Journal of Electrical and Computer Engineering

Algorithms for Multispectral and Hyperspectral Image Analysis

Guest Editors: Heesung Kwon, Xiaofei Hu, James Theiler, Alina Zare, and Prudhvi Gurram

Algorithms for Multispectral and Hyperspectral Image Analysis

Journal of Electrical and Computer Engineering

Algorithms for Multispectral and Hyperspectral Image Analysis

Guest Editors: Heesung Kwon, Xiaofei Hu, James Theiler, Alina Zare, and Prudhvi Gurram

Editorial Board

The editorial board of the journal is organized into sections that correspond to the subject areas covered by the journal.

Circuits and Systems

Communications

Signal Processing

Contents

Hindawi Publishing Corporation
Journal of Electrical and Computer Engineering
Volume 2013, Article ID 908906, 2 pages
http://dx.doi.org/10.1155/2013/908906

Editorial

Algorithms for Multispectral and Hyperspectral Image Analysis

Heesung Kwon,[1] Xiaofei Hu,[2] James Theiler,[3] Alina Zare,[4] and Prudhvi Gurram[1]

[1] *Sensors and Electron Devices Directorate, US Army Research Laboratory, Adelphi, MD 20783, USA*
[2] *Department of Mathematics, Wake Forest University, Winston-Salem, NC 27106, USA*
[3] *Space Data Systems Group, Los Alamos National Laboratory, Los Alamos, NM 87545, USA*
[4] *Department of Electrical and Computer Engineering, University of Missouri, Columbia, MO 65211, USA*

Correspondence should be addressed to Heesung Kwon; heesung.kwon.civ@mail.mil

Received 28 November 2012; Accepted 28 November 2012

Recent advances in multispectral and hyperspectral sensing technologies coupled with rapid growth in computing power have led to new opportunities in remote sensing—higher spatial and/or spectral resolution over larger areas leads to more detailed and comprehensive land cover mapping and more sensitive target detection. However, these massive hyperspectral datasets provide new challenges as well. Accurate and timely processing of hyperspectral data in large volumes must be treated in a nonconventional way in order to drastically enhance data modeling and representation, learning and inference, physics-based analysis, computational complexity, and so forth. Current practical issues in processing multispectral and hyperspectral data include robust characterization of target and background signatures and scene characterization [1–3], joint exploitation of spatial and spectral features [4], background modeling for anomaly detection [5, 6], robust target detection techniques [7], low-dimensional representation, fusion of learning algorithms, the balance of statistical and physical modeling, and real-time computation [8, 9].

The aim of this special issue is to advance the capabilities of algorithms and analysis technologies for multispectral and hyperspectral imagery by addressing some of the above-mentioned critical issues. We have received many submissions and selected six papers after careful and rigorous peer review. The accepted papers cover a wide range of topics, such as anomaly detection, target detection and classification, dimensionality reduction and reconstruction, fusion of hyperspectral detection algorithms, and non-Gaussian mixture modeling for hyperspectral imagery. The brief summaries of the accepted papers are as follows.

The paper "*Hyperspectral anomaly detection: comparative evaluation in scenes with diverse complexity*," by D. Borghys et al., provides a comprehensive review of popular hyperspectral anomaly detection methods, an important problem in hyperspectral signal processing, including the global Reed-Xiaoli (RX) method, subspace methods, local methods, and segmentation based methods. The extensive performance analysis of these methods is presented in scenes with various backgrounds and different representative targets. The comparative results reveal the superiority of some detectors in certain scenes over other detectors.

The paper "*Non-Gaussian linear mixing models for hyperspectral images*," by P. Bajorski, addresses the problem of modeling hyperspectral data using non-Gaussian distribution. It is done by assuming a linear mixing model consisting of nonrandom-structured background and random noise terms. The nonvariable part of a hyperspectral image (structured background) can be assumed to be deterministic because of the strong presence of certain known materials. The variable noise term is modeled as two different multivariate distributions in the paper. The model is tested on two sets of hyperspectral data, one AVIRIS and one HyMap image, to determine which model best fits the data. Comprehensive results are provided along with a complete analysis of how researchers can verify how well a particular model fits a particular dataset. The significance of this paper lies in the fact that, often, in applications such as detection, classification, and synthetic data generation, a Gaussian distribution cannot be used to model hyperspectral data distributions, and other multivariate distributions are required instead.

The paper "*Randomized SVD methods in hyperspectral imaging*," by J. Zhang et al., addresses the problem of dimensionality reduction, compression, classification, and reconstruction of massive hyperspectral datasets by using a recently developed novel probabilistic approach called a randomized singular value decomposition (rSVD) technique. In rSVD, a large data matrix is iteratively approximated by random projection and factorized into low-dimensional matrices. In this paper, it was also demonstrated that fast computation in compression and reconstruction of large HSI datasets can be effectively achieved using the rSVDs approach.

The paper "*Evaluating subpixel target detection algorithms in hyperspectral imagery*," by Y. Cohen et al., considers algorithms for subpixel target detection and emphasizes the importance of good evaluation protocols for assessing those algorithms. The choice of algorithms (and, just as importantly, of parameters within a given algorithm) depends on image statistics, the target's spectral signature, and spatial size. By artificially emplacing simulated targets in a scene of interest, they are able to evaluate the effectiveness of different algorithms at detecting those targets and do so in a way that avoids the anecdotal statistics and inherent uncertainties that arise with real targets and real (which is to say, imperfect) ground truth. This work, in particular, extends the authors' previous work in the field by making the emplacement more realistic by incorporating "pixel phasing" effects (which occur when the target straddles two or more pixels) and image blurring. The authors use this approach to identify good detection algorithms for subpixel targets in the RIT Blind Test dataset [10] and demonstrate their efficacy by obtaining excellent scores on the blind test challenge. Although they do not claim to have found a universally optimal detector, their experiments consistently preferred a local ACE detector with a 3 × 3 moving window.

The paper "*Target detection using nonsingular approximations for a singular covariance matrix*," by N. Gorelik et al., introduces nonsingular matrix approximation techniques to improve the performance of the Reed-Xiaoli (RX) hyperspectral anomaly target detection approach, which normally involves singular covariance matrices. In this paper, the performance evaluation of the RX techniques based on these two nonsingular matrix approximations instead of a singular covariance matrix is presented. The experimental results characterize the pros and cons of the two methods in different scenes.

The last paper "*A semiparametric model for hyperspectral anomaly detection*," by D. Rosario, addresses the problem of anomaly detection in hyperspectral imagery. Because anomalies are by definition undefined, this is a problem that is fraught with pitfalls. Nonetheless, a scheme is developed that incorporates both physical intuition and mathematical sophistication and is applied to both the ubiquitous Forest Radiance dataset and some forward-looking imagery in the visible and near infrared. One of the innovations in the algorithm is a conversion of high-dimensional pixel descriptors to scalar values, based on angular distances to an appropriate centroid.

This special issue provides the reader with an overview of many (though certainly not all) of the current critical issues in multispectral and hyperspectral image analysis. We thank all the authors who responded to the call for papers, and we are especially grateful to the anonymous reviewers for their considerable time and tremendous effort in evaluating the manuscripts and providing invaluable comments to improve the quality of the papers in this special issue.

Heesung Kwon
Xiaofei Hu
James Theiler
Alina Zare
Prudhvi Gurram

References

[1] N. Keshava and N. J. F. Mustard, "Spectral unmixing," *IEEE Signal Processing Magazine*, vol. 19, no. 1, pp. 44–57, 2002.

[2] B. Somers, G. P. Asner, L. Tits, and P. Coppin, "Endmember variability in Spectral Mixture Analysis: a review," *Remote Sensing of Environment*, vol. 115, no. 7, pp. 1603–1616, 2011.

[3] J. M. Bioucas-Dias, A. Plaza, N. Dobigeon et al., "Hyperspectral unmixing overview: geometrical, statistical, and sparse regression-based approaches," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 2, pp. 354–379, 2012.

[4] G. Camps-Valls, L. Gomez-Chova, J. Muñoz-Marí, J. Vila-Francés, and J. Calpe-Maravilla, "Composite kernels for hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 3, no. 1, pp. 93–97, 2006.

[5] D. W. J. Stein, S. G. Beaven, L. E. Hoff, E. M. Winter, A. P. Schaum, and A. D. Stocker, "Anomaly detection from hyperspectral imagery," *IEEE Signal Processing Magazine*, vol. 19, no. 1, pp. 58–69, 2002.

[6] S. Matteoli, M. Diani, and G. Corsini, "A tutorial overview of anomaly detection in hyperspectral images," *IEEE Aerospace and Electronic Systems Magazine*, vol. 25, no. 7, pp. 5–28, 2010.

[7] D. G. Manolakis and G. Shaw, "Detection algorithms for hyperspectral imaging applications," *IEEE Signal Processing Magazine*, vol. 19, no. 1, pp. 29–43, 2002.

[8] D. A. Landgrebe, *Theory Methods in Multispectral Remote Sensing*, John Wiley & Sons, Hoboken, NJ, USA, 2003.

[9] M. Eismann, *Hyperspectral Remote Sensing*, vol. PM 210, SPIE Press Monograph, 2012.

[10] D. Snyder, J. Kerekes, I. Fairweather, R. Crabtree, J. Shive, and S. Hager, "Development of a web-based application to evaluate target finding algorithms," in *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS '08)*, vol. 2, pp. II915–II918, July 2008.

Hindawi Publishing Corporation
Journal of Electrical and Computer Engineering
Volume 2012, Article ID 425947, 30 pages
doi:10.1155/2012/425947

Research Article

A Semiparametric Model for Hyperspectral Anomaly Detection

Dalton Rosario

SEDD/Image Processing Division, Army Research Laboratory, 2800 Powder Mill Road, Adelphi, MD 20783, USA

Correspondence should be addressed to Dalton Rosario, dalton.s.rosario.civ@mail.mil

Received 21 April 2012; Revised 19 July 2012; Accepted 1 August 2012

Academic Editor: James Theiler

Using hyperspectral (HS) technology, this paper introduces an autonomous scene anomaly detection approach based on the asymptotic behavior of a semiparametric model under a multisample testing and minimum-order statistic scheme. Scene anomaly detection has a wide range of use in remote sensing applications, requiring no specific material signatures. Uniqueness of the approach includes the following: (i) only a small fraction of the HS cube is required to characterize the unknown clutter background, while existing global anomaly detectors require the entire cube; (ii) the utility of a semiparematric model, where underlying distributions of spectra are not assumed to be known but related through an exponential function; (iii) derivation of the asymptotic cumulative probability of the approach making mistakes, allowing the user some control of probabilistic errors. Results using real HS data are promising for autonomous manmade object detection in difficult natural clutter backgrounds from two viewing perspectives: nadir and forward looking.

1. Introduction

Hyperspectral (HS) sensors collect the radiation over a wide range of contiguous spectral bands, with each band corresponding to a unique spectral value. The field of view of the sensor is broken into hundreds of thousands of pixels, with each pixel representing from less than one to many squared meters of the region of interest depending on the spatial resolution of the sensor and the height of the sensor during the data collection. A collection of spatial-spectral images is put together resulting in an HS data cube, where the length and width represent the spatial dimension, and the depth represents the spectral dimension [1]. The resulting HS data cube consists of hundreds of thousands of pixels. Each pixel has tens or hundreds of data points, each point corresponding to a unique spectral value. In theory, the spectral signature of each pixel should uniquely characterize the physical material in that spatial land area. In practice, the recorded spectral signatures will never be identical for samples of the same material. Owing to the different illumination conditions, atmospheric effects, sensor noise, and so forth, the resulting spectral signatures for HS imagery pixels containing similar materials will exhibit spectral variability.

The discriminant capability, however, of spectral signatures has led to two major applications: object classification and target detection. The former aims to assign all pixels of the image to thematic classes, the latter searches the image for the presence of specific material (the target). As highlighted in [2], from a theoretical point of view, target detection can be considered as a binary classification problem, aiming at classifying the image into the target class and the background class. But, since targets are usually sparsely populated in the scene, while the background is abundant and heterogeneous, a major distinction between the approaches designed for target detection and classification is that target detectors cannot use criteria based on the minimization of classification error since that would lead to labeling every pixel as background. So, a typical solution proposed in the literature for target detection is to use the Neyman-Pearson approach, as discussed by Manolakis in [3], where maximizing the detection probability for a fixed false-alarm rate is the adopted criterion for the algorithm design.

Due to the availability of spectral signature libraries for a wide range of materials, spectral signature-based target detectors have been widely examined [3, 4]. These methods assume the target spectral signature is both reliable and known a priori and aim at finding highly correlated spectra in the scene corresponding to the reference target spectra; these methods are also known as target matching.

