2 (z_{(j+1)} - z_{(j)}) + \frac{j}{n} z_v (z_{(j)}^2 - z_{(j+1)}^2) + \frac{z_{(j+1)}^3}{3} - \frac{z_{(j)}^3}{3} \right] + \left[z_v^2 z - z_v z^2 + \frac{z^3}{3} \right]_{z_{(n)}}^{z_v} \\ \frac{z_v^3 W_v^2}{n} &= \frac{z_{(1)}^3}{3} + \sum_{j=1}^{n-1} \left(\frac{j}{n} z_v \right)^2 (z_{(j+1)} - z_{(j)}) + \sum_{j=1}^{n-1} \frac{j}{n} z_v (z_{(j)}^2 - z_{(j+1)}^2) + \sum_{j=1}^{n-1} \frac{z_{(j+1)}^3}{3} - \sum_{j=1}^{n-1} \frac{z_{(j)}^3}{3} \\ &\quad + z_v^2 z_v - z_v z_v^2 + \frac{z_v^3}{3} - z_v^2 z_{(n)} + z_v z_{(n)}^2 - \frac{z_{(n)}^3}{3} \\ \frac{z_v^3 W_v^2}{n} &= \frac{z_{(1)}^3}{3} + \sum_{j=1}^{n-1} \frac{z_{(j+1)}^3}{3} - \sum_{j=1}^{n-1} \frac{z_{(j)}^3}{3} - \frac{z_{(n)}^3}{3} + \sum_{j=1}^{n-1} \left(\frac{j}{n} z_v \right)^2 z_{(j+1)} - \sum_{j=1}^{n-1} \left(\frac{j}{n} z_v \right)^2 z_{(j)} - z_v^2 z_{(n)} \\ &\quad + \sum_{j=1}^{n-1} \frac{j}{n} z_v z_{(j)}^2 - \sum_{j=1}^{n-1} \frac{j}{n} z_v z_{(j+1)}^2 + z_v z_{(n)}^2 + \frac{z_v^3}{3} \\ \frac{z_v^3 W_v^2}{n} &= \sum_{j=1}^n \left(\frac{j-1}{n} z_v \right)^2 z_{(j)} - \sum_{j=1}^n \left(\frac{j}{n} z_v \right)^2 z_{(j)} + \sum_{j=1}^n \frac{j}{n} z_v z_{(j)}^2 - \sum_{j=1}^n \frac{j-1}{n} z_v z_{(j)}^2 + \frac{z_v^3}{3} \\ \frac{z_v^3 W_v^2}{n} &= \sum_{j=1}^n \frac{j^2 - 2j + 1}{n^2} z_v^2 z_{(j)} - \sum_{j=1}^n \frac{j^2}{n^2} z_v^2 z_{(j)} + \sum_{j=1}^n \frac{j}{n} z_v z_{(j)}^2 - \sum_{j=1}^n \frac{j-1}{n} z_v z_{(j)}^2 + \frac{z_v^3}{3} \\ \frac{z_v^3 W_v^2}{n} &= \sum_{j=1}^n \frac{1-2j}{n^2} z_v^2 z_{(j)} + \sum_{j=1}^n \frac{1}{n} z_v z_{(j)}^2 + \frac{z_v^3}{3} \\ \frac{z_v^3 W_v^2}{n} &= \sum_{j=1}^n \frac{1-2j}{n^2} z_v^2 z_{(j)} + \sum_{j=1}^n \frac{1}{n} z_v z_{(j)}^2 + \frac{z_v^3}{3} \\ W_v^2 &= \frac{1}{n z_v} \sum_{j=1}^n (1-2j) z_{(j)} + \frac{1}{z_v^2} \sum_{j=1}^n z_{(j)}^2 + \frac{n}{3} \\ W_v^2 &= \sum_{j=1}^n \frac{(1-2j) z_{(j)}}{n z_v} + \sum_{j=1}^n \frac{z_{(j)}^2}{z_v^2} + \frac{n}{3} \end{aligned}$$

A.2.3 Derivation of the *Right-Truncated Anderson-Darling Goodness of Fit Test statistic*.

Following the method of Chernobai et al. [23] we use the probability integral transform to change to the uniformly distributed z values for the distribution:

$$A_v^2 = n \int_{-\infty}^v \frac{(F_n(x) - \hat{F}(x))^2}{\hat{F}(x)(1 - \hat{F}(x))} d\hat{F}(x) = \frac{n}{z_v} \int_0^{z_v} \frac{(F_n(z)(z_v) - z)^2}{(z_v - z)z} dz$$

From here, we can use the observations of the data set $x_j, j = 1, 2, \dots, n$ to define F_n and re-write the above as

$$\begin{aligned} \frac{z_v A_v^2}{n} &= \int_0^{z_{(1)}} \frac{z^2}{(z_v - z)z} dz + \sum_{j=1}^{n-1} \int_{z_{(j)}}^{z_{(j+1)}} \frac{\left(\frac{j}{n} z_v - z\right)^2}{(z_v - z)z} dz + \int_{z_{(n)}}^{z_v} \frac{(z_v - z)^2}{(z_v - z)z} dz \\ \frac{z_v A_v^2}{n} &= \int_0^{z_{(1)}} \frac{z}{(z_v - z)} dz + \sum_{j=1}^{n-1} \int_{z_{(j)}}^{z_{(j+1)}} \frac{\left(\frac{j}{n} z_v\right)^2 - 2\left(\frac{j}{n} z_v\right)z + z^2}{(z_v - z)z} dz + \int_{z_{(n)}}^{z_v} \frac{(z_v - z)}{z} dz \\ \frac{z_v A_v^2}{n} &= [-z_v \log(z_v - z) - z]_0^{z_{(1)}} + \sum_{j=1}^{n-1} \left[\frac{j^2 z_v (\log(z) - \log(z_v - z))}{n^2} \right]_{z_{(j)}}^{z_{(j+1)}} \\ &\quad - \sum_{j=1}^{n-1} \left[\frac{2j z_v \log(z_v - z)}{n} \right]_{z_{(j)}}^{z_{(j+1)}} + \sum_{j=1}^{n-1} [-z_v \log(z_v - z) - z]_{z_{(j)}}^{z_{(j+1)}} + [z_v \log(z) - z]_{z_{(n)}}^{z_v} \\ \frac{z_v A_v^2}{n} &= -z_v \log(z_v - z_{(1)}) - z_{(1)} + z_v \log(z_v) \\ &\quad + \sum_{j=1}^{n-1} \frac{j^2 z_v (\log(z_{(j+1)}) - \log(z_v - z_{(j+1)}))}{n^2} - \sum_{j=1}^{n-1} \frac{j^2 z_v (\log(z_{(j)}) - \log(z_v - z_{(j)}))}{n^2} \\ &\quad - \sum_{j=1}^{n-1} \frac{2j z_v \log(z_v - z_{(j+1)})}{n} + \sum_{j=1}^{n-1} \frac{2j z_v \log(z_v - z_{(j)})}{n} + \sum_{j=1}^{n-1} [-z_v \log(z_v - z_{(j+1)}) - z_{(j+1)}] \\ &\quad - \sum_{j=1}^{n-1} [-z_v \log(z_v - z_{(j)}) - z_{(j)}] + z_v \log(z_v) - z_v - z_v \log(z_{(n)}) + z_{(n)} \\ \frac{z_v A_v^2}{n} &= -z_v + 2z_v \log(z_v) - z_v \log(z_v - z_{(1)}) - z_{(1)} + \sum_{j=1}^n \frac{(j-1)^2 z_v (\log(z_{(j)}) - \log(z_v - z_{(j)}))}{n^2} \end{aligned}$$

$$\begin{aligned}
& - \sum_{j=1}^{n-1} \left[\frac{j^2 z_v (\log(z_{(j)}) - \log(z_v - z_{(j)}))}{n^2} \right] - \sum_{j=1}^n \frac{2(j-1)z_v \log(z_v - z_{(j)})}{n} \\
& + \sum_{j=1}^{n-1} \frac{2jz_v \log(z_v - z_{(j)})}{n} + \sum_{j=2}^n [-z_v \log(z_v - z_{(j)}) - z_{(j)}] - \sum_{j=1}^{n-1} [-z_v \log(z_v - z_{(j)}) - z_{(j)}] \\
& - z_v \log(z_{(n)}) + z_{(n)} \\
\frac{z_v A_v^2}{n} & = -z_v + 2z_v \log(z_v) + \sum_{j=1}^n \frac{(j-1)^2 z_v (\log(z_{(j)}) - \log(z_v - z_{(j)}))}{n^2} \\
& - \sum_{j=1}^{n-1} \left[\frac{j^2 z_v (\log(z_{(j)}) - \log(z_v - z_{(j)}))}{n^2} \right] - \sum_{j=1}^n \frac{2(j-1)z_v \log(z_v - z_{(j)})}{n} \\
& + \sum_{j=1}^{n-1} \frac{2jz_v \log(z_v - z_{(j)})}{n} - \sum_{j=1}^n z_v \log(z_v - z_{(j)}) - \sum_{j=1}^n z_{(j)} + \sum_{j=1}^{n-1} z_v \log(z_v - z_{(j)}) \\
& + \sum_{j=1}^n z_{(j)} - z_v \log(z_{(n)}) \\
\frac{z_v A_v^2}{n} & = -z_v + 2z_v \log(z_v) + \sum_{j=1}^n \frac{(j-1)^2 z_v (\log(z_{(j)}) - \log(z_v - z_{(j)}))}{n^2} \\
& - \sum_{j=1}^{n-1} \left[\frac{j^2 z_v (\log(z_{(j)}) - \log(z_v - z_{(j)}))}{n^2} \right] - \sum_{j=1}^n \frac{2(j-1)z_v \log(z_v - z_{(j)})}{n} \\
& + \sum_{j=1}^{n-1} \frac{2jz_v \log(z_v - z_{(j)})}{n} - z_v \log(z_{(n)}) - z_v \log(z_v - z_{(n)}) \\
\frac{z_v A_v^2}{n} & = -z_v + 2z_v \log(z_v) + \sum_{j=1}^n \frac{(j-1)^2 z_v (\log(z_{(j)}) - \log(z_v - z_{(j)}))}{n^2} \\
& - \sum_{j=1}^{n-1} \left[\frac{j^2 z_v (\log(z_{(j)}) - \log(z_v - z_{(j)}))}{n^2} \right] - \sum_{j=1}^n \frac{2(j-1)z_v \log(z_v - z_{(j)})}{n} \\
& + \sum_{j=1}^{n-1} \frac{2jz_v \log(z_v - z_{(j)})}{n} - z_v \log(z_{(n)}) + z_v \log(z_v - z_{(n)}) \\
\frac{z_v A_v^2}{n} & = -z_v + 2z_v \log(z_v) + \sum_{j=1}^n \frac{(j-1)^2 z_v (\log(z_{(j)}) - \log(z_v - z_{(j)}))}{n^2} \\
& - \sum_{j=1}^{n-1} \left[\frac{j^2 z_v (\log(z_{(j)}) - \log(z_v - z_{(j)}))}{n^2} \right] + \sum_{j=1}^n \frac{2z_v \log(z_v - z_{(j)})}{n} \\
& + 2z_v \log(z_v - z_{(n)}) - z_v \log(z_{(n)}) - z_v \log(z_v - z_{(n)})
\end{aligned}$$

$$\begin{aligned}
\frac{z_v A_v^2}{n} &= -z_v + 2z_v \log(z_v) + \sum_{j=1}^n \frac{(j-1)^2 z_v (\log(z_{(j)}) - \log(z_v - z_{(j)}))}{n^2} \\
&\quad - \sum_{j=1}^{n-1} \left[\frac{j^2 z_v (\log(z_{(j)}) - \log(z_v - z_{(j)}))}{n^2} \right] + \sum_{j=1}^n \frac{2z_v \log(z_v - z_{(j)})}{n} \\
&\quad - z_v \log(z_{(n)}) + z_v \log(z_v - z_{(n)}) \\
\frac{z_v A_v^2}{n} &= -z_v + 2z_v \log(z_v) + \sum_{j=1}^n \frac{(j-1)^2 z_v (\log(z_{(j)}) - \log(z_v - z_{(j)}))}{n^2} \\
&\quad - \sum_{j=1}^n \left[\frac{j^2 z_v (\log(z_{(j)}) - \log(z_v - z_{(j)}))}{n^2} \right] + \sum_{j=1}^n \frac{2z_v \log(z_v - z_{(j)})}{n} \\
\frac{z_v A_v^2}{n} &= -z_v + 2z_v \log(z_v) + \sum_{j=1}^n \frac{(1-2j)z_v (\log(z_{(j)}) - \log(z_v - z_{(j)}))}{n^2} + \sum_{j=1}^n \frac{2z_v \log(z_v - z_{(j)})}{n} \\
A_v^2 &= -n + 2n \log(z_v) + \sum_{j=1}^n \frac{(1-2j) (\log(z_{(j)}) - \log(z_v - z_{(j)}))}{n} + \sum_{j=1}^n 2 \log(z_v - z_{(j)}) \\
A_v^2 &= -n + 2n \log(z_v) + \sum_{j=1}^n \left[\frac{(1-2j)}{n} \log(z_{(j)}) - \frac{(1+2(n-j))}{n} \log(z_v - z_{(j)}) \right]
\end{aligned}$$