Target matching approaches are complicated by the large number of possible objects of interest and uncertainty as to the reflectance/emission spectra of these objects. For example, the surface of an object of interest may consist of several materials, and the spectra may be affected by weathering or other unknown factors. One may be interested in a large number of possible objects each with several signatures. Thus, the multiplicity of possible spectra associated with the objects of interest and the complications of atmospheric compensation, as well spectral calibration, acquisition geometry, and contamination from adjacent objects (see, for instance, the discussion in [5, 6]), have led to the development and application of anomaly detectors that seek to distinguish observations of unusual materials from typical background materials without reference to target signatures.

Anomalies are defined with reference to a model of the background. Typically, background models are developed adaptively using reference data from either a local neighborhood of the test pixel or a large section of the image. Local and global spectral anomalies are defined as observations that deviate in some way from the neighboring clutter background or the image-wide clutter background, respectively. Both approaches have their merits.

Local anomaly detectors are typically designed under the assumption that an anomalous material (the target) is spectrally distinct from local neighborhood spectra, which are assumed to be controlled by a known multivariate distribution (Gaussian); also, it is assumed that the scales of targets are known a priori, or the viewing perspective is assumed to be nadir and the altitude of the flying platform carrying the sensor is available for target scale estimation. This kind of detectors is susceptible to unknown spectral mixtures that are often obtained by sampling spectra through a moving window in the imagery, as the window is placed at a transition of distinct regions, forcing the neighboring spectral mixture to be compared against spectra of one of the regions in that transition; this may significantly increase the false alarm rate. The local anomaly detector is also susceptible to false alarms that are isolated spectral anomalies. For example, consider a scene containing isolated trees on a grass plain. Each separate tree may be detected as a local spectral anomaly even if the image contains a separate region with many pixels of trees. The most popular local anomaly detector in the HS research community is based on maximum likelihood estimation under the multivariate normal distribution; this detector is commonly known as the Reed-Xi (RX) algorithm [7] and has become a benchmark for comparison. Kwon and Nasrabadi proposed a *kernelized* version of RX in [8], and Matteoli et al. proposed a segmented version of RX in [9], for the estimation of the local background covariance matrix from global background statistics to cope with the small sample size problem in estimating covariance from local patches in the image; a small sample size problem occurs when the number of spectral observations is lesser than the number of spectral bands (see examination of this problem in [2]). Banerjee et al. in [10] leveraged the employment of kernel methods and the method known as support vector data description (SVDD) to propose the kernel SVDD. But, it is widely known that the performance of kernel methods crucially depends on the kernel function and its parameter(s) [11]. More recently, Gurram and Kwon in [12] and Khazai et al. in [13] have also addressed the parameter sensitivity of kernel-SVDD based detectors, which is still an open area of research. In the statistical based arena, Stein et al. in [14] presented an overview of the statistical anomaly detectors derived from three background models: local Gaussian model, Gaussian mixture model (GMM), and linear-mixing model. More recently, these models were compared against others approaches using the same dataset [15, 16]. In [15], the algorithms RX, GMM, and a cluster-based one were examined. Matteoli et al. in [16] extended the comparison to include the orthogonal subspace projection (OSP) detector and a deterministic signal subspace processing detector. Other classic approaches have also been adapted to the local anomaly detection problem, for example, Fisher's linear discriminant (see an implementation in [17]).

Global anomaly detectors are based on a simple universal distribution, which is designed to represent the background process in the whole image, thus, the name *global*. Example of these detectors is the GMM [14] or a different version of the RX algorithm, which estimates its required parameters (mean and covariance) not locally, but using spectra from the entire HS data cube; this version is informally known as the Global RX. By design, these methods are more resistant to the small sample size problem mentioned earlier, but they have limited ability to adapt to all nuances of the background class (sometimes referred to as an underfitting problem), which may result in both high false alarm and low anomaly detection rates. Other earlier versions of global detectors require that an HS data cube is first segmented into its constituent material classes, so detection is achieved by applying a cutoff threshold and automatically locating pixel clusters with pixel values above the threshold, representing the outliers of these classes. These hybrid algorithms vary in the method of segmentation, but also tend to use maximum likelihood detection under the multivariate normal distribution. The stochastic expectation maximization clustering algorithm by Stocker in [18] is a related example; see also Masson and Pieczynski in [19]. But, since segmentation results are known to be also sensitive to algorithms' parameters (see, for instance, [1]), utility of segmentation algorithms in the context of anomaly detection has not met expectations for varying real world scenarios.

Independently of the kind of anomaly detector in use, the following is a key consideration that should not be ignored: susceptibility to unknown spectral mixtures of unknown distributions often observed by sampling spectra through a moving window in the imagery, where the spectra in test belong to one of the components in that mixture (for instance, a local patch of canopy being compared against neighboring spectral mixture of the same canopy type and a patch of soil). Rosario in [20] examined this particular problem using near infrared HS imagery, where he showed that even on simple real case scenarios (e.g., motor vehicles parked at an open grassy field having also trees in the background on a sunny day) transitions of distinct regions may contribute to 18% of all of the locally observed spectra in the imagery, using a small moving window (10 by 10 pixels). Of course, with increased scene complexity (increased heterogeneity), this percentage reaches higher levels.

This paper presents a quasiglobal, semiparametric anomaly detection (QG-SemiP) approach that is less susceptible to problems and issues mentioned above for existing local or global anomaly detectors. In particular, the approach requires only a small fraction of the HS cube to characterize the unknown clutter background (hence, the term quasi-global), in contrast to existing global anomaly detectors, which require the entire cube. It does not use segmentation and it is less susceptible to spectral mixtures in local neighborhoods of the imagery.

The approach consists of three major parts: (i) a data dimension reduction method, which plays an important role on the overall approach, since it maps the data from their native multivariate space to a univariate domain in order to avoid the small sample size problem mentioned earlier, gaining in the process insensitiveness to illumination on objects, reducing the dominance of the blackbody response produced by Earth (if the longwave infrared region applies), among other benefits to be discussed later; (ii) a univariate semiparametric model [21], which is chosen because of its robustness to samples representing a mixture of material types; (iii) a parallel random sampling scheme that characterizes the unknown background clutter, using a binomial probability density function to model the likelihood of sampling targets by chance and erroneously labeling them as clutter, justifying multiple sampling trials in parallel in order to significantly decrease the labeling error probability.

The semiparametric model is neither parametric, since the specific distribution controlling the data is not assumed to be known, nor nonparametric, since other parameters must still be estimated relating two unknown distributions. The semiparametric model, however, assumes that the distributions of the samples to be tested are related to each other through an exponential function (a distortion), having two unknown parameters. As it will be shown later in detail, the model is appealing for many reasons, including the following: if two spectral samples under test belong to the same material type (i.e., they are not anomalous to each other), then the assumed exponential function relating both distributions is reduced to unity. If the two samples under test belong to different material types (i.e., they are anomalous to each other), then the exponential function will impose a significant weight relating both distributions, indicating that the samples are anomalous to each other; a key point here is that such an outcome is invariant to whether the assumption of exponential relationship between the distributions is satisfied or not—this will be discussed in more detail later. Finally, another benefit, although not recognized earlier by users of semiparametric models in other areas of study, is the model's natural ability to handle samples representing a mixture of different material types, which also will be shown later. As a note, samples representing a mixture of material types are known to significantly increase the false alarm rate in operational scenarios, requiring autonomous anomaly detection; they can produce dominant edges between spatial regions of different material classes, later to be detected as meaningless (false) anomalies [2].

The strength of the semiparametric model handling the mixture problem is attributed to the fact that a sample under test is expected to contribute to the estimation of the distribution function controlling the sample labeled as reference, where both samples are expected to *equally* contribute to the estimation (when both are in fact under the same distribution), only the reference sample will be able to contribute (when both samples are from different distributions), or both the reference sample and a portion of the test sample will contribute (when the reference is a mixture and the test sample represents a component in that mixture, or vice versa). These outcomes are produced naturally by the model because of the imposed exponential relationship between the two distributions, as shown later using simulated univariate data to make the point. Another appeal for using an exponential distortion assumption is that many of the known distribution functions can be expressed in terms of an exponential distortion of another distribution, including all of the exponential family of distributions [22].

Rosario in [23] published a much earlier and limited conference paper version of this work, where a two-step semiparametric detector (data transformation and semi-parametric test statistic) was introduced to the limited task of local anomaly detection (where prior knowledge of targets' scales was required, imposing the limitation) and its performance was compared only against the RX algorithm. Relative to the previous work in [23], this paper significantly extends both the overarching methodology and presents additional results using the extended approach to test significantly more challenging scenarios from the ground to ground viewing perspective, where targets' scales are unknown a priori. In other words, this work enables capability rather than just offers an incremental improvement. The specific contributions in this paper (method extension and additional results) are as follows.

(i) The two-step anomaly detector (data transformation and semiparametric test statistic) is employed for the first time in a quasi-global framework (which is also proposed herein), where only a fraction of the entire data cube being represented by blocks of data are randomly selected from the imagery and used as reference sets of spectra to test the entire imagery. The results are later fused using order statistics, as the sampling scheme is modeled by the Binomial distribution.

(ii) An analytical cutoff threshold is derived from the approach's asymptotic cumulative probability of rejecting a null hypothesis, when either the null or the alternative hypothesis is true (given that the null hypothesis is based on a multisample testing and order statistic scheme).

(iii) The extended approach is applied to additional real HS imagery to automatically find manmade objects in the scene, producing excellent results in difficult natural clutter scenarios viewed from both nadir and forward looking viewing perspectives.

(iv) Performance of the extended approach is compared via ROC curves against multiple methods found in the literature, for example, global methods (k-means, Gaussian mixture model, global RX), local methods (kernel RX, standard RX, Fisher's linear discriminant, and the local semiparametric detector discussed in [23]).

(v) An analytical study of the two-sample test framework, using the local semiparameter detector in [23] as the base detector.

(vi) An analytical study of the multi-sample fusion test framework, using the semiparametric model embedded in the quasi-global framework, which relaxes the prior knowledge requirement of target scales, hence, enabling scene anomaly detection tasks independently of the viewing perspective (nadir or forward looking).

(vii) A study of the semiparametric model's behavior in the presence of samples representing a mixture of two different material classes, which is the most common mixture case scenario given the sliding window sizes typically used in anomaly detection operational scenarios.

(viii) A subsection specially devoted to discuss the motivation and give a qualitative assessment of the data transformation used in the two-step semiparametric anomaly detection of [23].

For convenience, a list of notations is available after the appendix.

This paper is organized as follows: Section 2 describes spatial data window models for HS sensing, a semiparametric model, and a hypothesis test; Section 3 discusses the sampling method, its probabilistic model, and introduces QG-SemiP; Section 4 discusses performance of QG-SemiP testing nadir and forward looking HS imagery, consisting of manmade objects in difficult natural background scenarios; Section 5 concludes the paper.

2. Problem Formulation, Data Transformation, and Semiparametric Model Description

The main goal of anomaly detection algorithms testing incoming imagery is to detect objects that are spectrally anomalous to its surroundings, yielding in the process a tolerable number of *nuisance* detections. In many surveillance and reconnaissance applications, it is desired that manmade objects are detected as being anomalous to the surrounding natural clutter. Both format and model of the data play a significant role in attempting to achieve the intended goal.

2.1. Remote Sensing Data and Data Format. Experiment was carried out on data sets from two distinct sensors and viewing perspectives: (i) the hyperspectral digital imagery collection experiment (HYDICE) sensor, from a nadir looking perspective; (ii) the SOC-700 hyperspectral sensor, from a forward looking perspective. Data from these sensors will be referred to in this paper as nadir looking imagery and forward looking imagery, respectively.

The HYDICE sensor records 210 spectral bands in the visible to near infrared (VNIR: bands between 0.38 and 0.97 μm) and short-wave infrared (SWIR: bands between 1.0 and 2.50 μm). An extended description of this dataset can be found, for example, in Schweizer and Moura (2000). The results shown in this section for one data subcube are representative for other sub-cubes in the HYDICE (forest radiance) dataset. An illustrative subcube (shown as an average of 150 bands; 640×100 pixels) is depicted in Figure 1 (Cube 1, top). We discarded water absorption and low signal to noise ratio bands; the bands used are the 23rd–101st, 109th–136th, and 152nd–194th. The scene consists of 14 stationary motor vehicles (targets near a treeline) in the presence of natural background clutter (e.g., trees, dirt roads, grasses). Each target consists of about 7×4 pixels, and each pixel corresponds to an area of about 1.3×1.3 square meters at the given altitude.