A.2.4 Derivation of the Right-Truncated Upper Tail Modified Anderson-Darling Goodness of Fit Test statistic.

Following the method of Chernobai et al. [23] we use the probability integral tranform to change to the uniformly distributed z values for the distribution:

$$A_{v,up}^2 = n \int_{-\infty}^v \frac{(F_n(x) - \hat{F}(x))^2}{(1 - \hat{F}(x))} d\hat{F}(x) = \frac{n}{z_v} \int_0^{z_v} \frac{(F_n(z^*)(z_v) - z)^2}{(z_v - z)} dz$$

From here, we can use the observations of the data set $x_j, j = 1, 2, \dots, n$ to define F_n and re-write the above as

$$\begin{aligned}
\frac{z_v AD_{up}^{2*}}{n} &= \int_0^{z_v} \frac{(F_n(z)z_v - z)^2}{z_v - z} dz \\
&= \int_0^{z_1} \frac{z^2}{z_v - z} + \sum_{j=1}^{n-1} \int_{z_j}^{z_{j+1}} \frac{(\frac{z_v j}{n} - z)^2}{z_v - z} dz + \int_{z_n}^{z_v} \frac{(z_v - z)^2}{z_v - z} dz \\
&= \left[-zz_v - z_v^2 \log(z_v - z) - \frac{z^2}{2} \right]_0^{z_1} + \sum_{j=1}^{n-1} \left[-\log(z_v - z) \left(\frac{jz_v}{n} \right)^2 \right]_{z_j}^{z_{j+1}}
\end{aligned}$$

$$\begin{aligned}
& - \sum_{j=1}^{n-1} \left[\left(\frac{2z_v j}{n} \right) (-z - z_v \log(z_v - z)) \right]_{z_j}^{z_{j+1}} + \sum_{j=1}^{n-1} \left[-z z_v - z_v^2 \log(z_v - z) - \frac{z^2}{2} \right]_{z_j}^{z_{j+1}} \\
& + \left[z_v z - \frac{z^2}{2} \right]_{z_n}^{z_v} \\
& = -z_1 z_v - z_v^2 \log(z_v - z_1) - \frac{z_1^2}{2} + z_v^2 \log(z_v) - \sum_{j=1}^{n-1} \log(z_v - z_{j+1}) \left(\frac{z_v j}{n} \right)^2 \\
& + \sum_{j=1}^{n-1} \log(z_v - z_j) \left(\frac{z_v j}{n} \right)^2 - \sum_{j=1}^{n-1} \left(\frac{2z_v j}{n} \right) (-z_{j+1} - z_v \log(z_v - z_{j+1})) \\
& + \sum_{j=1}^{n-1} \left(\frac{2z_v j}{n} \right) (-z_j - z_v \log(z_v - z_j)) - \sum_{j=1}^{n-1} z_v z_{j+1} - \sum_{j=1}^{n-1} z_v^2 \log(z_v - z_{j+1}) \\
& - \sum_{j=1}^{n-1} \frac{z_{j+1}^2}{2} + \sum_{j=1}^{n-1} z_v z_j + \sum_{j=1}^{n-1} z_v^2 \log(z_v - z_j) + \sum_{j=1}^{n-1} \frac{z_j^2}{2} + z_v^2 - \frac{z_v^2}{2} - z_v z_n + \frac{z_n^2}{2} \\
& = z_v^2 \log(z_v) - \sum_{j=1}^n \log(z_v - z_j) \left(\frac{z_v(j-1)}{n} \right)^2 + \sum_{j=1}^{n-1} \log(z_v - z_j) \left(\frac{z_v j}{n} \right)^2 \\
& - \sum_{j=1}^n \left(\frac{2z_v(j-1)}{n} \right) (-z_j - z_v \log(z_v - z_j)) + \sum_{j=1}^{n-1} \left(\frac{2z_v j}{n} \right) (-z_j - z_v \log(z_v - z_j)) \\
& - z_v^2 \log(z_v - z_1) - \sum_{j=2}^n z_v^2 \log(z_v - z_j) + \sum_{j=1}^{n-1} z_v^2 \log(z_v - z_j) - \sum_{j=1}^n \frac{z_j^2}{2} + \sum_{j=1}^n \frac{z_j^2}{2} \\
& + \frac{z_v^2}{2} - 2z_v z_n \\
& = z_v^2 \log(z_v) - \sum_{j=1}^n \log(z_v - z_j) \left(\frac{z_v^2(1-2j)}{n^2} \right) - z_v^2 \log(z_v - z_n) \\
& + \sum_{j=1}^n \left(\frac{2z_v}{n} \right) (-z_j - z_v \log(z_v - z_j)) - (2z_v) (-z_n - z_v \log(z_v - z_n)) \\
& - z_v^2 \log(z_v - z_n) + \frac{z_v^2}{2} - 2z_v z_n \\
& = \frac{z_v^2}{2} + z_v^2 \log(z_v) - \sum_{j=1}^n \left(\frac{2z_v}{n} \right) z_j - \sum_{j=1}^n \log(z_v - z_j) \left(\frac{2z_v^2}{n} - \frac{z_v^2(2j-1)}{n^2} \right) \\
& - z_v^2 \log(z_v - z_n) + 2z_v z_n + 2z_v^2 \log(z_v - z_n) - z_v^2 \log(z_v - z_n) - 2z_v z_n \\
& = \frac{z_v^2}{2} + z_v^2 \log(z_v) - \sum_{j=1}^n \left(\frac{2z_v}{n} \right) z_j - \sum_{j=1}^n \log(z_v - z_j) \left(\frac{2z_v^2}{n} - \frac{z_v^2(2j-1)}{n^2} \right)
\end{aligned}$$

$$AD_{up}^{2*} = \frac{nz_v}{2} + z_v n \log(z_v) - \sum_{j=1}^n 2z_j - \sum_{j=1}^n \log(z_v - z_j) \left(2z_v - \frac{z_v(2j-1)}{n}\right)$$

A.2.5 Derivation of the Right-Truncated Lower Tail Modified Anderson-Darling Goodness of Fit Test statistic.

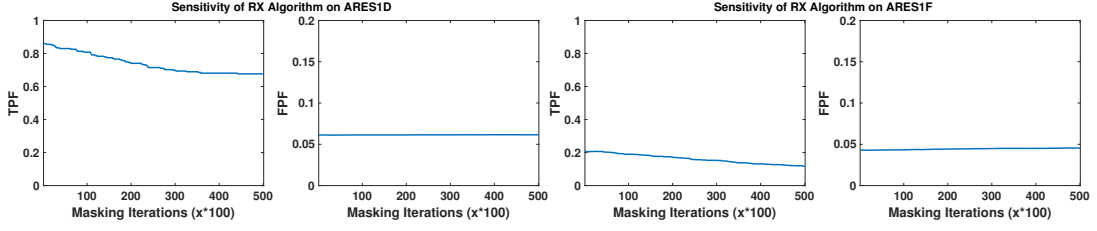
The calculations follow from the Upper Tail statistic above.

$$\begin{aligned} \frac{z_v AD_{low}^{2*}}{n} &= \int_0^{z_v} \frac{(F_n(z)z_v - z)^2}{z} dz \\ &= \int_0^{z_1} \frac{z^2}{z} + \sum_{j=1}^{n-1} \int_{z_j}^{z_{j+1}} \frac{(\frac{z_v j}{n} - z)^2}{z} dz + \int_{z_n}^{z_v} \frac{(z_v - z)^2}{z} dz \\ &= \left[\frac{z^2}{2} \right]_0^{z_1} + \sum_{j=1}^{n-1} \left[\log(z) \left(\frac{jz_v}{n} \right)^2 - \left(\frac{2z_v j}{n} \right) z + \frac{z^2}{2} \right]_{z_j}^{z_{j+1}} + \left[z_v^2 \log z - 2z_v z + \frac{z^2}{2} \right]_{z_n}^{z_v} \\ &= \frac{z_1^2}{2} + \sum_{j=1}^{n-1} \log(z_{j+1}) \left(\frac{jz_v}{n} \right)^2 - \sum_{j=1}^{n-1} \log(z_j) \left(\frac{jz_v}{n} \right)^2 - \sum_{j=1}^{n-1} \left(\frac{2z_v j}{n} \right) z_{j+1} + \sum_{j=1}^{n-1} \left(\frac{2z_v j}{n} \right) z_j \\ &\quad + \sum_{j=1}^{n-1} \frac{z_{j+1}^2}{2} - \sum_{j=1}^{n-1} \frac{z_j^2}{2} + z_v^2 \log z_v - 2z_v^2 + \frac{z_v^2}{2} - z_v^2 \log z_n + 2z_v z_n - \frac{z_n^2}{2} \\ &= \frac{z_1^2}{2} + \sum_{j=2}^n \frac{z_j^2}{2} - \sum_{j=1}^{n-1} \frac{z_j^2}{2} - \frac{z_n^2}{2} + \sum_{j=1}^n \log(z_j) \left(\frac{(j-1)z_v}{n} \right)^2 - \sum_{j=1}^{n-1} \log(z_j) \left(\frac{jz_v}{n} \right)^2 \\ &\quad - \sum_{j=1}^n \left(\frac{2z_v(j-1)}{n} \right) z_j + \sum_{j=1}^{n-1} \left(\frac{2z_v j}{n} \right) z_j + z_v^2 \log z_v - 2z_v^2 + \frac{z_v^2}{2} - z_v^2 \log z_n + 2z_v z_n \\ &= \sum_{j=1}^n \log(z_j) \left(\frac{(j-1)z_v}{n} \right)^2 - \sum_{j=1}^n \log(z_j) \left(\frac{jz_v}{n} \right)^2 - \sum_{j=1}^n \left(\frac{2z_v(j-1)}{n} \right) z_j \\ &\quad + \sum_{j=1}^{n-1} \left(\frac{2z_v j}{n} \right) z_j + z_v^2 \log z_v - \frac{3z_v^2}{2} + 2z_v z_n \\ &= \sum_{j=1}^n \log(z_j) \left(\frac{(1-2j)z_v^2}{n^2} \right) - \sum_{j=1}^n \left(\frac{2z_v(j-1)}{n} \right) z_j + \sum_{j=1}^n \left(\frac{2z_v j}{n} \right) z_j + z_v^2 \log z_v - \frac{3z_v^2}{2} \\ &= \sum_{j=1}^n \log(z_j) \left(\frac{(1-2j)z_v^2}{n^2} \right) + \sum_{j=1}^n \left(\frac{2z_v}{n} \right) z_j + z_v^2 \log z_v - \frac{3z_v^2}{2} \end{aligned}$$

$$AD_{low}^{2*} = -\frac{3nz_v}{2} + nz_v \log z_v + \sum_{j=1}^n \log(z_j) \left(\frac{(1-2j)z_v}{n} \right) + \sum_{j=1}^n 2z_j$$