The forward looking imagery used for this work was recorded using the SOC-700 VNIR HS spectral imager, which is commercially available off the shelf. The system produces HS data cubes of fixed dimensions $R = 640$ by $C = 640$ pixels by $K = 120$ spectral bands between 0.38 and 0.97 μm. Figure 1 (Cube 2 through 4, bottom) depicts examples of the forward looking imagery, where each pixel in any of the three cube examples corresponds to the average of 120 band values. Data cubes Cube 2 and Cube 3 were collected during the summer of 2004 in California, USA; Cube 4 was collected during the spring of 2008 in New Jersey, USA. From actual ground truth, it is known that the scene in Cube 2 (see Figure 1) contains three motor vehicles and a standing person in the center of the scene (i.e., two pick-up trucks to the left in proximity to each other, a man slightly forward from the vehicles in the center, and a sport utility to the right). Although the natural clutter in Cube 2 and Cube 3 is dominated by Californian valley-type trees and/or terrain at the same general geolocation, the data in Cube 3 depict a significantly more complex scenario. From actual ground truth, it is known that in Cube 3 a sport utility vehicle is inconspicuously deployed in the shades of a large cluster of trees. Portions of the shadowed vehicle can be observed near the center in Cube 3. Cube 4 was recorded in a wooded region at Picatinny, where (according to the available ground truth) a sport car is located behind multiple tree trunks, partially obscuring the vehicle; see Figure 1 (Cube 4).

The four data cubes in Figure 1 are independently displayed as intensity images after linear mapping the gray scale of each to the range 0–255. Pixel intensities shown in each individual surface is only relative to corresponding values in that surface; in other words, pixel values representing the same material (e.g., general terrain) may be displayed with different intensities in another surface. This fact explains,

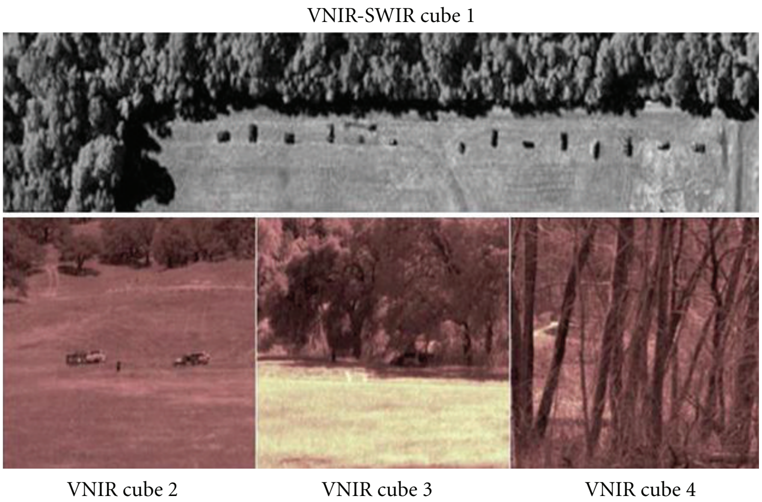


Figure 1: Examples of nadir looking imagery (Cube 1) and forward looking imagery (Cubes 2 through 4). An effective anomaly detection algorithm suite would allow a machine to automatically detect the presence of manmade objects (targets), while suppressing the cluttered environment, using no prior information about what constitutes clutter background or target in the imagery.

for instance, the difference in brightness between the same terrain under similar atmospheric conditions shown in Cube 2 and Cube 3. (The strong reflections from certain parts of the vehicles captured by the sensor in Cube 2 are not as dominant in Cube 3 because the vehicle in Cube 3 is in tree shades; the open field in Cube 3 is then the strongest reflector in the scene).

Next, we present a model of observed data using a sliding $n \times n$ window in $\mathbf{X}$ (a data cube). The data format of $\mathbf{X}$ is shown in (1), where r ($r = 1, \ldots, R$) and c ($c = 1, \ldots, C$) index pixels $\mathbf{x}_{rc}$ in the $R \times C$ spatial area $\mathbf{X}$, where $n \ll R$ and $n \ll C$. Pixels within a fixed $n \times n$ block of data in $\mathbf{X}$ are indexed from the upper left corner of this block using ij relative to rows and columns in $\mathbf{X}$, where $i = 1, \ldots, (R-n-1)$ and $j = 1, \ldots, (C-n-1)$. A representation of an $n \times n$ window at pixel location $(i, j) = (2, 2)$ in $\mathbf{X}$ is as follows:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_{11}, & \mathbf{x}_{12}, & \ldots, \mathbf{x}_{1C} \\ \begin{matrix} \mathbf{x}_{21}, \\ \mathbf{x}_{(2+1)1}, \\ \vdots \\ \mathbf{x}_{(2+n-1)1}, \end{matrix} & \overbrace{\begin{bmatrix} \mathbf{x}_{ij}, \mathbf{x}_{i(j+1)}, \ldots, \mathbf{x}_{i(j+n-1)} \\ \mathbf{x}_{(i+1)j}, \mathbf{x}_{(i+1)(j+1)}, \ldots, \mathbf{x}_{(i+1)(j+n-1)} \\ \vdots \\ \mathbf{x}_{(i+n-1)j}, \mathbf{x}_{(i+n-1)(j+1)}, \ldots, \mathbf{x}_{(i+n-1)(j+n-1)} \end{bmatrix}}^{n \times n \text{ window, where in this case } i=j=2} & \begin{matrix} , \ldots, \mathbf{x}_{2C} \\ , \ldots, \mathbf{x}_{(2+1)C} \\ \vdots\vdots \\ , \ldots, \mathbf{x}_{(2+n-1)C} \end{matrix} \\ \vdots\vdots\vdots & & \\ \mathbf{x}_{R1}, & \mathbf{x}_{R2}, & \ldots, \mathbf{x}_{RC} \end{bmatrix}. \tag{1}$$

Before pixels within a block of data can be used by a detector, they need to be rearranged to a sequence of multivariate variables. The rearrangement is made by concatenating individual rows in the $n \times n$ window in (1) as follows:

$$\begin{aligned} \mathbf{W}_1 = \big\lfloor & \mathbf{x}_{ij}, \ldots, \mathbf{x}_{i(j+n-1)}, \mathbf{x}_{(i+1)j}, \ldots, \mathbf{x}_{(i+1)(j+n-1)}, \ldots, \\ & \mathbf{x}_{(i+n-1)j}, \ldots, \mathbf{x}_{(i+n-1)(j+n-1)} \big\rfloor \\ = {} & [\mathbf{y}_{11}, \ldots, \mathbf{y}_{1n_1}], \end{aligned} \tag{2}$$

where $\mathbf{W}_1 \in \mathbf{R}^{K \times n_1}$, $n_1 = n^2$, and $\mathbf{y}_{1h} \in \mathbf{R}^K$ ($h = 1, \ldots, n_1$), such that $\mathbf{y}_{11} = \mathbf{x}_{ij}$, $\mathbf{y}_{12} = \mathbf{x}_{i(j+1)}$, and so forth until finally $\mathbf{y}_{1n_1} = \mathbf{x}_{(i+n-1)(j+n-1)}$. Since a window can be anywhere in $\mathbf{X}$ and $\mathbf{X}$ represents any HS data cube, $\{\mathbf{y}_{1h}\}_{h=1}^{n_1}$ are considered random vectors and the entire set of spectra that constitutes $\mathbf{X}$ will be observed through the $n \times n$ window.

Using the assumption that the random vectors in $\mathbf{W}_1$ are independent and identically distributed (i.i.d.), the distribution of data within the window, using (2), can be simplified to the following:

$$\mathbf{y}_{11}, \ldots, \mathbf{y}_{1n_1} \sim \text{i.i.d. } g_1(\mathbf{y} \mid \boldsymbol{\theta}), \tag{3}$$

where $g_1(\mathbf{y} \mid \boldsymbol{\theta})$ is a conditional multivariate probability density function (PDF) and $\boldsymbol{\theta}$ is its parameter set; both $g_1(\mathbf{y} \mid \boldsymbol{\theta})$ and $\boldsymbol{\theta}$ are typically unknown for real applications.

Normally, an anomaly detector requires two input sets of spectra ($\mathbf{W}_1 \in \mathbf{R}^{K\times n_1}$) and ($\mathbf{W}_0 \in \mathbf{R}^{K\times n_0}$) to perform its task on $\mathbf{X}$. The test sample ($\mathbf{W}_1$) is obtained at a fixed location ij in $\mathbf{X}$, as shown in (1) and (2); but, the source to obtain a reference sample ($\mathbf{W}_0$) will depend on the application, or viewing perspective.

For the nadir looking imagery, the most popular sampling method is to use pixel vectors surrounding a $n\times n$ block of data to construct $\mathbf{W}_0$, where $\mathbf{W}_1$ is constructed using the block of data to be tested. Notice that this sampling method is not suitable for the forward looking imagery because a priori knowledge of target scales in the imagery are required to set the size of a separation (guard) region between the block of data to be tested and locally surrounding samples.

For the forward looking imagery, the reference input set $\mathbf{W}_0$ could be made available from a spectral library, or be randomly selected from the testing data cube. In either case, $\mathbf{W}_0$ would be a rearranged version of a $n \times n$ block of data. The latter is addressed in Section 3, where (in order to make such a test useful for real applications) $\mathbf{W}_1$ is independently compared to multiple spectral sets $\mathbf{W}_0^{(f)} \in \mathbf{R}^{K\times n_0}$ ($f = 1,\ldots,N$); fusing thereafter these comparison results in order to yield a decision (output) surface, as described in Section 3.

Both input sets $\mathbf{W}_1$ and $\mathbf{W}_0$ feed the anomaly detector.

As mentioned earlier, whether the viewing perspective is nadir or forward looking, mixtures of different materials in $\mathbf{W}_1$ and/or in $\mathbf{W}_0$ can significantly degrade anomaly detectors' performances, as examined by Manolakis and Shaw in [2]. It is customary to assume normality in (3), or mixture of Gaussian distributions, but experience has shown (see [1]) that relaxation of these assumptions is needed.

We discuss next a two-step approach for anomaly detection, as introduced in [23], comprising of spectral transformation followed by the application of a univariate semiparametric model.

2.2. Data Transformation. This subsection discusses the employed method to transform spectra from their native multivariate space to a univariate domain, satisfying the univariate data requirement of the semiparametric model. We also provide justification for choosing the employed transformation and give some example cases to reinforce its use.

We consider a data transformation approach in two parts: (i) spectral differencing and (ii) angle mapping.

The rationale for (i) is twofold: (a) since HS samples are contiguous in the spectral domain (i.e., typical spectral resolution is of the order of 10 nanometers), more discriminant and independent information pertaining to a particular material type may be found between adjacent bands, which could augment the statistical power of detectors (this is specially the case in LWIR (longwave infrared) HS imagery where the radiance values observed in calibrated data (collected outdoor) are overwhelmingly influenced by the Planck's blackbody equation [1]), as the Earth's landscape (primarily, canopy and soil) behaves as a blackbody in the LWIR region of the electromagnetic spectrum (note: there is a whole topic of study in mathematical statistics on feature exploitation by *zero crossing*, which uses the output of random variable differencing as used herein for spectra, see [24]); (b) spectral differencing also puts significantly less weight in the importance of spectral magnitude (or bias) in anomaly detection applications, putting focus on the importance of spectral shape, instead. Spectral magnitude relates to the mean average of all measured radiance within a spectral sample, and spectral shape relates to the plotted curve of measured radiance as a function of wavelength. Existing classification and detection algorithms directly or indirectly exploit magnitude and/or shape of spectra in order to perform their tasks.

The benefit of (ii) is that it reduces the detection problem from a multivariate dimensional space to a univariate domain, avoiding the undesired problem of *singularity* issues during inverse estimations of covariance matrices. Singularity is known to occur when the sample size of a spectral sample is less than its number of spectral bands. Although there are approaches proposed in the literature to overcome this issue, the output of these approaches is not always desirable (see, for instance, [3]), since a typical HS sensor usually delivers between 120 and 1,000 bands, but targets may vary in number of pixels from as large as in the thousands to as small as 1 to 4 pixels, depending on the actual physical sizes of these targets and/or distance between the sensor and targets.

This paper is concerned about ensuring that the data transformation method can in fact reduce algorithm sensitivity to spectral magnitude (which can be achieved via (i)), so that a manmade object, for instance, deployed in tree shades can be considered as much as an anomaly relative to a dominant natural clutter background in the same way that the manmade object would have been if it were deployed, instead, out in the open field. All of this, while simultaneously preserving both a high sensitivity to spectral shape and the discriminant characteristics among spectra of distinct material types (which can be achieved via (ii)). If these requirements are matched, then the data transformation approach has achieved the overall desired goal—some example results are shown later in this subsection to reinforce those points.