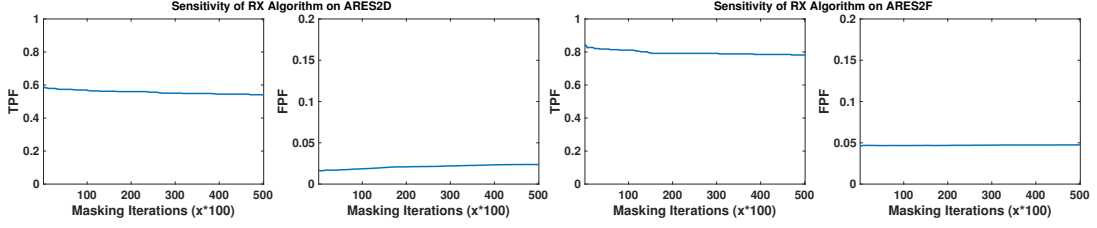
Appendix B: Quantitative Sensitivity Analysis Charts

In this Appendix we present the graphics of each algorithm's performance on the selection of ten images.



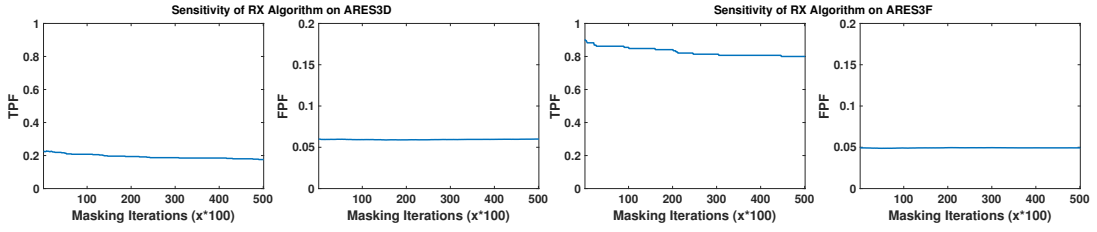
(a) ARES1D

(b) ARES1F



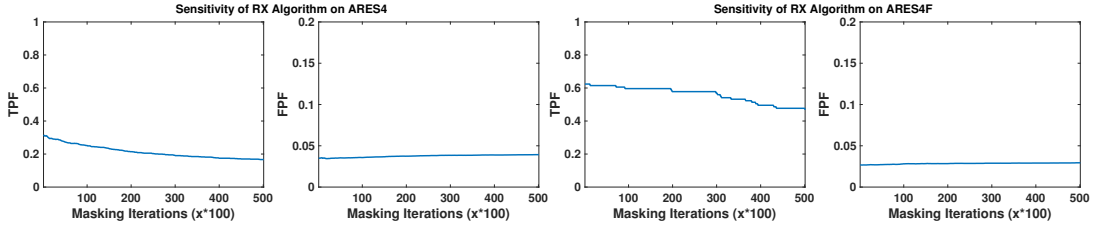
(c) ARES2D

(d) ARES2F



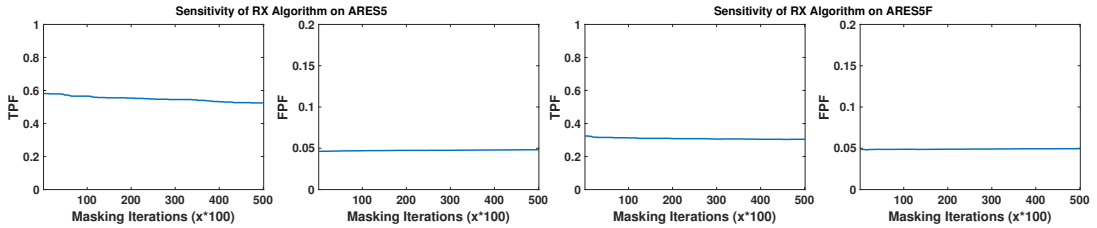
(e) ARES3D

(f) ARES3F



(g) ARES4

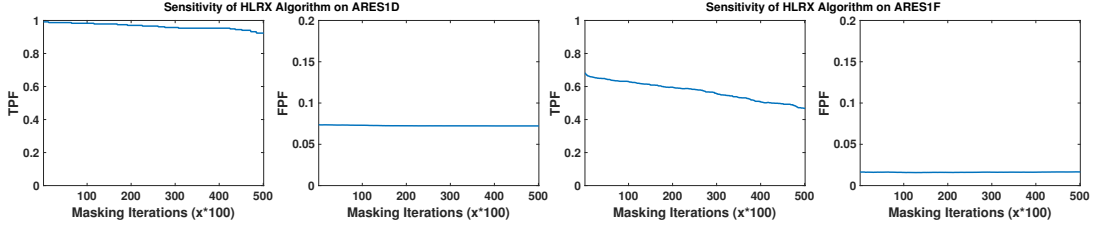
(h) ARES4F



(i) ARES5

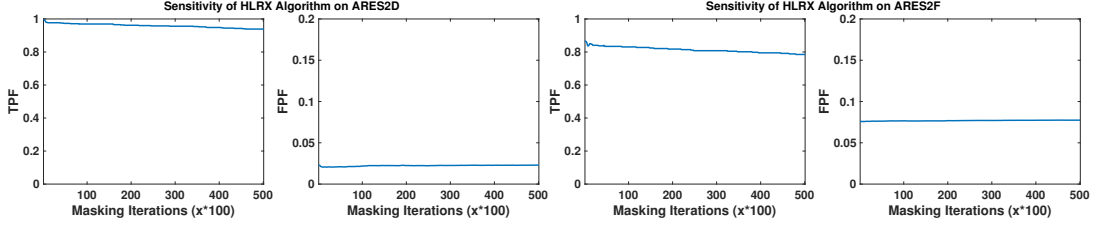
(j) ARES5F

Figure B.1: Mahalanobis Masking Performance of RX Algorithm



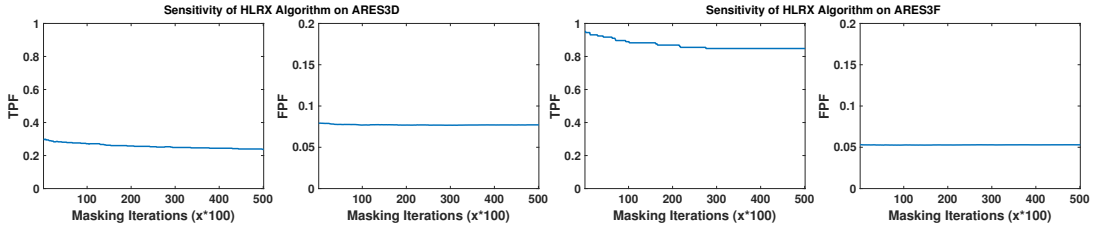
(a) ARES1D

(b) ARES1F



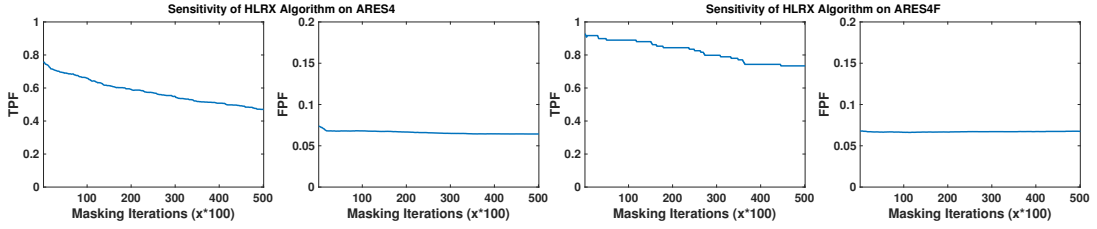
(c) ARES2D

(d) ARES2F



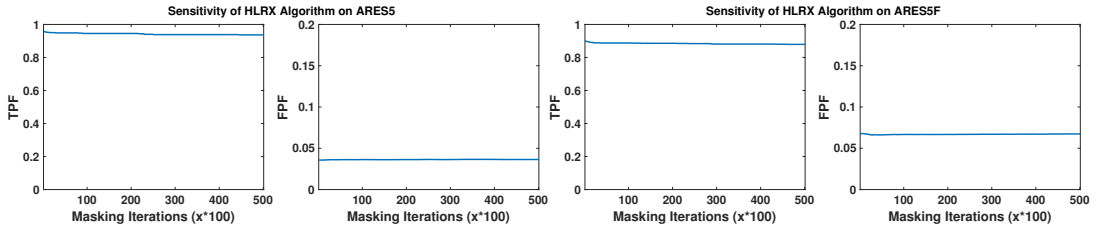
(e) ARES3D

(f) ARES3F



(g) ARES4

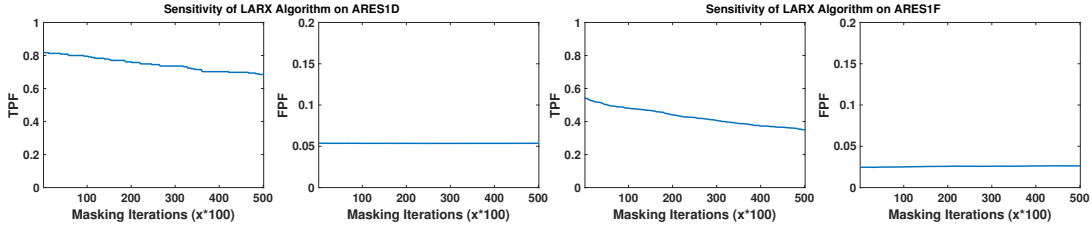
(h) ARES4F



(i) ARES5

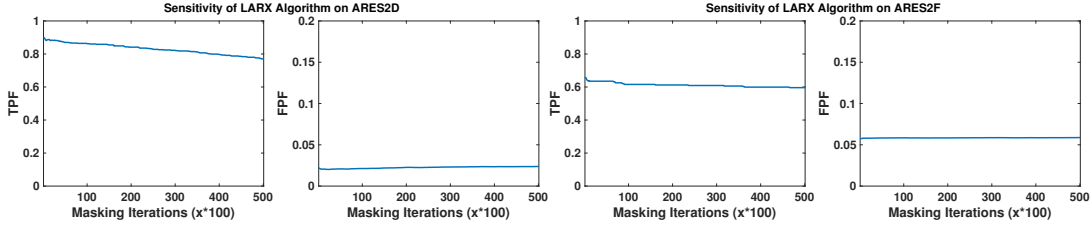
(j) ARES5F

Figure B.2: Mahalanobis Masking Performance of HLRX Algorithm



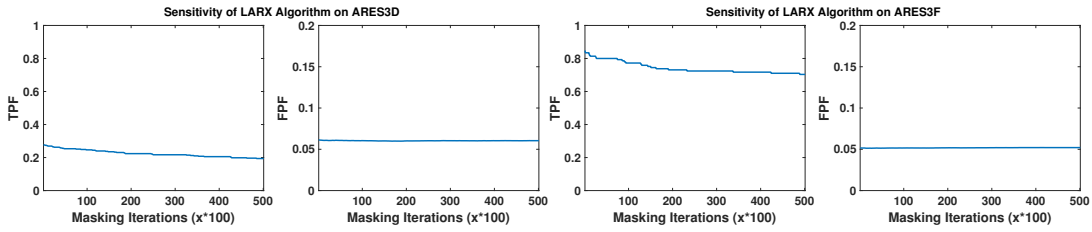
(a) ARES1D

(b) ARES1F



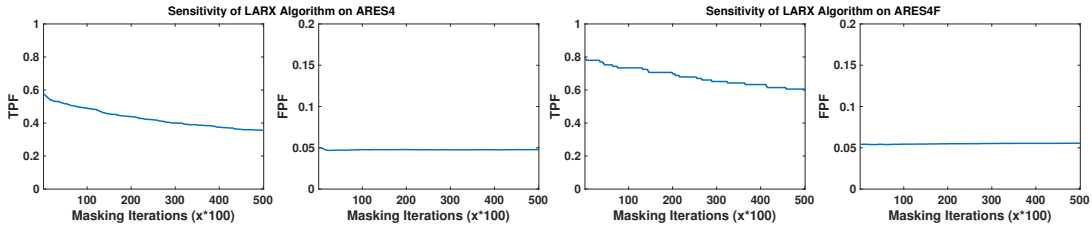
(c) ARES2D

(d) ARES2F



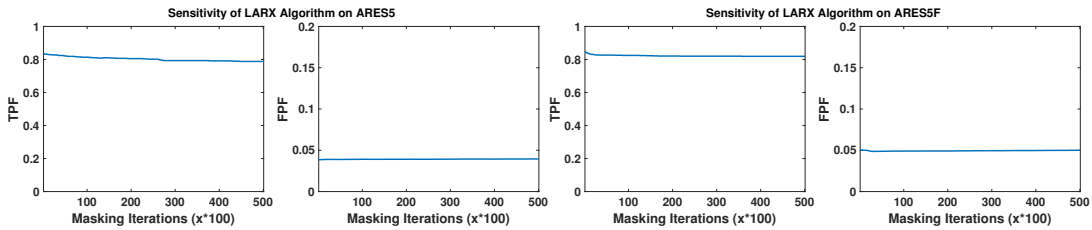
(e) ARES3D

(f) ARES3F



(g) ARES4

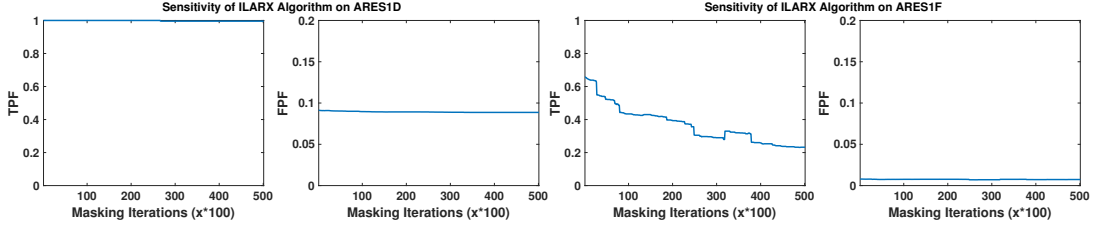
(h) ARES4F



(i) ARES5

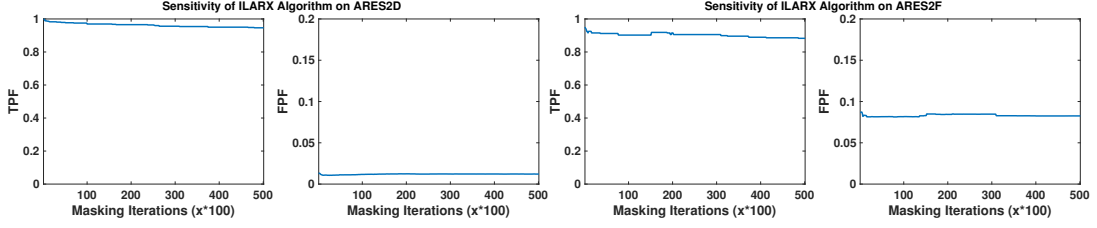
(j) ARES5F

Figure B.3: Mahalanobis Masking Performance of LRX Algorithm



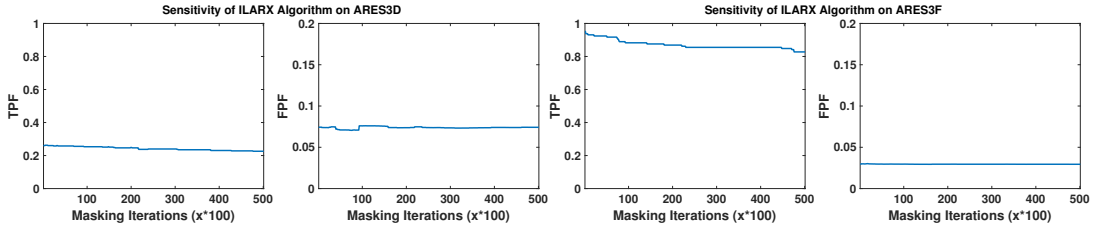
(a) ARES1D

(b) ARES1F



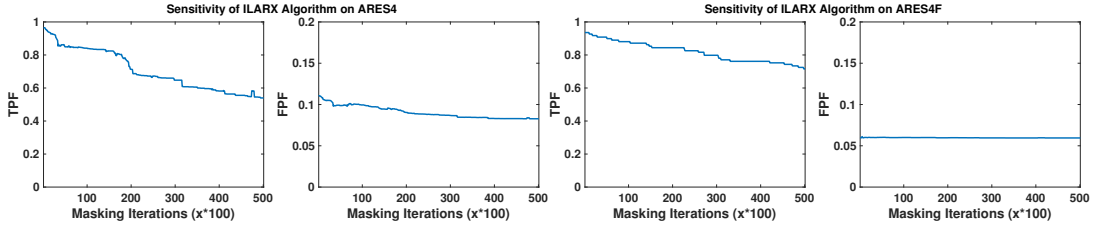
(c) ARES2D

(d) ARES2F



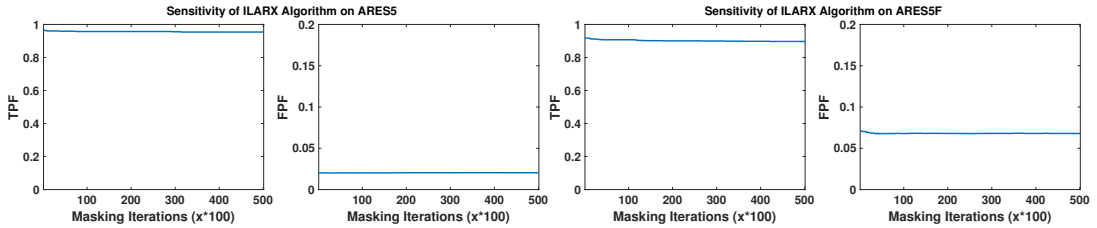
(e) ARES3D

(f) ARES3F



(g) ARES4

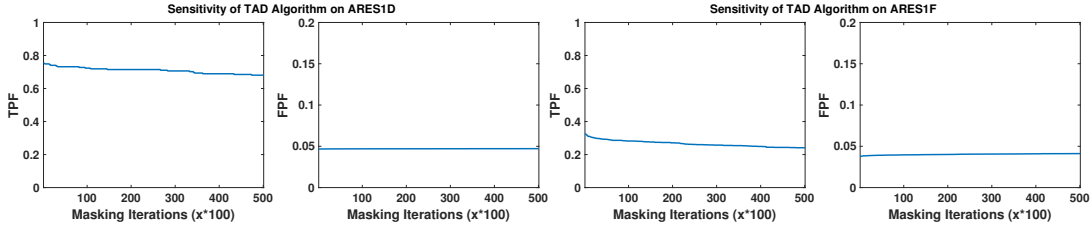