Before those examples are presented, however, consider the following: let two spectra—having K spectral bands—be available for comparison, $\mathbf{y}_0 = [L_{01},\ldots,L_{0K}]$ and $\mathbf{y}_1 = [L_{11},\ldots,L_{1K}]$, where L_{ij} ($i = 0,1$; $j = 1,\ldots,K$) are nonnegative radiance values. Further assume that $\mathbf{y}_1$ is twenty percent stronger in magnitude than $\mathbf{y}_0$. One way to formalize the disparity in magnitude is to let $\{L_{0j} = \mu + \delta_{0j}\}_{j=1}^{K}$, $\{L_{1j} = 1.2\mu + \delta_{1j}\}_{j=1}^{K}$, and $\mu > 0$. Two key points are worth noticing: (a) the difference $L_{0(j+1)} - L_{0j} = (\mu + \delta_{0(j+1)}) - (\mu + \delta_{0j}) = \delta_{0(j+1)} - \delta_{0j}$ would provide more discriminant and independent information ($\delta_{0(j+1)} - \delta_{0j}$) than jointly using the highly correlated $L_{0(j+1)}$ and L_{0j}, ditto for $L_{1(j+1)} - L_{1j}$; (b) the difference $L_{1(j+1)} - L_{1j} = \delta_{1(j+1)} - \delta_{1j}$ would remove from consideration the 20 percent stronger average magnitude of L_{1j} over L_{0j}, if the $\{L_{0(j+1)} - L_{0j}\}_{j=1}^{K-1}$ were used instead for comparison against $\{L_{1(j+1)} - L_{1j}\}_{j=1}^{K-1}$.

In essence, (a) replaces the need for using—for instance—PCA (principal component analysis) to uncorrelate the data, as it is commonly employed in the HS community [2]; from (b), if $\mathbf{y}_1$ and $\mathbf{y}_0$ are observations of the same material under different illumination conditions (e.g., spectrum $\mathbf{y}_0$ representing a shaded object and spectrum $\mathbf{y}_1$ representing the same object but not shaded), then the average magnitude difference between the two spectra would not play a role in the comparison test, which is desired. For those readers who may have some concerns about what may be lost in the process of transforming the data from multivariate to univariate in the context of anomaly detection, as it will be shown shortly, the loss—although difficult to quantify—is not relevant to the anomaly detection problem, since an effective anomaly detector is not expected to distinguish a material type that is spectrally similar to another material type. If the detector is designed to be that sensitive, it would likely also produce an unacceptably high false alarm rate due to expected within class variability of the same material types dominating the scene (for instance, the expected within class variability of tree clusters across the scene).

The two-part transformation approach is described next.

Borrowing from the discussion in Section 2.1, the transformation approach requires two sets of spectra, a reference set ($\mathbf{W}_0$),

$$\begin{aligned}\mathbf{W}_0 &= [\mathbf{y}_{01}, \dots, \mathbf{y}_{0n_0}]\\ &= \begin{bmatrix} L_{011}, \dots, L_{01n_0} \\ \vdots \\ L_{0K1}, \dots, L_{0Kn_0} \end{bmatrix},\end{aligned} \tag{4}$$

where $\mathbf{y}_{0i} = [L_{01i}, \dots, L_{0Ki}]^t$ is calibrated spectra from a pixel-size location at the scene observed by the K-band sensor, during a particular set of atmospheric and illumination conditions; $(\cdot)^t$ is the vector transposed operator; L_{0ki} ($k = 1, \dots, K$) are radiance values, such that, adjacent radiance values are usually highly correlated; $i = 1, \dots, n_0$ and n_0 is the sample size of $\mathbf{W}_0$; and an independent test set ($\mathbf{W}_1$) that most likely has the same atmospheric condition captured in (4), but not necessarily the same illumination condition captured in (4),

$$\begin{aligned}\mathbf{W}_1 &= [\mathbf{y}_{11}, \dots, \mathbf{y}_{1n_1}]\\ &= \begin{bmatrix} L_{111}, \dots, L_{11n_1} \\ \vdots \\ L_{1K1}, \dots, L_{1Kn_1} \end{bmatrix},\end{aligned} \tag{5}$$

where all of the qualifying comments made for (4) also apply to $\mathbf{y}_{1i} = [L_{11i}, \dots, L_{1Ki}]^t$ with $i = 1, \dots, n_1$.

Letting u denote the index that distinguish both sets $\mathbf{W}_u$ ($u = 0, 1$), the magnitude of L_{uki} in (4) and (5) depends on the amount of illumination (e.g., shaded or nonshaded objects) and the illumination environment, this dependence can be significantly reduced by applying the first order differentiation—an approximation of the first derivative—to the columns of $\mathbf{W}_u$, or

$$\begin{aligned}\nabla_0 &= \begin{bmatrix} (L_{021} - L_{011}), \dots, (L_{02n_0} - L_{01n_0}) \\ \vdots \\ (L_{0K1} - L_{0(K-1)1}), \dots, (L_{0Kn_0} - L_{0(K-1)n_0}) \end{bmatrix},\\ \nabla_1 &= \begin{bmatrix} (L_{121} - L_{111}), \dots, (L_{12n_0} - L_{11n_1}) \\ \vdots \\ (L_{1K1} - L_{1(K-1)1}), \dots, (L_{1Kn_0} - L_{1(K-1)n_1}) \end{bmatrix}.\end{aligned} \tag{6}$$

Notice in (6) that $\nabla_0 \in \mathbf{R}^{(K-1)\times n_0}$ and $\nabla_1 \in \mathbf{R}^{(K-1)\times n_1}$, and the sample means of ∇_0 and ∇_1 are, respectively,

$$\begin{aligned}\overline{\nabla}_0 &= \frac{1}{n_0} \nabla_0 \mathbf{1}_{n_0\times 1},\\ \overline{\nabla}_1 &= \frac{1}{n_1} \nabla_1 \mathbf{1}_{n_1\times 1},\end{aligned} \tag{7}$$

where $\mathbf{1}_{d\times 1}$ is a column vector of dimension d filled with real values of 1's.

Denoting the columns of ∇_0 (which corresponds to the reference set) as $\{\nabla_{0i} \in \mathbf{R}^{(K-1)}\}_{i=1}^{n_0}$, then the multivariate reference and test samples can be transformed to univariate reference and test samples through the following angle-mapping formulas:

$$x_{0i} = \frac{180}{\pi} \arccos\left(\frac{\nabla_{0i}^t \overline{\nabla}_0}{\| \nabla_{0i} \| \left\| \overline{\Delta}_0 \right\|} \right), \tag{8}$$

$$x_{1i} = \frac{180}{\pi} \arccos\left(\frac{\nabla_{0i}^t \overline{\nabla}_1}{\| \nabla_{0i} \| \left\| \overline{\Delta}_1 \right\|} \right), \tag{9}$$

where $0^\circ \leq x_{0i} \leq 90^\circ$, $0^\circ \leq x_{1i} \leq 90^\circ$, the operator $\| \cdot \|$ using a column vector $\mathbf{x}$ is the square root of $\mathbf{x}^t\mathbf{x}$ (note: although we prefer to use a metric that yields a number having a geometric interpretation, the reader who is concerned about algorithm speed may replace the angle mapper metric with any other comparable metric of choice, for instance, the correlation metric [2] or the normalized dot product showed inside the arccos operator in (8) and (9). The most important aspect about the employed metric is that it must be able to preserve the discriminant characteristics among spectra of different material types, as it will be shown shortly).

From (8) and (9), the following two univariate sequences are constructed:

$$\begin{aligned}x_0 &= (x_{01}, x_{02}, \dots, x_{0n_0}),\\ x_1 &= (x_{11}, x_{12}, \dots, x_{1n_0}),\end{aligned} \tag{10}$$

where x_0 (reference) and x_1 (test) are the input sequences to be used by the univariate based anomaly detection technique discussed in Section 2.3. Note that both samples end up having the same sample size, n_0.

As mentioned earlier, the employed data transformation was specifically chosen to offer a number of desired properties, including reduced sensitivity to spectral magnitude

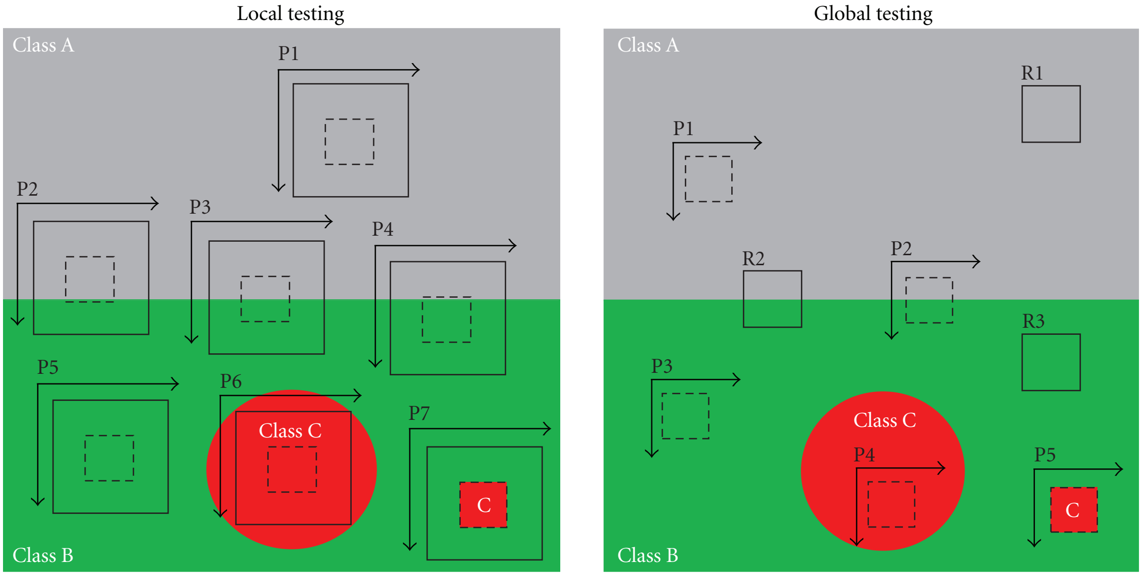


Figure 2: Most common window-based testing scenarios in anomaly detection problems, assuming for simplification that the scene background consists of two distinct material types (Class A and Class B) and a third material (Class C) also distinct from Class A and Class B depicts a anomalous material relative to the background.

and preservation of discriminant features among spectra of distinct material types. Regarding the latter, notice in (10) that, for the proposed transformation to work as advertised, when both multivariate samples $\mathbf{W}_0$ and $\mathbf{W}_1$ happen to be observations of the same material type, the component values in x_0 and x_1 are expected to be comparable and closer to 0° in the scale between 0° and 90°; however, when $\mathbf{W}_0$ and $\mathbf{W}_1$ are observations of distinct material types, the component values in x_0 and x_1 should be proportionally apart, where values in x_0 are expected to be closer to 0° while values in x_1 are closer to 90°. In addition, when the observation represents a mixture of two different material types, the proposed transformation should yield a univariate sample that is representative of the mixture.

Now, we will take a closer look at these expectations, using (8) and (9) to transform two sets of real HS spectra for a qualitative assessment, addressing first the most common sliding window-based testing scenarios naturally occurring in anomaly detection problems: *local testing*, which requires a priori knowledge of object scales (range dependent), and *global testing*, which does not require a priori knowledge of object scales (range invariant). Figure 2 depicts these scenarios (in this context, local testing means comparing clustered spectra against neighboring spectra, while global testing means comparing clustered spectra against spectra located elsewhere across the same imagery).

Figure 2 illustrates the same data-cube spatial representation under the two-test case scenarios, where for simplification the scene background is spatially dominated by only two distinct material types (Class A and Class B) and a third material (Class C—also distinct from Class A and Class B) illustrates the presence of an anomalous material relative to the background. Notice also that the two objects of Class C in the scene have significant size and shape differences, so that, the advantage of approaching the anomaly detection problem from a global rather than a local perspective can be pointed out.

The left hand side image in Figure 2 shows the overlaid sliding window locations of the standard approach to local anomaly detection in the HS research community (see [3]), where a sliding window consisting of an interior square (inner window) is concentrically located along with a larger square so that spectra observed through the inner window can be compared against spectra observed through the outer portion (outer window) of the larger square (i.e., the area of the larger square minus the inner window). Furthermore, the spectral set observed through the outer window is labeled as the *reference sample*, while the spectral set observed through the inner window is labeled as the *test sample*.

As the test window slides across the image one pixel location per algorithmic test, the labels P1 through P7 (left hand side image in Figure 2) highlight seven key test locations in the image. Table 1 summarizes a list of *plausible* anomaly declarations by an anomaly detector versus the *desired* declarations, using the known ground truth information about the scene.

The last column of Table 1 (desired anomaly) shows that out of the seven most distinct window locations for local testing, only two (P6 and P7) are desired, which may contradict declarations made by anomaly detection models currently found in the literature. In fairness, these detectors would be performing the job they are employed to do, as the plausible anomaly declarations shown in Table 1 are indeed correct in the strict sense. For instance, P2 shows a test between observations from Class A (test sample) and observations from a mixture (reference sample: Class A and Class B), while P6 shows a test between observations from

Table 1: Plausible versus desired declarations of *local anomalies*.

Window location	Plausible anomaly	Desired anomaly
P1	No	No
P2	Yes	No
P3	No	No
P4	Yes	No
P5	No	No
P6	No	Yes
P7	Yes	Yes

Table 2: Plausible versus desired declarations of *global anomalies* (Using R1, R2, and R3 as reference samples, see Figure 2).

Window location	Plausible anomaly	Desired anomaly
P1	No (R1)	No (R1)
	Yes (R2)	No (R2)
	Yes (R3)	Yes (R3)
P2	Yes (R1)	No (R1)
	No (R2)	No (R2)
	Yes (R3)	No (R3)
P3	Yes (R1)	Yes (R1)
	Yes (R2)	No (R2)
	No (R3)	No (R3)
P4	Yes (R1)	Yes (R1)
	Yes (R2)	Yes (R2)
	Yes (R3)	Yes (R3)
P5	Yes (R1)	Yes (R1)
	Yes (R2)	Yes (R2)
	Yes (R3)	Yes (R3)

the same material type (the intended anomaly: Class C). The reason, however, these declarations are not desired is that locations similar to P2 will likely accentuate edges as anomalies across the image, increasing the probability of false alarms, and locations similar to P6 will suppress the intended anomaly, decreasing the probability of correct detections. Location P6 also emphasizes the lack of robustness using the local testing approach to find anomalies in the scene, as a priori knowledge of object scales (consisting of the anomalous material) are not always available.

Regarding global testing shown in Figure 2 (the illustration at the right hand side), the window locations denoted as P1 through P5 depict distinct testing locations (observed test samples), while locations denoted as R1 through R3 depict distinct observations (fixed reference samples) that may have been randomly sampled from the image (prior to testing) and used to test every possible window-sized regions across the image, including P1 through P5. Table 2 shows the plausible declarations of anomalies versus the desired declarations, using the same ground truth information about the scene.

The last column of Table 2 (desired anomaly) in essence shows that any test that involves a mixture of classes and a component of that mixture should not be declared as an anomaly so that only truly anomalous material types (in this case Class C) relative to the background will be accentuated. An implementation scheme for the global testing approach, which requires fusion of declarations, will be discussed later. For now consider the following: the final declaration for any given window location is to retain the declaration NO, if it is there, out of all of the results produced by the particular testing location. Using this rule, locations P1, P2, and P3 would produce a final declaration of NO, and P4 and P5 would produce a final declaration of YES, as it would be desired by a global detection scheme. Notice also that P4 ensures that the circular object (Class C) would be accentuated as an anomaly, using the same test window size that would have detected the other anomaly of different scale shown in location P5; this is also desired.

Using real spectral samples, we will now qualitatively demonstrate the behavior of transformation output equations (10), when exposed to the key window positions shown in Figure 2. In anticipation, we would like to see that the data transformation will preserve distinction between samples of two different classes and produce results corresponding to a mixture of classes, when such a mixture is being observed. If successful, those results would give us some level of confidence that not much is being lost in terms of class distinction and that examples of mixtures would be shown as mixtures by the data transformation (demonstration of the desired anomaly detection declarations will be shown later, when the semiparametric model is discussed in Section 2.3. We also defer answering questions about what would happen when other possible window location cases appear, besides the ones shown in Figure 2, until discussion of results from testing real HS imagery in Section 4).

Figure 3 shows spectral transformation results using two sets of real spectra (Class A and Class B), where in this case spectral band differencing was not used as input for angle estimation among spectra and spectral means (see (6)), the actual individual radiance values (L's) were used instead of the difference between adjacent radiance values (L's). Figure 4 shows results using the same two sets of spectra but this time around using band differencing, following exactly the path from (6) to (10).

Class A in Figure 3 consists of 200 real spectra representing a grassy area in an open field; Class B consists of 200 real spectra representing a cluster of tree leaves in the same geographic location of Class A. The employed HS sensor operates in the VNIR (visible to near infrared) region, producing 120 spectral bands per spectrum. To observe the behavior of the proposed data transformation as it processes spectra equivalent to the window location examples shown in Figure 2, Class A is denoted as the reference sample and processed with an independent test set representing Class A (200 spectra), it was also processed with a second test set representing Class B (200 spectra). The spectral sample means of both spectral sets are also shown in Figure 3.

Using both the reference and test samples from the same class (Class A) exemplifies window locations P1, P3, P5,

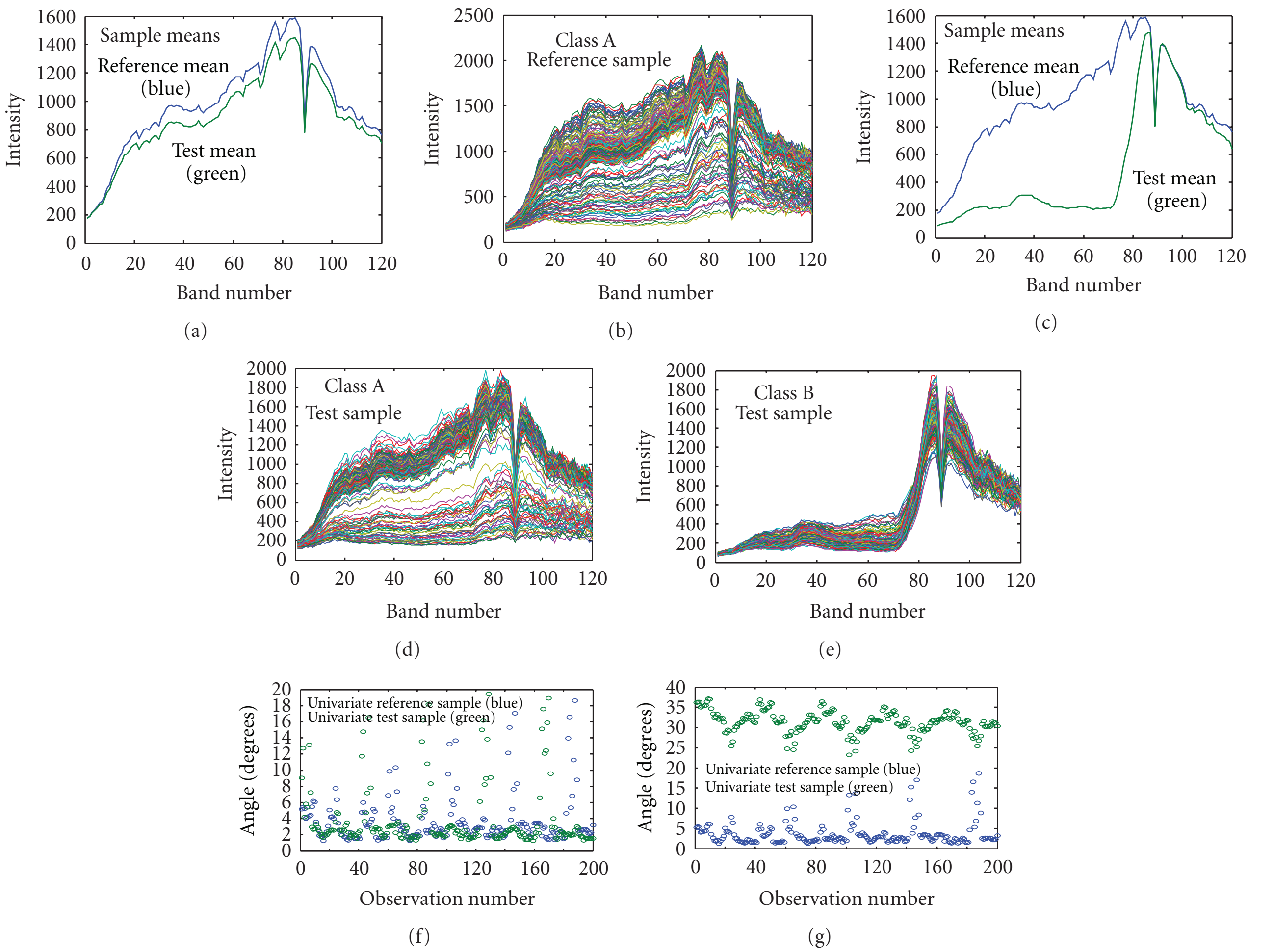


Figure 3: Spectral transformation from multivariate to univariate sample using the actual spectral values instead of spectral differencing.

and P6 in the local testing image in Figure 2; for the global testing illustration in Figure 2, it also represents the cases when spectra from R1 are processed with spectra from P1 and when spectra from R3 are processed with spectra from P3. Using spectra Class A as the reference and spectra from Class B as the test exemplifies the local testing locations P7 (i.e., the presence of an anomaly) and global testing location duals (R1, P4), (R1, P5), (R3, P4), and (R3, P5). The key point to notice in Figure 3 is the angle mapped plots shown in the bottom portion of the figure, where the plot at the left hand side shows that both the univariate reference sample (blue bubbles) and the univariate test sample (green bubbles) are comparable, preserving the fact that both samples belong to the same class. On the other hand, the angle mapped plot shown on the right hand side shows that the spectral transformation preserves the distinction between the samples from Class A and Class B. Both results are desired and would be passed as input to the employed detector for a decision.

The same experiment was held, but instead differentiating data in the spectral direction in order to check whether such a step would change the desired outcome from the spectral transformation shown in Figure 3. Figures 4(a) and 4(b) shows the band differencing means using (7) choosing samples from Class A to be both the reference and test sets (Figure 4(a)), and choosing a sample from Class A as reference and a sample of Class B as the test (Figure 4(b)).

The angle plots shown in Figures 4 (c) and 4(d) affirm that the differentiation step shown in (6) and (7), followed by output transformation from multivariate to univariate data, do preserve the lack of distinction between samples from the same class (Figure 4(c)) and a strong distinction between samples from Class A (reference) and samples from Class B (test), see plots in (Figure 4(d)) . This example provides a direct assertion that what is lost due to the transformation is not relevant to the anomaly detection problem, since the goal of an anomaly detector is to find those material types that are truly distinct from the material types spatially dominating the background scene, as other factors will conspire against the detector's effectiveness (e.g., mixture of distinct material classes, within material class variability). In other words, if a specific material type (target) is desired by the user to be considered as an anomaly relative to a natural environment but the target is not sufficiently distinct spectrally from

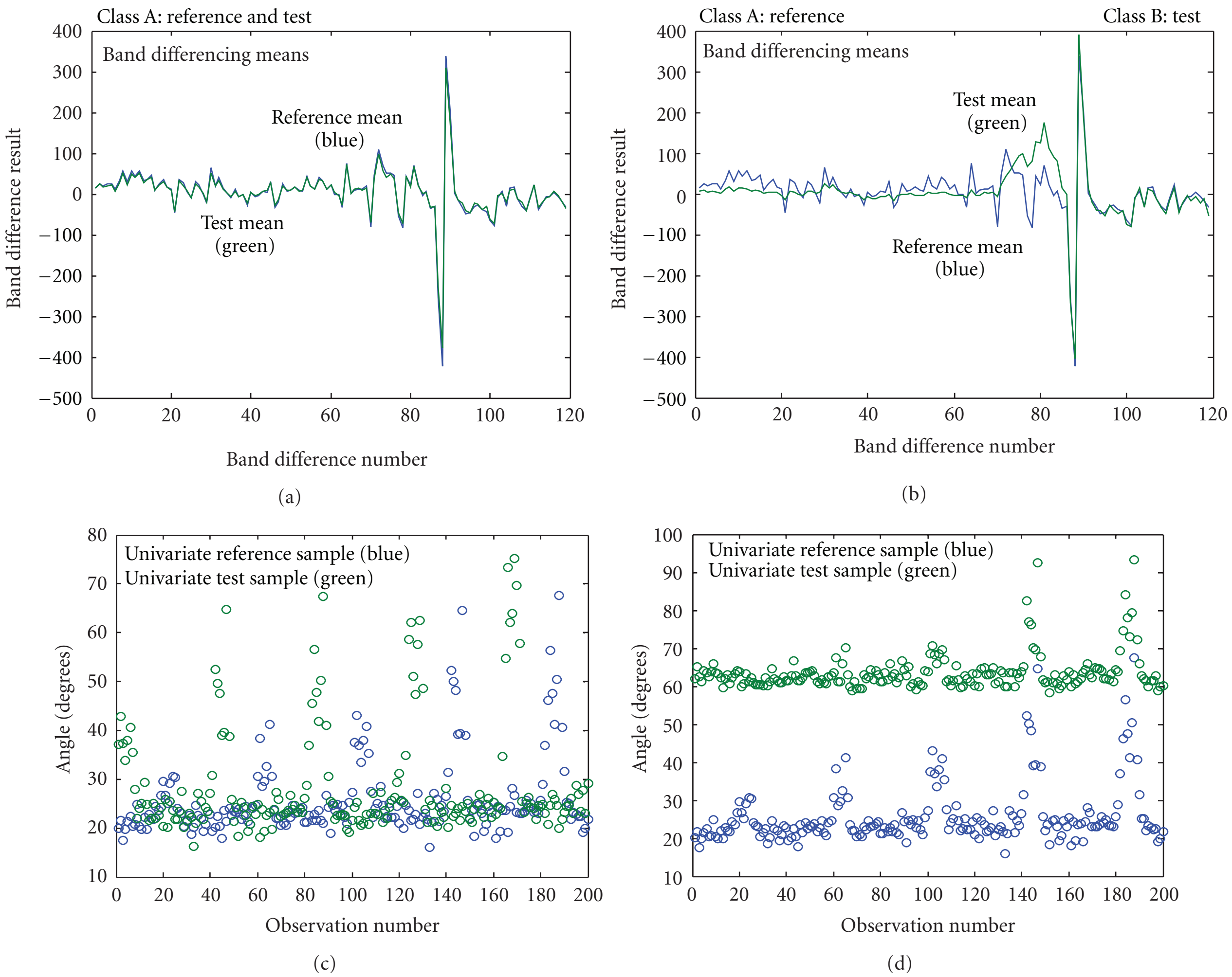


Figure 4: Spectral transformation from a multivariate to univariate sample using spectral differencing.

the background, then an effective anomaly detector would most likely not declare the target as an anomaly. Ironically, as mentioned earlier, this is a virtue because a detector that is too sensitive in distinguishing two spectrally similar material types is also likely to produce an unacceptably high number of false alarms due to expected within class variability of dominant background material types in the scene.