(h) ARES4F



(i) ARES5

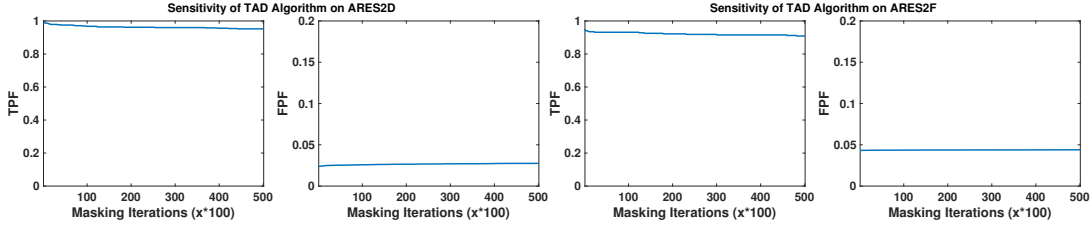
(j) ARES5F

Figure B.4: Mahalanobis Masking Performance of ILRX Algorithm



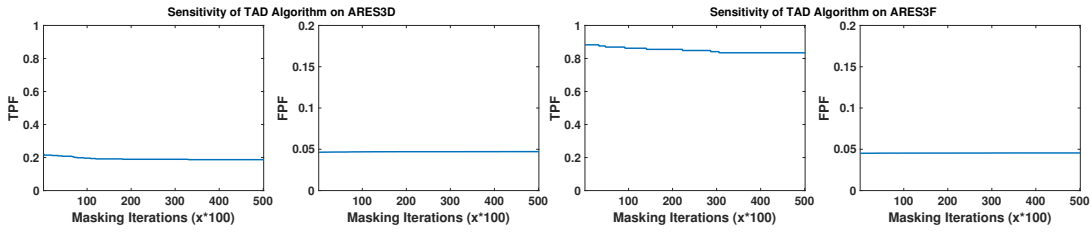
(a) ARES1D

(b) ARES1F



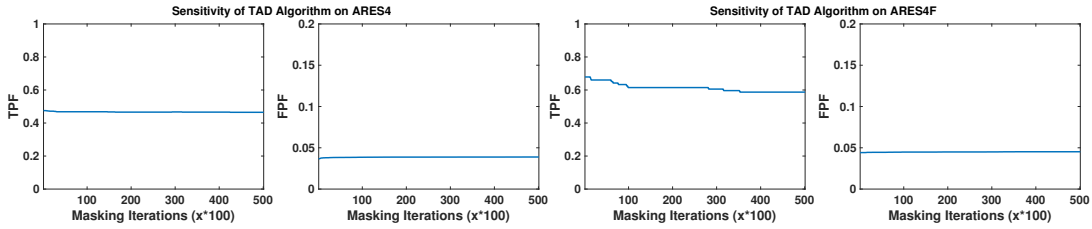
(c) ARES2D

(d) ARES2F



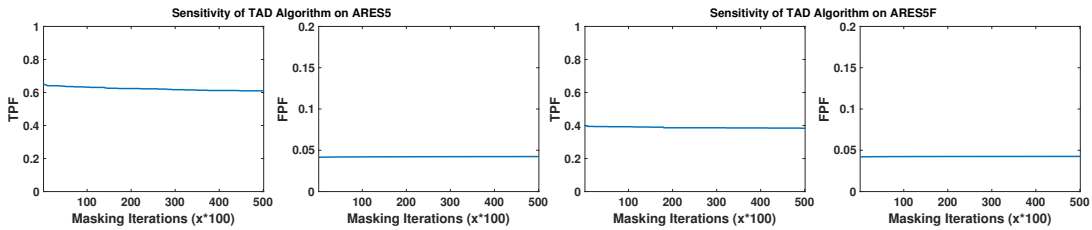
(e) ARES3D

(f) ARES3F



(g) ARES4

(h) ARES4F



(i) ARES5

(j) ARES5F

Figure B.5: Mahalanobis Masking Performance of TAD Algorithm

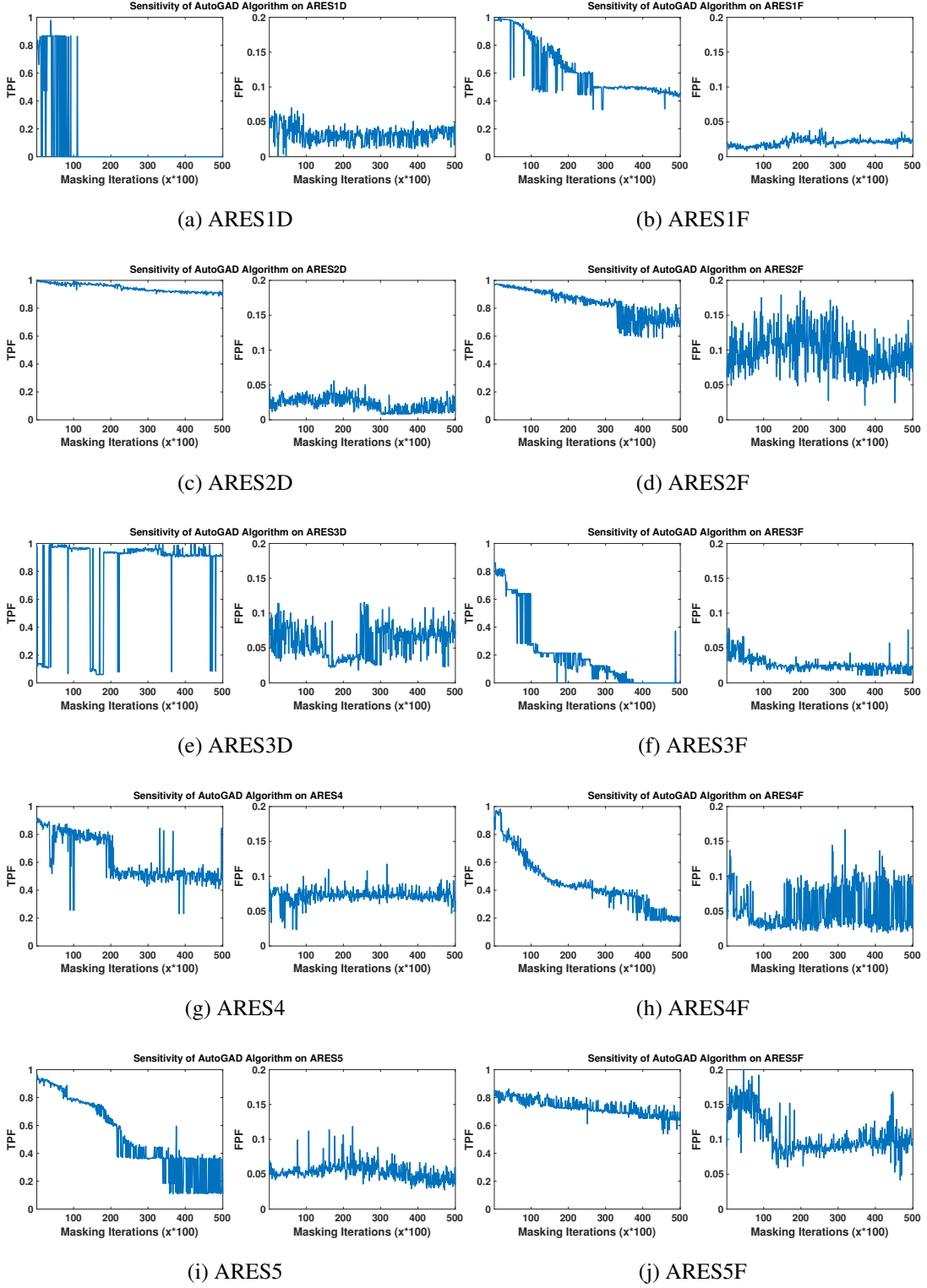


Figure B.6: Mahalanobis Masking Performance of AutoGAD Algorithm

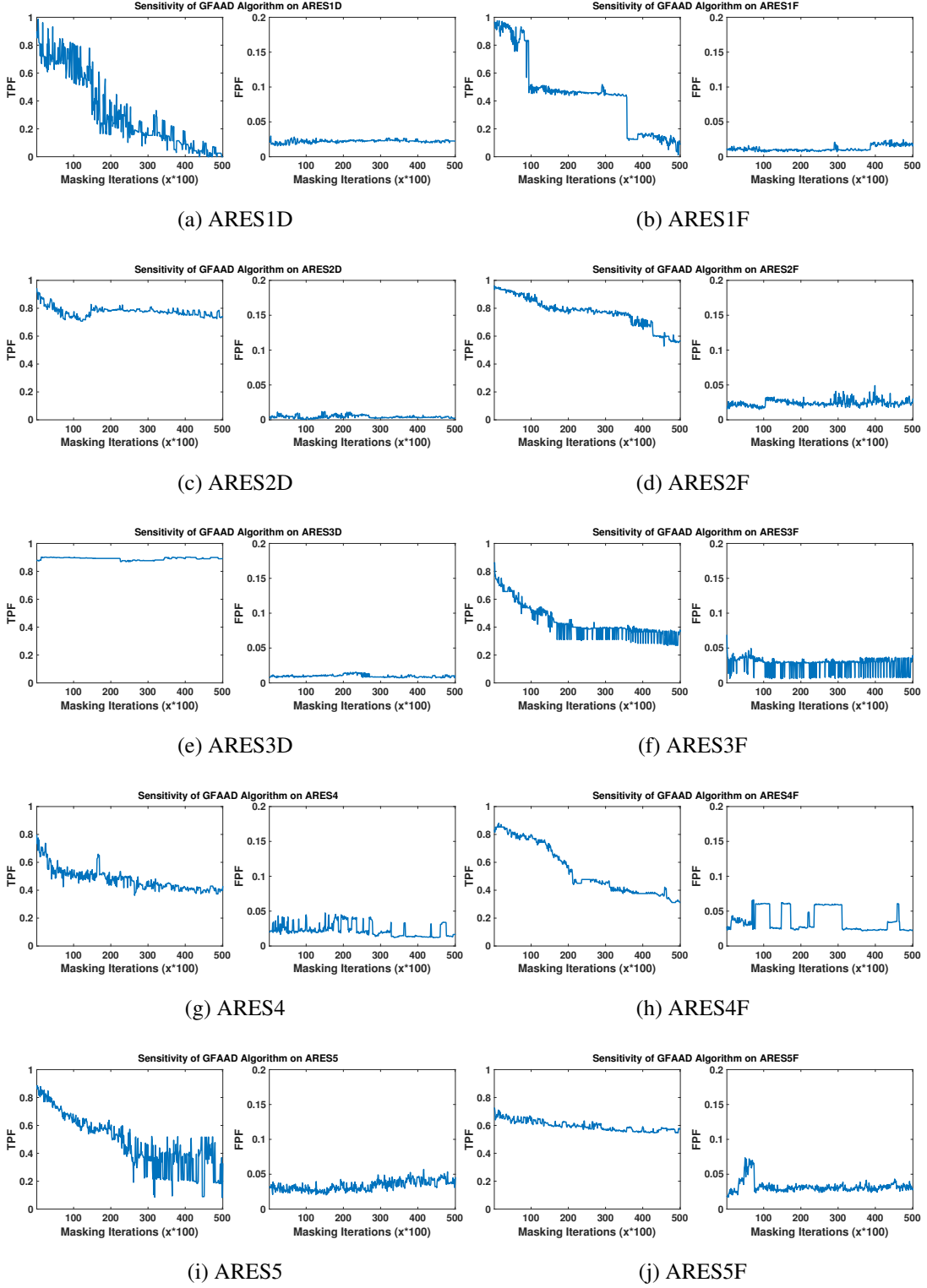


Figure B.7: Mahalanobis Masking Performance of GFAAD Algorithm

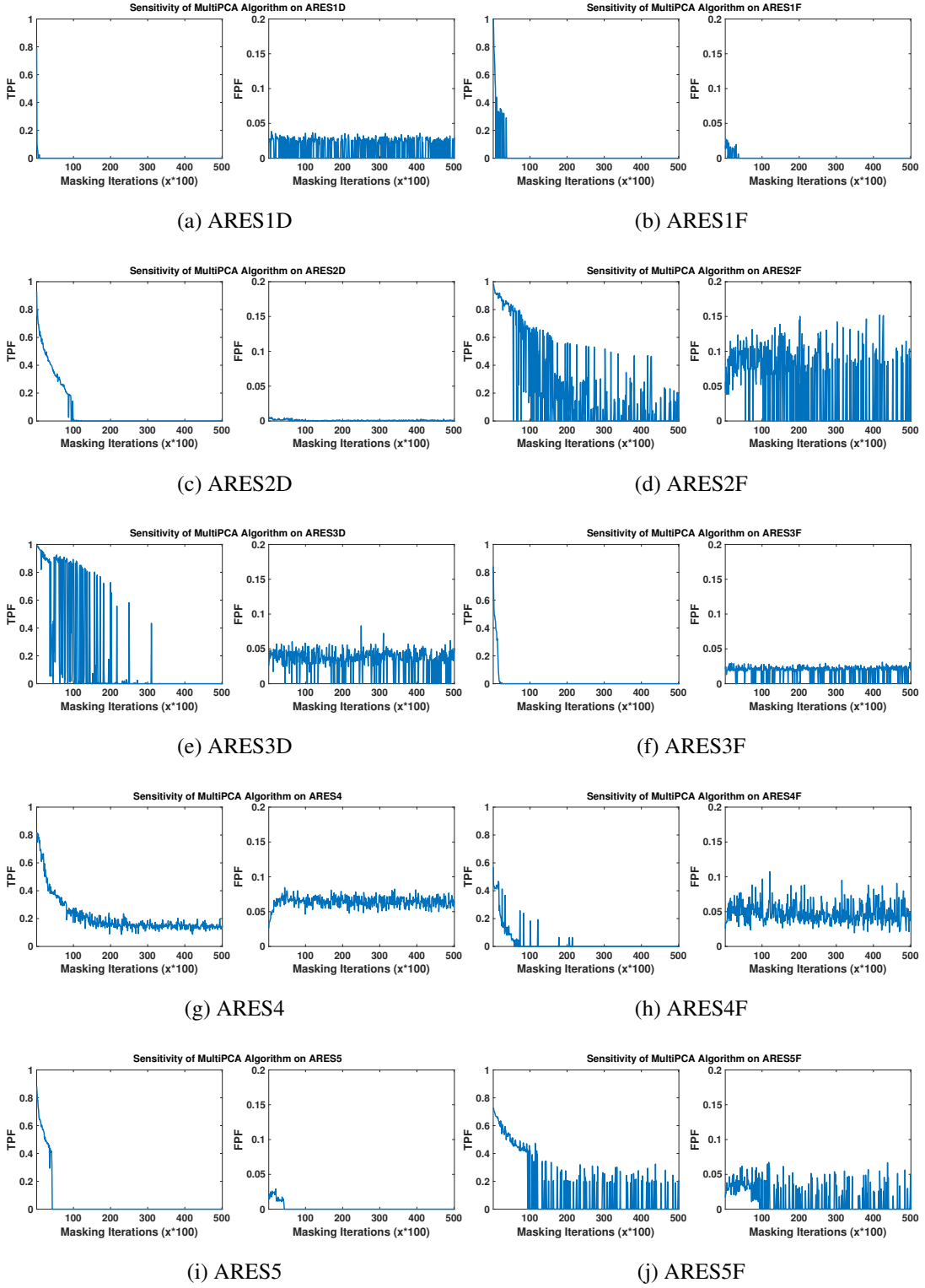
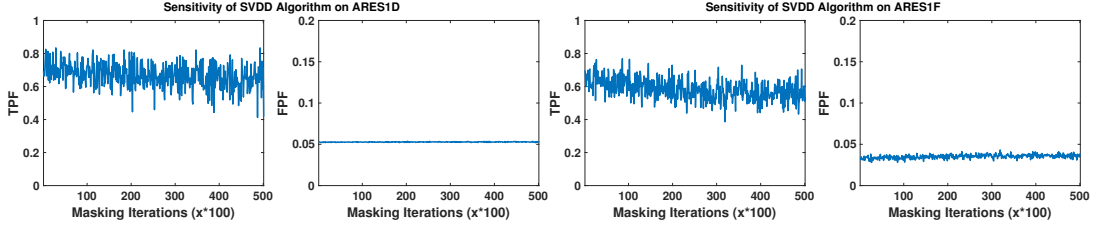
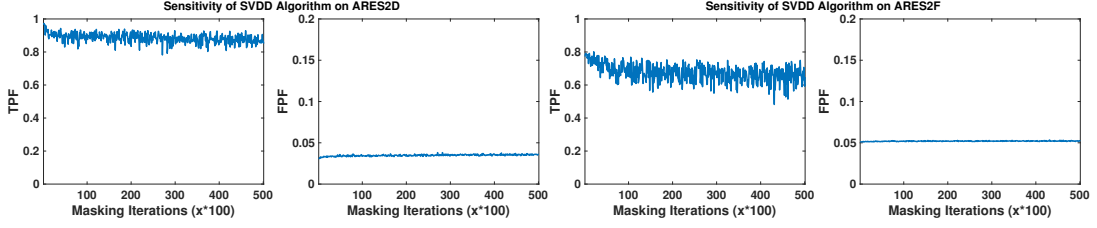