Next, we would like to check how a spectral set, representing a mixture, is altered by the data transformation. For this demonstration, we constructed a reference sample by combining 100 spectra of Class A with 100 spectra of Class B, so that, the reference sample represented a mixture of two classes, and arbitrarily chose the test set to be represented by 200 spectra of Class B (the latter were independent from the ones used to construct the reference mixture). Figures 5(a) and 5(b) show both the mixture (reference construct) and the component of that mixture (Class B: test sample), and the resulting angle mapped plots using spectra without band differencing (Figure 5(c)) and spectra after band differencing (Figure 5(d)).

The key message from Figures 5(c) and 5(d) is that the data transformation with or without the band differencing step does preserve in the univariate domain the fact that the material class of the multivariate test spectra is a component of the mixture of classes represented by the multivariate reference set of spectra. In this particular case, the univariate reference sample (blue bubbles) clearly shows the presence of two classes (i.e., half of the observations has lower angle values and the other half has higher angle values), while the test univariate sample (green bubbles) shows observation values commensurate to the one of the mixture class component (lower angle values). Although this fact is not necessarily desired for anomaly detection (see, for instance, P2 and P4 in Figure 2 (local testing, left image), the data transformation at least does not seem to alter such a scenario involving a mixture, which is fine as long as the employed anomaly detector is designed to handle similar challenging cases.

In summary, the proposed data transformation does preserve distinctions or similarities that exist between spectral sets and also offers additional benefits, as highlighted earlier in this subsection. What is needed now is a model that is more effective finding anomalies, while simultaneously managing the expected negative-impact nuisances naturally occurring in real operational scenarios (e.g., the presence

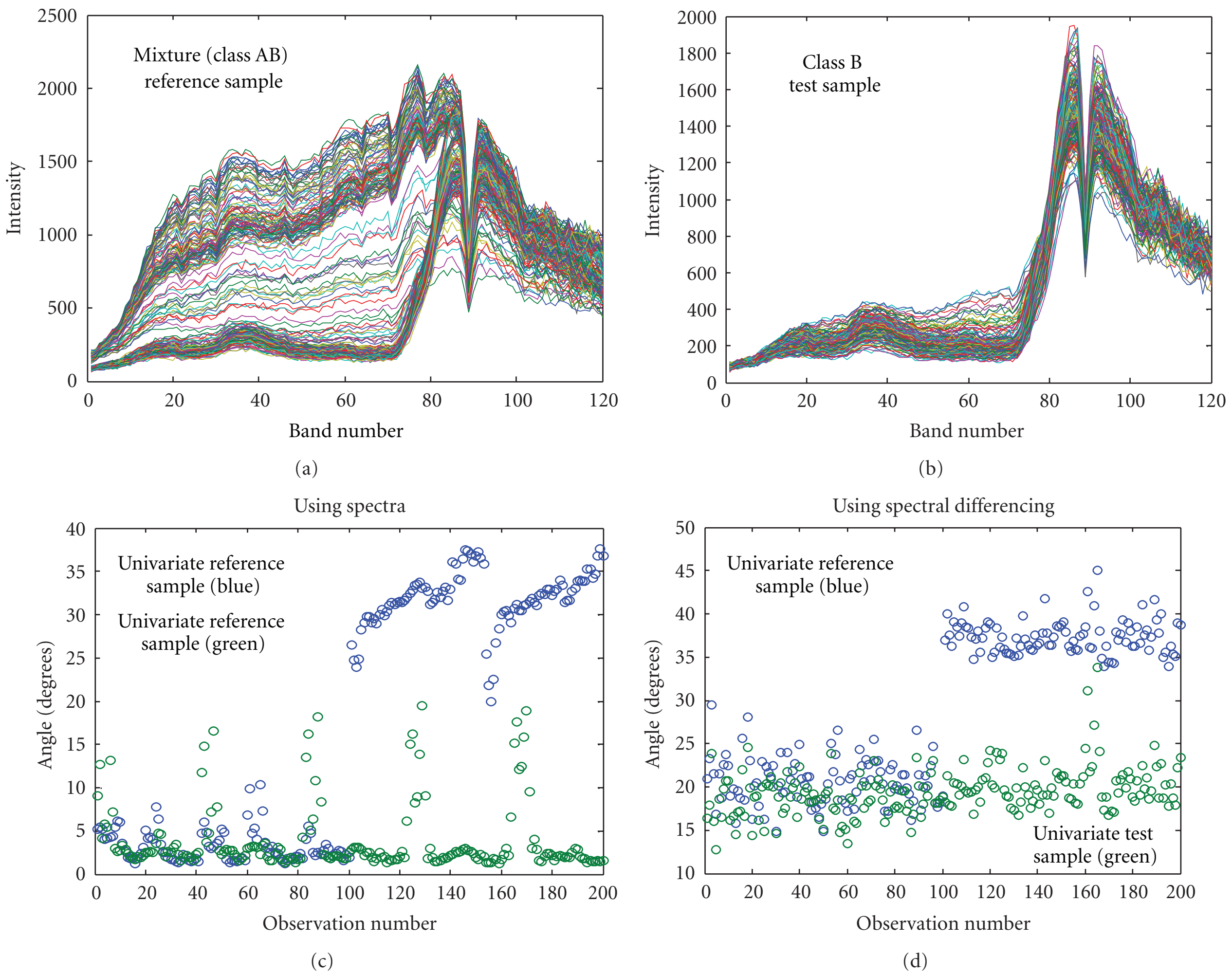


Figure 5: Spectral transformation from a multivariate to univariate sample for a mixture and a component of that mixture.

of mixtures). A semiparametric model is described next for that purpose.

2.3. Univariate Semiparametric Model. Let two multivariate samples $\mathbf{W}_1$ and $\mathbf{W}_0$ be transformed to $x_0 = (x_{01}, x_{02}, \ldots, x_{0n_0}) \sim f_0(x)$ and $x_1 = (x_{11}, x_{12}, \ldots, x_{1n_1}) \sim f_1(x)$, respectively (using, for instance, (4) through (9)), where $f_0(x)$ and $f_1(x)$ are unknown joint PDFs.

To simplify the anomaly detection problem using the transformed data, suppose the two random samples (real valued, not vector valued) x_0 (reference) and x_1 (test) are independent, consisting of i.i.d. (see mathematical notation list after the appendix) random variables controlled by unknown marginal PDFs g_0 and g_1, respectively. Moreover, let g_0 and g_1 be related through the following model:

$$\begin{aligned} x_1 &= (x_{11}, \ldots, x_{1n_1}) \text{ i.i.d} \sim g_1(x) \\ x_0 &= (x_{01}, \ldots, x_{0n_0}) \text{ i.i.d} \sim g_0(x), \end{aligned} \tag{11}$$

$$\frac{g_1(x)}{g_0(x)} = \exp(\alpha + \beta x). \tag{12}$$

The model in (11)-(12) is appealing for many reasons, consider the following examples.

If x_0 and x_1 are samples of the same distribution, then the assumed exponential relationship is reduced to *unity* so that $g_1 = g_0$ (whether g_1 or g_0 is known or not), indicating that x_0 is not anomalous to x_1. If x_0 and x_1 are samples of different distributions, then the exponential function will impose a significant weight relating both distributions, indicating that x_1 and x_0 are anomalous to each other. The key point here is that the latter outcome is invariant to and independent of whether the assumption of exponential relationship between the distributions is satisfied or not; that is, if the exponential relationship assumption is satisfied, then x_1 is anomalous to x_0, but if this assumption is not satisfied, x_1 is still anomalous to x_0. Since the application of interest only requires a determination of whether x_1 is anomalous to x_0, satisfying the imposed assumption of (12) is inconsequential. So, a hypothesis test could just be designed to check whether $\exp(\alpha + \beta x) = 1$ in order to determine the presence of anomalies.

However, the relationship assumption in (12) plays a major role in the mathematical development of the statistical test and, more importantly for the application in context, it also plays an important role in determining whether a portion or the entire test sample (x_1) can contribute to the estimation of the reference distribution (g_0). The latter is a subtle feature never before recognized in other areas of study, for example, pharmaceutical [22], where semiparametric models are more commonly employed for their utility. The implication of this subtlety is that when one of the samples represents a mixture of different material types and the other represents a component of that mixture, the model in (11)-(12) allows both samples, through the assumed relationship, to contribute to the distribution estimation of the chosen reference sample; in essence stating that the assumption may also be partially satisfied as long as a portion of the test sample belongs to the reference distribution. This is manifested when the test produces an estimation of $\exp(\alpha + \beta x)$ that is significantly closer to unity than it would produce when the test sample has absolutely no relationship with the reference sample. In the practical scenario this is the difference between having a detector capable or not of naturally handling samples representing a mixture, as it will be shown later in this subsection. As mentioned earlier, samples representing a mixture of material types are known to significantly increase the false alarm rate in operational scenarios for autonomous anomaly detection—they can produce dominant edges between regions of different material classes, later to be detected as false alarms [2].

Notice in (12) that, since g_1 is a density, $\beta = 0$ must imply that $\alpha = 0$, as α merely functions as a normalizing parameter, following from the requirement that a PDF (in this case g_1) must integrate to unity, see PDF properties in [25]. Notice also that if $\beta = 0$ then x_0 and x_1 must belong to the same population (i.e., $g_1 = g_0$). Using this fact, a test statistic can be designed to test the following hypotheses:

$$\begin{aligned} &H_0 : \beta = 0 \quad (g_1 = g_0) \quad \text{anomaly absent}, \\ &H_1 : \beta \neq 0 \quad (g_1 \neq g_0) \quad \text{anomaly present}. \end{aligned} \tag{13}$$

In order to estimate β, denote the union of x_0 and x_1 (combined data) by t,

$$t = (x_{11}, \ldots, x_{1n_1}, x_{01}, \ldots, x_{0n_0}) \equiv (t_1, \ldots, t_n), \tag{14}$$

and following the construction by Qin and Zhang in [21] and Fokianos et al. in [22], a maximum likelihood estimator of $G_0(x)$—the continuous cumulative distribution function (CDF) corresponding to the reference $g_0(x)$, can be obtained by maximizing the likelihood over the class of step CDF's with jumps at the observed values $t_1, \ldots, t_n$. Accordingly, if $\tilde{g}_0(t_i) = dG(t_i)$, where $d(\cdot)$ denotes the differentiation operator, $i = 1, \ldots, n$, the likelihood becomes,

$$\begin{aligned} \zeta(\alpha, \beta, \tilde{g}_0) &= \prod_{i=1}^{n_0} \tilde{g}_0(x_{0i}) \prod_{j=1}^{n_1} \exp\left(\alpha + \beta x_{1j}\right) \tilde{g}_0\left(x_{1j}\right) \\ &= \prod_{i=1}^{n=n_1+n_0} \tilde{g}_0(t_i) \prod_{j=1}^{n_1} \exp\left(\alpha + \beta x_{1j}\right). \end{aligned} \tag{15}$$