Figure B.8: Mahalanobis Masking Performance of MultiPCA Algorithm



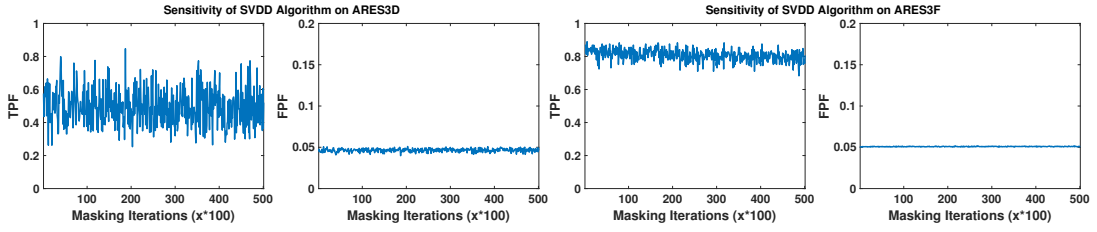
(a) ARES1D

(b) ARES1F



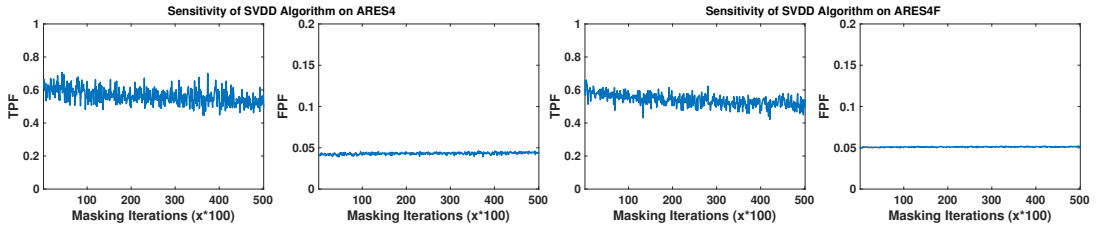
(c) ARES2D

(d) ARES2F



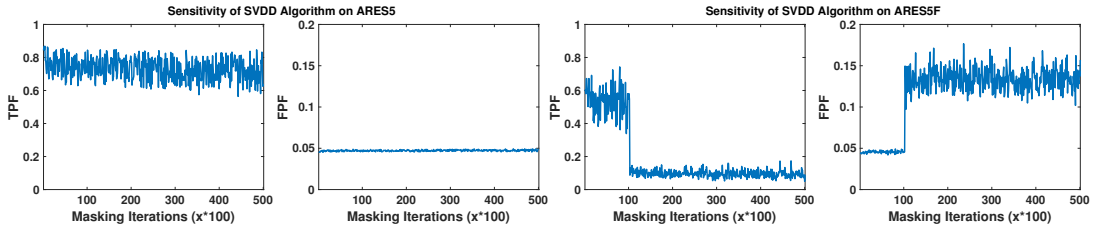
(e) ARES3D

(f) ARES3F



(g) ARES4

(h) ARES4F



(i) ARES5

(j) ARES5F

Figure B.9: Mahalanobis Masking Performance of SVDD Algorithm

Appendix C: List of Acronyms

AutoGAD	Autonomous Global Anomaly Detector
ATEEVT	Adaptive Threshold Estimation via Extreme Value Theory
AUC	Area Under the Curve
BACON	Blocked Adaptive Computationally Efficient Outlier Nominators
BIC	Bayesian Information Criterion
CDF	Cumulative Distribution Function
CFAR	Constant False Alarm Rate
ED	Euclidean Distance
EFA	Exploratory Factor Analysis
EVT	Extreme Value Theory
ExtRX	Extreme RX
ExtLRX	Extreme LRX
ExtILRX	Extreme Iterative Linear RX (ILRX)
ExtHLRX	Extreme HLRX
FA	Factor Analysis
FAR	False Alarm Rate
FN	False Negative
FP	False Positive
FPF	False Positive Fraction
GEV	Generalized Extreme Value
GFAAD	Global Factor Analysis-based Anomaly Detector
GLRT	Generalized Likelihood Ratio Test
GMRF	Gauss–Markov Random Field
GoF	Goodness-of-Fit
GPD	Generalized Pareto Distribution
GRASP	Greedy Randomized Adaptive Search Procedures
HLRX	High-Low FAR LRX
HSI	Hyperspectral Imaging
HYDICE	Hyperspectral Digital Imagery Collection Equipment
ICA	Independent Component Analysis
ILRX	Iterative Linear RX
IRX	Iterative RX
KKT	Karush-Kuhn-Tucker
LA	Label Accuracy
LRX	Linear RX
MCC	Matthew’s Correlation Coefficient
MD	Mahalanobis Distance
MDM	Multivariate Delta Method
MLE	Maximum Likelihood Estimation
MSE	Mean-Squared Error

MultiPCA	Multiple PCA Anomaly Detector
nm	nanometers
NPV	Negative Predictive Value
NN	nearest neighbor
PA-SNR	potential anomaly signal to noise ratio
PCA	Principal Component Analysis
POT	<i>points over threshold</i>
pdf	probability density function
PPV	Positive Predictive Value
PUT	Pixel Under Test
RBF	Radial Basis Function
RBFNN	Radial Basis Function Neural Network
RSM	Response Surface Methodology
ROC	Response Operating Characteristic
RPD	Robust Parameter Design
RX	Reed–Xiaoli
SNR	Signal-to-Noise Ratio
SVD	Singular Value Decomposition
SVDD	Support Vector Data Description
SVM	support vector machine
TAD	Topological Anomaly Detector
TAD-RX	Topological Anomaly Detector-RX
TEVAT	Truncated Extreme Value Anomaly Thresholding
TN	True Negative
TP	True Positive
TPF	True Positive Fraction
tGoF	truncated Goodness-of-Fit (GoF)
tGPD	truncated Generalized Pareto distribution

Bibliography

- [1] Acito, N., M. Diani, and G. Corsini. "On the CFAR property of the RX algorithm in the presence of signal-dependent noise in hyperspectral images." *IEEE Transactions on Geoscience and Remote Sensing*, 51(6):3475–3491, June 2013, 0196-2892.
- [2] Ahmad, Muhammad Idrees, CD Sinclair, and BD Spurr. "Assessment of flood frequency models using empirical distribution function statistics." *Water Resources Research*, 24(8):1323–1328, 1988.
- [3] Baldi, Pierre, Søren Brunak, Yves Chauvin, Claus AF Andersen, and Henrik Nielsen. "Assessing the accuracy of prediction algorithms for classification: an overview." *Bioinformatics*, 16(5):412–424, 2000.
- [4] Balkema, August A and Laurens De Haan. "Residual life time at great age." *The Annals of Probability*, 2(5):792–804, October 1974.
- [5] Banerjee, Amit, Philippe Burlina, and Chris Diehl. "A support vector method for anomaly detection in hyperspectral imagery." *IEEE Transactions on Geoscience and Remote Sensing*, 44(8):2282–2291, 2006.
- [6] Barrett Jr, Carl R. "Adaptive Thresholding and Automatic Detection." Jerry L. Eaves and Edward K. Reedy (editors), *Principles of Modern Radar*, 368–393. Springer, 1987.
- [7] Bartlett, Brent D, Ariel Schlamm, Carl Salvaggio, and David W Messinger. "Anomaly detection of man-made objects using spectropolarimetric imagery." *Proc. SPIE*, vol. 8048, 80480B. 2011.
- [8] Basener, William F, Emmett J. Ientilucci, and David W. Messinger. "Anomaly detection using topology." *Proc. SPIE*, vol. 6565, 65650J. 2007.
- [9] Basener, William F and David W Messinger. "Enhanced detection and visualization of anomalies in spectral imagery." *Proc. SPIE*, vol. 7334, 73341Q. 2009.
- [10] Basener, William F, Eric Nance, and John Kerekes. "The target implant method for predicting target difficulty and detector performance in hyperspectral imagery." *Proc. SPIE*, vol. 8048, 80481H. 2011.
- [11] Bates, Ron A, Ron S Kenett, David M Steinberg, and Henry P Wynn. "Achieving robust design from computer simulations." *Quality Technology & Quantitative Management*, 3(2):161–177, 2006.

- [12] Bekkar, Mohamed, Hassiba Khelouane Djema, and Taklit Akrouf Alitouche. "Evaluation measures for models assessment over imbalanced data sets." *Journal of Information Engineering and Applications*, 3(10), 2013.
- [13] Bellucci, Joseph P. *Non-linear metamodeling extensions to the robust parameter design of computer simulations*. Ph.D. thesis, Air Force Institute of Technology, WPAFB, 2016.
- [14] Bellucci, Joseph P. and Kenneth W. Bauer Jr. "A Taylor series approach to the robust parameter design of computer simulations using Kriging and radial basis function neural networks." *International Journal of Quality Engineering and Technology*, 2017 (to appear).
- [15] Bellucci, Joseph P., Timothy E. Smetek, and Kenneth W. Bauer. "Improved Hyperspectral Image Processing Algorithm Testing Using Synthetic Imagery and Factorial Designed Experiments." *IEEE Transactions on Geoscience and Remote Sensing*, 48(3):1211–1223, March 2010, 0196-2892.
- [16] Benediktsson, Jón Atli, Jón Aegvar Palmason, and Johannes R Sveinsson. "Classification of hyperspectral data from urban areas based on extended morphological profiles." *IEEE Transactions on Geoscience and Remote Sensing*, 43(3):480–491, 2005.
- [17] Billor, Nedret, Ali S. Hadi, and Paul F. Velleman. "BACON: blocked adaptive computationally efficient outlier nominators." *Computational Statistics & Data Analysis*, 34(3):279 – 298, 2000, 0167-9473.
- [18] Boser, Bernhard E., Isabelle M. Guyon, and Vladimir N. Vapnik. "A Training Algorithm for Optimal Margin Classifiers." *Proceedings of the Fifth Annual Workshop on Computational Learning Theory, COLT '92*, 144–152. ACM, 1992.
- [19] Broadwater, Joshua B and Rama Chellappa. "Adaptive threshold estimation via extreme value theory." *IEEE Transactions on Signal Processing*, 58(2):490–500, 2010.
- [20] Cannistraci, Carlo V, Franco M Montevicchi, and Massimo Alessio. "Median-modified Wiener filter provides efficient denoising, preserving spot edge and morphology in 2-DE image processing." *Proteomics*, 9(21):4908–4919, 2009.
- [21] Casella, George and Roger L Berger. *Statistical Inference*, vol. 2. Duxbury, 2002.
- [22] Chandola, Varun, Arindam Banerjee, and Vipin Kumar. "Anomaly Detection: A Survey." *ACM Computer Surveys*, 41(3):15:1–15:58, July 2009, 0360-0300.
- [23] Chernobai, Anna, Svetlozar T Rachev, and Frank J Fabozzi. "Composite goodness-of-fit tests for left-truncated loss samples." *Handbook of Financial Econometrics and Statistics*, 575–596. Springer, 2015.

- [24] Chiang, Shao-Shan, Chein-I Chang, and Irving W Ginsberg. “Unsupervised target detection in hyperspectral images using projection pursuit.” *IEEE Transactions on Geoscience and Remote Sensing*, 39(7):1380–1391, 2001.
- [25] Choi, Seungjin, Andrzej Cichocki, Hyung-Min Park, and Soo-Young Lee. “Blind source separation and independent component analysis: a review.” *Neural Information Processing-Letters and Reviews*, 6(1):1–57, 2005.
- [26] Clauset, Aaron, Cosma Rohilla Shalizi, and Mark EJ Newman. “Power-law distributions in empirical data.” *SIAM Review*, 51(4):661–703, 2009.
- [27] D’Agostino, Ralph B and Michael A. Stephens. *Goodness-of-Fit Techniques*, vol. 68 of *Statistics, Textbooks and Monographs*. Marcel Dekker, Inc., 1986.
- [28] De Haan, Laurens and Ana Ferreira. *Extreme Value Theory: an Introduction*. Springer Science & Business Media, New York, New York, 2007.
- [29] Dillon, William R and Matthew Goldstein. *Multivariate Analysis: Methods and Applications*. John Wiley & Sons: New York, NY, 1984.
- [30] DIRS Laboratory. “DIRSIG Reference Docs.” Website, January 2017. URL <http://www.dirsig.org/docs/new/intro.html>.
- [31] Duda O., Richard, Peter Hart E., and David Stork G. *Pattern Classification*, vol. 2nd. John Wiley & Sons: New York, NY, 2000.
- [32] DuMouchel, William H and Richard A Olshen. “On the distributions of claims costs.” *Statistics Department Technical Report. University of Michigan, EE. UU*, 1974.
- [33] Ensafi, Eskandar and Alan D Stocker. “An adaptive CFAR algorithm for real-time hyperspectral target detection.” *Proc. SPIE*, vol. 6966, 696605. 2008.
- [34] Fauvel, Mathieu, Jón Atli Benediktsson, Jocelyn Chanussot, and Johannes R. Sveinsson. “Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles.” *IEEE Transactions on Geoscience and Remote Sensing*, 46(11):3804–3814, 2008.
- [35] Fauvel, Mathieu, Yuliya Tarabalka, Jón Atli Benediktsson, Jocelyn Chanussot, and James C. Tilton. “Advances in spectral-spatial classification of hyperspectral images.” *Proceedings of the IEEE*, 101(3):652–675, 2013.
- [36] Ferri, Cesar, José Hernández-Orallo, and Peter A Flach. “A coherent interpretation of AUC as a measure of aggregated classification performance.” *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 657–664. 2011.

- [37] Fisher, R. A. and L. H. C. Tippett. “Limiting forms of the frequency distribution of the largest or smallest member of a sample.” *Mathematical Proceedings of the Cambridge Philosophical Society*, 24(2):180–190, Apr 1928.
- [38] Funk, Christopher C, James Theiler, Dar A Roberts, and Christoph C Borel. “Clustering to improve matched filter detection of weak gas plumes in hyperspectral thermal imagery.” *IEEE Transactions on Geoscience and Remote Sensing*, 39(7):1410–1420, 2001.
- [39] Gnedenko, Boris. “Sur la distribution limite du terme maximum d’une serie aleatoire.” *Annals of Mathematics*, 43(3):423–453, July 1943.
- [40] Guida, Maurizio, Domenico Iovino, and Maurizio Longo. “Comparative performance analysis of some extrapolative estimators of probability tails.” *IEEE Journal on Selected Areas in Communications*, 6(1):76–84, Jan 1988, 0733-8716.
- [41] He, Yang, Ma Ben, Du Qian, and Zhang Liangpei. “Comparison of spectral-spatial classification for urban hyperspectral imagery with high resolution.” *2009 Joint Urban Remote Sensing Event*. 2009.
- [42] Heo, Jun-Haeng, Hongjoon Shin, Woosung Nam, Juseong Om, and Changsam Jeong. “Approximation of modified Anderson–Darling test statistics for extreme value distributions with unknown shape parameter.” *Journal of Hydrology*, 499:41–49, 2013.
- [43] Holmes, JD and WW Moriarty. “Application of the generalized Pareto distribution to extreme value analysis in wind engineering.” *Journal of Wind Engineering and Industrial Aerodynamics*, 83(1):1–10, 1999.
- [44] Hyvärinen, Aapo. “Fast and robust fixed-point algorithms for independent component analysis.” *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999.
- [45] Hyvärinen, Aapo. “Gaussian moments for noisy independent component analysis.” *IEEE Signal Processing Letters*, 6(6):145–147, 1999.
- [46] Jablonski, James. *Reconstruction Error and Principal Component Based Anomaly Detection in Hyperspectral Imagery*. Masters thesis, Air Force Institute of Technology, 2014.
- [47] Jablonski, James A, Trevor J Bihl, and Kenneth W Bauer. “Principal component reconstruction error for hyperspectral anomaly detection.” *IEEE Geoscience and Remote Sensing Letters*, 12(8):1725–1729, 2015.
- [48] Jackson, J. Edward and Robert H. Morris. “An Application of Multivariate Quality Control to Photographic Processing.” *Journal of the American Statistical Association*, 52(278):186–199, 1957.

- [49] Johnson, Robert J., Jason P. Williams, and Kenneth W. Bauer. "AutoGAD: An improved ICA-based hyperspectral anomaly detection algorithm." *IEEE Transactions on Geoscience and Remote Sensing*, 51(6):3492–3503, 6 2013.
- [50] Kaiser, Henry F. "The varimax criterion for analytic rotation in factor analysis." *Psychometrika*, 23(3):187–200, 1958.
- [51] Khodadadzadeh, Mahdi, Roozbeh Rajabi, and Hassan Ghassemian. "A Novel Approach for Spectral-Spatial Classification of Hyperspectral Data Based on SVM-MRF Method." *IEEE International Geoscience and Remote Sensing Symposium*, 1890–1893. IEEE, 2011.
- [52] Kohonen, Teuvo. "Learning vector quantization." *Self-Organizing Maps*, 203–217. Springer, 1997.
- [53] Kuiper, Nicolaas H. "Tests concerning random points on a circle." *Proc. Koninkl. Nederl. Akad. Van Wetenschappen, Series A*, vol. 63, 38–47. 1960.
- [54] Kwon, Heesung, Sandor Z Der, and Nasser M Nasrabadi. "Adaptive anomaly detection using subspace separation for hyperspectral imagery." *Optical Engineering*, 42(11):3342–3351, 2003.
- [55] Li, Jun, José M Bioucas-Dias, and Antonio Plaza. "Spectral-Spatial Hyperspectral Image Segmentation Using Subspace Multinomial Logistic Regression and Markov Random Fields." *IEEE Transactions on Geoscience and Remote Sensing*, 50(3):809–823, March 2012, 0196-2892.
- [56] Lin, Dennis KJ and Wanzhu Tu. "Dual response surface optimization." *Journal of Quality Technology*, 27(1):34–39, 1 1995.
- [57] Luceño, Alberto. "Fitting the generalized Pareto distribution to data using maximum goodness-of-fit estimators." *Computational Statistics & Data Analysis*, 51(2):904 – 917, 2006, 0167-9473.
- [58] Ludeman, Lonnie C. *Random Processes: Filtering, Estimation, and Detection*. John Wiley & Sons, Inc., 2003.
- [59] Mackay, Edward BL, Peter G Challenor, and AbuBakr S Bahaj. "A comparison of estimators for the generalised Pareto distribution." *Ocean Engineering*, 38(11):1338–1346, 2011, 0029-8018.
- [60] Mahalanobis, Prasanta Chandra. "On the generalised distance in statistics." *Proceedings National Institute of Science, India*, 1, 49–55. April 1936.
- [61] Manolakis, D, D Zhang, M Rossacci, R Lockwood, T Cooley, and J Jacobson. "Maintaining CFAR operation in hyperspectral target detection using extreme value distributions." *Proc. SPIE*, vol. 6565, 65651W. 2007.

- [62] Manolakis, Dimitris. “Hyperspectral image processing for automatic target detection applications.” *Lincoln Laboratory* , 14(1):79–116, 2003.
- [63] Manolakis, Dmtrios, R. Lockwood, T. Cooley, and J. Jacobson. “Is there a best Hyperspectral detection algorithm?.” *Proc. SPIE*, vol. 7334, 733402. 2009.
- [64] Matteoli, S., M. Diani, and J. Theiler. “An Overview of Background Modeling for Detection of Targets and Anomalies in Hyperspectral Remotely Sensed Imagery.” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(6):2317–2336, June 2014, 1939-1404.
- [65] Matthews, B.W. “Comparison of the predicted and observed secondary structure of T4 phage lysozyme.” *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442 – 451, 1975, 0005-2795.
- [66] Merkwirth, Christian, Ulrich Parlitz, and Werner Lauterborn. “Fast nearest-neighbor searching for nonlinear signal processing.” *Physical Review E*, 62(2):2089, 2000.
- [67] Messer, Adam J. *Contextual Detection of Anomalies within Hyperspectral Images*. Master’s thesis, Air Force Institute of Technology, March 2011.
- [68] Messer, Adam J and Kenneth W Bauer. “Method of sensitivity analysis in anomaly detection algorithms for hyperspectral images.” *Proc. SPIE*, vol. 10198: Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XXIII, 101980U. May 2017.
- [69] Mindrup, Frank M, Trevor J Bihl, and Kenneth W Bauer. “Modeling noise in a framework to optimize the detection of anomalies in hyperspectral imaging.” C.H. Dagli (editor), *Intelligent Engineering Systems through Artificial Neural Networks: Computational Intelligence in Architecting Complex Engineering Systems*, vol. 20, 517–524. 2010.
- [70] Minkler, Gary and J Minkler. *CFAR: The Principles of Automatic Radar Detection in Clutter*. Magellan Book Company, Baltimore, MD, 1990.
- [71] Murphy, Kevin P. *Machine Learning: a Probabilistic Perspective*. MIT Press: Cambridge, MA, 2012.
- [72] Myers, Raymond, Douglas Montgomery, and Christine Anderson-Cook. *Response Surface Methodolgy: Process and Product Optimization Using Design Experiments*, vol. 2nd. Wiley, 2009.
- [73] Ozturk, Aydin, Prakash R Chakravarthi, and Donald D Weiner. “On determining the radar threshold for non-Gaussian processes from experimental data.” *IEEE Trans. Inf. Theory*, 42(4):1310–1316, 1996.