One can now express each $\tilde{g}_0(t_i)$ in terms of α and β and then substitute the terms $\tilde{g}_0(t_i)$ back into the likelihood to produce a function of α and β only. When α and β are fixed, (15) is maximized by maximizing only the product term $\prod_{i=1}^{n} \tilde{g}_0(t_i)$, subject to the constraints

$$\sum_{i=1}^{n} \tilde{g}_0(t_i) = 1, \qquad \sum_{i=1}^{n} \tilde{g}_0(t_i)[\exp(\alpha + \beta t_i) - 1] = 0. \tag{16}$$

Denoting $\rho = n_1/n_0$, Qin and Zhang in [21] showed that

$$\tilde{g}_0(t_i) = \frac{1}{n_0} \frac{1}{1 + \rho \exp(\alpha + \beta t_i)}, \tag{17}$$

and that the value of the profile log-likelihood $\log[\zeta(\alpha, \beta, \tilde{g}_0]$ up to a constant can be expressed as a function of α and β only, or

$$l(\alpha, \beta) = \sum_{j=1}^{n_1} (\alpha + \beta x_{1i}) - \sum_{i=1}^{n} \log[1 + \rho \exp(\alpha + \beta t_i)]. \tag{18}$$

A system of score equations that maximizes (18) over (α, β) is shown below [21],

$$\begin{aligned} \frac{\partial l(\alpha, \beta)}{\partial \alpha} &= -\sum_{i=1}^{n} \frac{\rho \exp[\alpha + \beta t_i]}{1 + \rho \exp[\alpha + \beta t_i]} + n_1 = 0, \\ \frac{\partial l(\alpha, \beta)}{\partial \beta} &= -\sum_{i=1}^{n} \frac{t_i \rho \exp(\alpha + \beta t_i)}{1 + \rho \exp(\alpha + \beta t_i)} + \sum_{j=1}^{n_1} x_{1j} = 0. \end{aligned} \tag{19}$$

The solution of the score equations yields the maximum likelihood estimators $(\hat{\alpha}, \hat{\beta})$ and consequently by substitution also yields an estimator of $\tilde{g}_0(t_i)$, or [21]

$$\hat{\tilde{g}}_0(t_i) = \frac{1}{n_0} \frac{1}{1 + \rho \exp\left(\hat{\alpha} + \hat{\beta} t_i\right)}. \tag{20}$$

So, in summary, by using profiling, an estimator (20) for $\tilde{g}_0(t_i)$ is attained in addition to score equations (19), where both the reference and test samples as shown in (14) are used to estimate g_0 (the reference PDF). This is only possible because the model in (12) implies that g_1 can be expressed in terms of g_0. This feature allows this model to be robust when either g_0 or g_1 is bimodal or multimodal (representing a sample mixture) while the other represents a component of the same mixture—a key factor for handling transitions of distinct regions in the anomaly detection application, as it will be shown later.

Using results from Fokianos et al. in [22], the estimator $\hat{\beta}$ has the normal asymptotic behavior, as the sample size tends to infinity ($n \to \infty$), or

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \xrightarrow[n \to \infty]{} N\left(0, \frac{\rho^{-1}(1 + \rho)^2}{\nu^2}\right), \tag{21}$$

where β_0 denotes the true parameter, ν^2 is the variance (a scalor) using components from the combined sequence t, $\rho = n_1/n_0$, $n = n_1 + n_0$, and → means *converges to*—in this case to a normal distribution having 0 mean and variance equals to $\rho^{-1}(1 + \rho)^2/\nu^2$.

Both estimators $\widehat{\alpha}$ and $\widehat{\beta}$ are required to estimate v^2 in (21) via $\widehat{\widetilde{g}}_0(t_i)$. Denoting $\widehat{v}^2$ the estimator of v^2 and using results from [22],

$$\widehat{v}^2 = \sum_i t_i^2 \widehat{\widetilde{g}}_0(t_i) - \left(\sum_i t_i \widehat{\widetilde{g}}_0(t_i)\right)^2, \tag{22}$$

where $\widehat{\widetilde{g}}_0(t_i)$ is shown in (20).

Normalizing the left side of (21) with $\rho^{-1}(1+\rho)^2/\widehat{v}^2$, setting $\beta_o = 0$ and squaring the final result, and using known properties of the family of Chi square distributions [25], one can test H_0 in (13) with the following expression:

$$Z_{\text{SemiP}} = n\rho(1+\rho)^{-2}\widehat{\beta}^2\widehat{v}^2 \xrightarrow[n\to\infty]{} \chi_1^2, \tag{23}$$

which has the Chi square distribution asymptotic behavior with 1 degree of freedom, χ_1^2. Under the idealized assumptions of model (11) and (12), a decision can be based on the value of Z_{SemiP} in (23), that is, high values of Z_{SemiP} reject hypothesis H_o, declaring the presence of local anomalies (note: Regarding (23), as typical from any asymptotic expression, the larger the value of n is the more accurate and precise the approximation of the expression will be. Since n in this context coincides with twice the sample size of the reference sample, sample sizes typically used in anomaly detection applications will suffice (greater than 100, yielding n greater than 200). Practitioners in statistics usually recommend that for univariate variables, asymptotic expressions should use at least thirty two observations, indicating that n in this case should be at least 32 or greater).

The test statistic in (23) will be referred to from here forward as the SemiP test statistic or SemiP anomaly detector, which has two steps: data transformation and test statistic estimation.

2.4. Implementation Notes for the Standing Alone SemiP Detector

2.4.1. Function Maximization. In order to implement (23), we perform an unconstrained maximization of the log maximum likelihood function in (18), or alternatively one could minimize the negative version of (18), to obtain the extremum estimators $\widehat{\alpha}$ and $\widehat{\beta}$. We used one of the conventional unconstrained nonlinear optimization algorithms—the simplex method [26], which is available in Matlab software (Release 13) under the function name fminsearch. The simplex method is a direct search method that does not use numerical or analytic gradients. If n is the length of x, a simplex in n-dimensional space is characterized by the $n + 1$ distinct vectors that are its vertices. For instance, in two-space, a simplex is a triangle; in three-space, it is a pyramid. At each step of the search, a new point in or near the current simplex is generated. The function value at the new point is compared with the function's values at the vertices of the simplex and, usually, one of the vertices is replaced by the new point, giving a new simplex. This step is repeated until the diameter of the simplex is less than the specified tolerance. A limitation using such a method is that it may find a local extremum, so the choices of initial parameters may prove to be critical in some cases; however, we found in practice that by setting the initial values to ($\alpha = 0$, $\beta = 0$), the method converges reasonably fast and works very well for all of the cases that we have observed, independently of whether anomalies were present or not in the tests.

The term $\widehat{v}^2$ in (23) is computed using $\widehat{v}^2 = \widehat{E}(t^2) - \widehat{E}^2(t)$, where $\widehat{E}(t^k) = \sum_i t_i^k \widehat{g}_0(t_i)$ and $\widehat{\widetilde{g}}_0(t_i)$ is shown in (20).

2.4.3. Decision Threshold. As mentioned earlier, using (23), high values of Z_{SemiP} reject hypothesis H_o in (13), declaring that x_1 is an anomaly relative to x_0. Using this detector as a standing alone unit, one could set a decision threshold based on the *type I error*, that is, based on the probability of rejecting H_o given that H_o is true. Using a standard integral table for the Chi square distribution, with 1 degree of freedom, find a threshold that yields an acceptable probability of error (e.g., 0.001).

2.5. Model Behavior in the Presence of Sample Mixtures. We show in this subsection the robustness of the semiparametric model toward an asymmetric test, that is, when a sample of a mixture is compared against a component of that mixture, which is found locally across the image in the form of spatial transitions. More specifically, we would like to show that $\widehat{\beta}$ (estimator for β in (13)) is significantly closer to *ZERO* when, for two PDFs $g_A(y) \neq g_B(y)$, $y_1 \sim g_B(y)$ and $y_0 \sim g_B(y)$ or $y_0 \sim (g_A(y) \cup g_B(y))$ (representing the union $\cup$ or a mixture of two PDFs) than when $y_1 \sim g_B(y)$ and $y_0 \sim g_A(y)$. We illustrate this fact using simulated data and focusing on three specific case studies.

Case 1. $y_0 \sim g_A(y)$ versus $y_1 \sim g_B(y)$.

Case 2. $y_0 \sim (g_A(y) \cup g_B(y))$ versus $y_1 \sim g_B(y)$.

Case 3. $y_0 \sim g_B(y)$ versus $y_1 \sim g_B(y)$.

According to [20], Case 2, which represents a transition of distinct regions in the image, appears some 20% of the entire image, or higher, as local patches are observed through a small moving window across typical images. Therefore, Case 2 is a major source of false alarms that could be avoided using a more robust model of the background than the typical models used in the target community. In anomaly detection applications, it is desired that the employed detector declares the presence of an anomaly for Case 1 but *no* anomalies for Cases 2 and 3; a declaration of *no* anomalies present is also desired, if Case 2 were reversed, that is, $y_0 \sim g_B(y)$ versus $y_1 \sim (g_A(y) \cup g_B(y))$, although this case is not shown here, its results are consistent with Table 3.

The results shown in Table 3 were computed using 100 simulated random samples from an i.i.d. Gaussian distribution to represent the reference sequence y_0 and another 100 samples to represent the test sequence y_1. A sequence representing a mixture of two classes consists of

Table 3: Behavior of $\widehat{\beta}$ on different case studies.

Case studies	Simulated samples Parameters $\mu_A=2000;\ \sigma_A^2=200$ $\mu_B=1000;\ \sigma_B^2=100$	$\widehat{\beta}$ $(\widehat{\alpha})$	Mean estimates $\widehat{\mu}=\sum_{i=1}^{n} t_i\widehat{\widetilde{g}}_0(t_i)$ $\widehat{\mu}_2=(1/n_0)\sum_{i=1}^{n_0} y_{0i}$ $\widehat{\mu}_3=(1/n)\sum_{i=1}^{n} t_i$	Variance estimates $\widehat{v}^2=\sum_{i=1}^{n} t_i^2\widehat{\widetilde{g}}_0(t_i)-\widehat{\mu}^2$ $\widehat{v}_2^2=(1/n_0)\sum_{i=1}^{n_0}(y_{0i}-\widehat{\mu}_2)^2$ $\widehat{v}_3^2=(1/n)\sum_{i=1}^{n}(t_i-\widehat{\mu}_3)^2$
Case 1	$y_0 \sim$ i.i.d $N(\mu_A,\sigma_A^2)$ $y_1 \sim$ i.i.d $N(\mu_B,\sigma_B^2)$	-0.7500 (848.75)	$1.9967e+003$ $1.9967e+003$ $1.4983e+003$	151.9466 153.4815 $2.4982e+005$
Case 2	$y_0 \sim \begin{cases}\text{i.i.d } N(\mu_A,\sigma_A^2)\\ \text{i.i.d } N(\mu_B,\sigma_B^2)\end{cases}$ $y_1 \sim$ i.i.d $N(\mu_B,\sigma_B^2)$	-0.0073 (8.0110)	$1.4990e+003$ $1.4990e+003$ $1.2494e+003$	$2.4997e+005$ $2.5316e+005$ $1.8859e+005$
Case 3	$y_0 \sim$ i.i.d $N(\mu_B,\sigma_B^2)$ $y_1 \sim$ i.i.d $N(\mu_B,\sigma_B^2)$	-0.0046 (4.5900)	999.8392 999.8392 999.6346	89.2284 89.2574 89.5693

50 samples for each class resulting in a total of 100 samples. The formulation and parameters used to generate these sequences are shown in Table 3 for different case studies, where the samples in row 2 (starting from the left upper corner in Table 3) simulates a local test between a genuine isolated object (y_1) and its homogeneous surrounding background (y_0)—*Case 1*, the samples in row 3 simulates a local test at a transition between two classes, where the test sample belongs to one of these classes—*Case 2*; the samples in row 4 simulates a local test within a homogeneous region—*Case 3*. Practical implementation details of the SemiP detector, which includes the estimation of (α,β), are shown in Section 2.3. The parameters (α,β) were estimated by maximizing the log likelihood function, using an optimization subroutine initialized to (0,0), (0,0), and (0,0) for *Cases 1, 2,* and *3*, respectively, so that convergence to a solution down to a tolerable error could be achieved using the subroutine.

Since the solution of the semiparametric model uses the union of samples t and estimators $\widehat{\alpha}$ and $\widehat{\beta}$ to estimate $\widetilde{g}_0$ (which itself is an estimator of g_0), we also included in Table 3 the mean estimates $\widehat{\mu}$, $\widehat{\mu}_2$, and $\widehat{\mu}_3$ and the variance estimates $\widehat{v}^2$, $\widehat{v}_2^2$, $\widehat{v}_3^2$; where $\widehat{v}^2$ estimates variance from the solution of the semiparametric model using the union of samples $t = (y_0, y_1)$ and $\widehat{\widetilde{g}}_0$; $\widehat{v}_2^2$ estimates variance using only the reference sample y_0; and $\widehat{v}_3^2$ is the sample variance using t. The mean estimates were computed accordingly, see Table 3.