- [74] Paciencia, Todd J. *Improving Non-Linear Approaches to Anomaly Detection, Class Separation, and Visualization*. Ph.D. thesis, Air Force Institute of Technology, Wright-Patterson AFB, OH, 2014.
- [75] Paciencia, Todd J. and Kenneth W. Bauer. “Hyperspectral anomaly detection using enhanced global factors.” *Proc. SPIE*, vol. 9844, 98440O. 2016.
- [76] Pelleg, Dan and Andrew W Moore. “X-means: Extending K-means with Efficient Estimation of the Number of Clusters..” *Proceedings of the Seventeenth International Conference on Machine Learning*, 727–734. 2000.
- [77] Pickands III, James. “Statistical inference using extreme order statistics.” *The Annals of Statistics*, 3(1):119–131, January 1975.
- [78] Pignatiello Jr., Joseph J. “An overview of the strategy and tactics of Taguchi.” *IIE Transactions*, 20(3):247–254, 1988.
- [79] Pimentel, Marco A F, David A Clifton, Lei Clifton, and Lionel Tarassenko. “A review of novelty detection.” *Signal Processing*, 99:215–249, 2014.
- [80] Qi, Ding and Eric D. Kolaczyk. “A Compressed PCA Subspace Method for Anomaly Detection in High-Dimensional Data.” *IEEE Transactions on Information Theory*, 59(11):7419–7433, Nov 2013, 0018-9448.
- [81] Reed, Irving S and Xiaoli Yu. “Adaptive multiple-band CFAR detection of an optical pattern with unknown spectral distribution.” *IEEE Trans. Acoust., Speech, Signal Process.*, 38(10):1760–1770, 1990.
- [82] Rickard, Lee J., Robert W. Basedow, Edward F. Zalewski, Peter R. Silvergate, and Mark Landers. “HYDICE: an airborne system for hyperspectral imaging.” *Proc. SPIE*, vol. 1937, 173–179. 1993.
- [83] Roberts, S. J. “Novelty detection using extreme value statistics.” *IEE Proceedings - Vision, Image and Signal Processing*, 146(3):124–129, June 1999, 1350-245X.
- [84] Roberts, S. J. “Extreme value statistics for novelty detection in biomedical data processing.” *IEE Proceedings - Science, Measurement and Technology*, 147(6):363–367, Nov 2000, 1350-2344.
- [85] Scarrott, Carl and Anna MacDonald. “A review of extreme value threshold estimation and uncertainty quantification.” *REVSTAT–Statistical Journal*, 10(1):33–60, 2012.
- [86] Schaum, Alan P. “Joint subspace detection of hyperspectral targets.” *Aerospace Conference, 2004. Proc. IEEE*, vol. 3, 1824 Vol.3. March 2004, 1095-323X.
- [87] Schaum, Alan P. “Advanced Methods of Multivariate Anomaly Detection.” *2007 IEEE Aerospace Conference*, 1–7. March 2007, 1095-323X.

- [88] Schweizer, Susan M. *The GMRF detector for hyperspectral imagery: an efficient fully-adaptive maximum likelihood detector*. Ph.D. thesis, Carnegie Mellon University, 1999.
- [89] Schweizer, Susan M. and Jos M. F. Moura. “Efficient detection in hyperspectral imagery.” *IEEE Transactions on Image Processing*, 10(4):584–597, 2001.
- [90] Shekhar, Shashi, Chang-Tien Lu, and Pusheng Zhang. “Detecting graph-based spatial outliers: algorithms and applications (a summary of results).” *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 371–376. ACM, 2001.
- [91] Sinclair, C.D., B.D. Spurr, and M.I. Ahmad. “Modified Anderson Darling test.” *Communications in Statistics - Theory and Methods*, 19(10):3677–3686, 1990.
- [92] Stein, David W. J., Scott G. Beaven, Lawrence E. Hoff, Edwin M. Winter, Alan P. Schaum, and Alan D. Stocker. “Anomaly detection from hyperspectral imagery.” *IEEE Signal Processing Magazine*, 19(1), 2002.
- [93] Sundaram, Srini, Iain GD Strachan, David A Clifton, Lionel Tarassenko, and Steve King. “Aircraft engine health monitoring using density modelling and extreme value statistics.” *Proceedings of The Sixth International Conference on Condition Monitoring and Machinery Failure Prevention Technologies*. The British Institute of Non-Destructive Testing, 2009.
- [94] Taguchi, Genichi. *Introduction to quality engineering: designing quality into products and processes*. Quality Resources, White Plains, NY, 1986.
- [95] Taguchi, Genichi. *The System of Experimental Design: Engineering Methods to Optimize Quality and Minimize Cost*. Kraus International Publications, White Plains, NY, 1987.
- [96] Taitano, Yuri P., Brian A. Geier, and Kenneth W. Bauer. “A Locally Adaptable Iterative RX Detector.” *EURASIP Journal on Advances in Signal Processing*, 2010(1):341908, 2010.
- [97] Tarabalka, Yuliya, Ja Benediktsson, and Jocelyn Chanussot. “Spectral-Spatial Classification of Hyperspectral Imagery Based on Partitional Clustering Techniques.” *IEEE Transactions on Geoscience and Remote Sensing*, 47(8):2973–2987, 2009.
- [98] Theiler, James. “By definition undefined: Adventures in anomaly (and anomalous change) detection.” *Proc. 6th IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*. 2014. URL <http://public.lanl.gov/jt/Papers/defundef.pdf>.

- [99] Theiler, James. “Anomalousness: How to Measure What You Can’t Define.” *Fourier Transform Spectroscopy and Hyperspectral Imaging and Sounding of the Environment*, JT1A.2. Optical Society of America, 2015.
- [100] Torralba, Antonio, Kevin P. Murphy, and William T. Freeman. “Contextual models for object detection using boosted random fields.” *Advances in Neural Information Processing Systems (NIPS)*, 17:1401–1408, 2005.
- [101] Voss, James, Luis Rademacher, and Mikhail Belkin. “Fast Algorithms for Gaussian Noise Invariant Independent Component Analysis.” *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, 1–9. 2015.
- [102] Wiener, Norbert. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*, vol. 2. MIT Press, 1949.
- [103] Williams, Jason P, Kenneth W Bauer, and Mark A Friend. “Further extensions to robust parameter design: three factor interactions with an application to hyperspectral imagery.” *International Journal of Quality Engineering and Technology*, 3(3):204–218, 2013.
- [104] Williams, Jason P, Trevor J Bihl, and Kenneth W Bauer. “Towards the mitigation of correlation effects in anomaly detection for hyperspectral imagery.” *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology*, 10(3):263–273, 2013.
- [105] Youden, William J. “Index for rating diagnostic tests.” *Cancer*, 3(1):32–35, 1950.
- [106] Yuan, Yuan, Qi Wang, and Guokang Zhu. “Fast hyperspectral anomaly detection via high-order 2-D crossing filter.” *IEEE Transactions on Geoscience and Remote Sensing*, 53(2):620–630, 2015.
- [107] Zortea, Maciel, Michaela De Martino, and Sebastiano Serpico. “A SVM ensemble approach for spectral-contextual classification of optical high spatial resolution imagery.” *2007 IEEE International Geoscience and Remote Sensing Symposium*, 1489–1492. July 2007, 2153-6996.

Vita

Major Adam Messer graduated from Heritage High School in Vancouver, WA in 2002. He then graduated from the United States Air Force Academy with a Bachelor of Science degree in Operations Research and was commissioned into the Air Force on 31 May 2006. Maj Messer served his first assignment as a Personnel Policy Analyst for the Headquarters Air Force, Personnel Policy Division. In August of 2009, he entered the Graduate School of Engineering and Management, Air Force Institute of Technology at Wright-Patterson AFB, Ohio. Upon completion of his Masters of Science degree in Operations Research, he was assigned to Headquarters, United States Air Force in Europe and Air Forces Africa (USAFE-AFAFRICA) Studies and Analysis Division (A9) as a Theater Analyst and subsequently the Deputy Chief Analyst. In August of 2012, he deployed to Prishtinë, Kosovo as a Plans Staff Officer in the NATO Headquarters Kosovo Forces (KFOR), Plans, Policy, & Assessment Division (J5). In August 2014, he was again assigned Air Force Institute of Technology at Wright-Patterson AFB, Ohio for Ph.D studies. After completion of his PhD in Operations Research he will be assigned as an assignment officer at the Air Force Personnel Center, Randolph AFB, Texas.

REPORT DOCUMENTATION PAGE					<i>Form Approved</i> <i>OMB No. 0704-0188</i>	
The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.						
1. REPORT DATE (DD-MM-YYYY) 14-09-2017		2. REPORT TYPE Doctoral Dissertation			3. DATES COVERED (From — To) Aug 2014 – Sep 2017	
4. TITLE AND SUBTITLE On the Development of Robust Anomaly Detection Algorithms with Limited Labeled Data				5a. CONTRACT NUMBER		
				5b. GRANT NUMBER		
				5c. PROGRAM ELEMENT NUMBER		
				5d. PROJECT NUMBER		
6. AUTHOR(S) Messer, Adam J., Major, USAF				5e. TASK NUMBER		
				5f. WORK UNIT NUMBER		
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Air Force Institute of Technology Graduate School of Engineering and Management (AFIT/EN) 2950 Hobson Way WPAFB, OH 45433-7765					8. PERFORMING ORGANIZATION REPORT NUMBER AFIT-ENS-DS-17-S-040	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) Headquarters Air Force Studies, Analyses, and Assessments Campaign Analyses Division Maj Todd Paciencia 1570 AF Pentagon Washington, D.C. 20330-1570 todd.j.paciencia.mil@mail.mil					10. SPONSOR/MONITOR'S ACRONYM(S) AF/A9FC	
					11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION / AVAILABILITY STATEMENT Distribution Statement A. Approved For Public Release; Distribution Unlimited.						
13. SUPPLEMENTARY NOTES This work is declared a work of the U.S. Government and is not subject to copyright protection in the United States.						
14. ABSTRACT Anomaly Detectors are used in a number of multivariate applications to identify data that do not belong within a dataset. A significant problem in the development of anomaly detection algorithms is the insufficient availability of truth-labeled data. This can lead to instances where the algorithms might not generalize to real-world operation scenarios. This research focuses on the sensitivity of anomaly detection algorithms to the anomalies present. First, a method to gradually modify the multivariate characteristics of the anomalies to approach the structure of the background is developed. In doing so, algorithms are tested with the expectation that gradual degradation in performance will occur as the anomalies become more difficult to differentiate from the background. This allows for a comparison between algorithms with both graphical and numerical techniques. The anomaly adaptation approach is extended into a robust parameter design framework to identify robust algorithm parameters that take into account the many possible anomalies present in extended operation scenarios. Additionally, a novel method for the selection of anomaly thresholds from possibly anomaly-contaminated datasets is presented. This method is shown to have superior performance on both simulated and real-world data. The method is also used to create new versions of a number of popular anomaly detectors in Hyperspectral Images.						
15. SUBJECT TERMS Anomaly detection, Sensitivity Analysis, hyperspectral, Robust Parameter Design, Anomaly Thresholding, Extreme Value Theory						
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON	
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr. Kenneth W. Bauer Jr., AFIT/ENS	
U	U	U	UU	145	19b. TELEPHONE NUMBER (include area code) (937) 255-3636 x4328 kenneth.bauer@afit.edu	