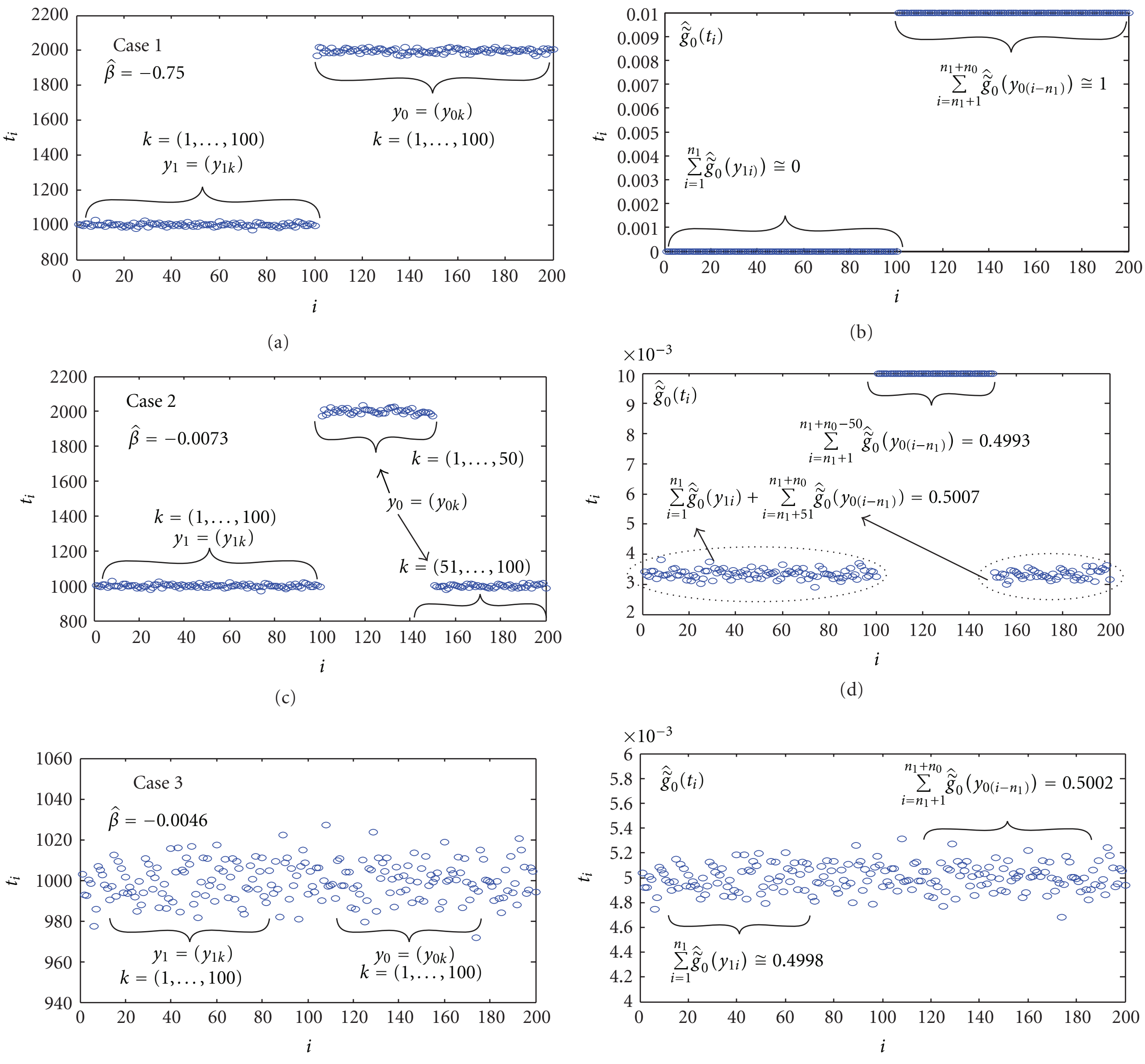
In reference to results shown in Table 3, recall that the null hypothesis is $\beta = 0$, and notice that the value of $\widehat{\beta}$ in Table 3 are significantly closer to *zero* for *Case 3* (homogeneous region) and *Case 2* (a transition of two different region) than for *Case 1* (genuine local anomaly), where in *Case 1* y_0 and y_1 do belong to different classes. Notice also that the disparity between the values of $\widehat{v}_2^2$ and $\widehat{v}_3^2$ for each case study also reflects how close $\widehat{\beta}$ is to *zero*. For instance, the disparity between $\widehat{v}_2^2$, and $\widehat{v}_3^2$ for *Case 1* is quite large compared to corresponding disparities for *Cases 2* and *3*.

The semiparametric model handles mixture by showing sensitivity on the estimation of β and α, such that when the test sample has strong statistical information about one of the subclasses in the reference sample, the semiparametric method responds by keeping both $\widehat{\beta}$ and $\widehat{\alpha}$ relatively close to *zero* in order to maximize the log likelihood function in (18). The estimates $\widehat{\alpha}$, $\widehat{\beta}$ affect directly the computation of $\widehat{\widetilde{g}}_0$, which in turn is used to compute $\widehat{v}^2$.

To shed some light on the effect of $\widehat{\beta}$ and $\widehat{\alpha}$ on $\widehat{\widetilde{g}}_0$, we present some results in Figure 2. The plots shown in Figure 2 (row 1) corresponds to *Case* 1; where, the plots on the left depict the values of t_i as a function of index i, for convenience we have marked where the sequences y_0 and y_1 are relative to each other within t; and the plots on the right depict $\widehat{\widetilde{g}}_0(t_i)$ versus i. Likewise for *Case* 2 (row 2) and *Case* 3 (row 3).

Let us consider first *Cases* 3 and 1. As mentioned earlier, because of the semiparametric model in (11) and (12), the union of samples $t = (y_0, y_1)$, where y_0 is the reference sequence, is used to estimate the reference PDF estimator $\widetilde{g}_0$. Circumstances when both samples belong to the same population (*Case* 3), the estimated cumulative *weight* for the test sample y_1, that is, $w_1 = \sum_{i=1}^{n_1}\widehat{\widetilde{g}}_0(y_{1i})$, is expected to be approximately equal to the estimated cumulative weight for the reference sample y_0, that is, $w_0 = \sum_{i=n_1+1}^{n_1+n_0}\widehat{\widetilde{g}}_0(y_{0(i-n_1)})$; because the constraint $\sum_{i=1}^{n=n_1+n_0}\widetilde{g}_0(t_i) = 1$ was used in the profiling method to attain $\widetilde{g}_0(t_i)$ in terms of α,β, we would expect both w_1 and w_0 to be near 0.500. Using simulated samples for the normal distributions and parameters shown in Table 3 for *Case* 3, we obtained $w_1 = 0.4998$ and $w_0 = 0.5002$, which are very close to our expectations, see Figure 6. We interpret the actual values of $\widetilde{g}_0(t_i)$ shown for *Case* 3 in Figure 6(f) to be an indication that the semiparametric method regards the test sample y_1 to be as *important* in the estimation of $\widetilde{g}_0$ as the reference sample y_0 is in that estimation, for the right justification, as both y_0 and y_1 happens to be governed by the same distribution.

In contrast, when both y_0 and y_1 belong to clearly different classes (see *Case* 1 in Table 3), the semiparametric method *recognizes* this difference and virtually *shuts down* the contribution of y_1 in the estimation of $\widetilde{g}_0$. The way it shuts down the contribution of y_1 is by maximizing the log likelihood function with values of $\widehat{\beta}$ and $\widehat{\alpha}$ that allow the estimates of atomic exponential distortions $\exp(\widehat{\alpha} + \widehat{\beta}t_i)$



(a)

(b)

(c)

(d)

(e)

(f)

Figure 6: Effect of estimators $\hat{\alpha}$, $\hat{\beta}$ on the computation of $\hat{\tilde{g}}_0$.

to be relatively *high* when t_i corresponds to components of y_1.

And since $\exp(\hat{\alpha}+\hat{\beta}t_i)$ are inversely proportional to $\hat{\tilde{g}}_0(t_i)$, see (20), the contributions of corresponding components of y_1 in t_i estimating $\tilde{g}_0$ are shut down as an indication of *nonimportance* to this estimation. The implication of this shut down is that the value of $\hat{\beta}$ is relatively away from *zero* (relative to *Case* 3, for instance), which rejects the null hypothesis as desired. Figures 6(a) and 6(b) show the combined samples for *Case* 1 and the resulting cumulative weights for the test sample $w_1 = \sum_{i=1}^{n_1} \hat{\tilde{g}}_0(y_{1i}) \cong 0.0$ and for the reference sample $w_0 = \sum_{i=n_1+1}^{n_1+n_0} \hat{\tilde{g}}_0(y_{0(i-n_1)}) \cong 1.0$, where $\cong$ denotes *approximately equal to*. The shutdown is reflected in the results for w_1.

In reference to *Case 2*, where the information carried in y_1 is also contained in y_0, as half of the random components in y_0 are governed by the same distribution of the random components in y_1—see Table 3 and Figures 6(c) and 6(d), the semiparametric method *recognizes* this fact by holding the value of $\hat{\beta}$ at near *zero*, and interestingly by allowing the contributions of y_1 estimating $\tilde{g}_0$ to be comparable to those contributions of the portion of y_0 that are similar to y_1. In other words, since both y_1 and y_0 are used to estimate $\tilde{g}_0$ and half of the random components in y_0 are governed by the same distribution of all of the random components in y_1 and the other half are governed by a different distribution, the semiparametric method will not *discriminate* between y_1 and the portion of y_0 that is similar to y_1. The outcome of this behavior is that, in order to maximize the log likelihood function, the values of $\hat{\beta}$ and $\hat{\alpha}$ are kept statistically close to *zero* (in this case $\hat{\beta} = -0.0073$) to reflect that the information in y_1 is *important* in the estimation of $\tilde{g}_0$. The way this method explicitly shows this nondiscrimination is by yielding the cumulative weight using y_1 and the portion of y_0 that is similar to y_1 to hold half of the power, while the other half of the power is allocated to the portion of y_0 that is dissimilar to y_1. In other words, using results from Figure 6(d) and this claim, we would expect a value of 0.5 (half power) for the cumulative weight using y_1 and the portion of y_0 that is *similar* to y_1; we obtained $\sum_{i=1}^{n_1} \hat{\tilde{g}}_0(y_{1i}) + \sum_{i=n_1+51}^{n_1+n_0} \hat{\tilde{g}}_0(y_{0(i-n_1)}) = 0.5007$. And we would expect the other half of the power to be in the cumulative weight using the portion of y_0 that is *dissimilar* to y_1; we obtained $\sum_{i=n_1+1}^{n_1+n_0+50} \hat{\tilde{g}}_0(y_{0(i-n_1)}) = 0.4993$. Notice that adding 0.5007 and 0.4993 yields exactly 1.0 as expected because the constraint $\sum_{i=1}^{n=n_1+n_0} \tilde{g}_0(t_i) = 1$ was used in the profiling method in order to attain a representation of $\tilde{g}_0$ in terms of free parameters β and α.

A conclusion that we can draw from this discussion is that the semiparametric method will indirectly compare two samples y_0 (reference) and y_1 (test) by assuming that the distribution of $y_1(g_1)$ and the distribution of $y_0(g_0)$ are related (exponentially) to each other and that, therefore, the information content in both samples can be used to estimate one of these distributions, in particular, g_0.

We found this indirect comparison method to be highly sensitive to the cumulative contribution of y_1 estimating g_0. This sensitivity has an important practical value in the anomaly detection application for three reasons.

First, if $g_1 = g_0$, both samples y_0 and y_1 are expected to equally contribute to the estimation of g_0, which in fact would improve that estimation due to the increase of sample size. Result: y_1 would be labeled as *not* being anomalous to y_0 in this application.

Second, if $g_1 \neq g_0$, sample y_1 is *not* expected to be allowed to contribute to the estimation of g_0, thus, this estimation would solely rely on the cumulative contribution of y_0. Result: y_1 would be labeled as being anomalous to y_0 in this application.

And third, if g_0 is a mixture of densities, such that, g_1 is a component in that mixture, we found that the contribution of y_1 would not be suppressed, but proportional to the weight of g_1 in that mixture (see Figure 6). Result: y_1 would be labeled as *not* being anomalous to y_0 in this application.

This behavior of the semiparametric test statistic is highly desired in the target community because it conforms with be behavior of a human analyst performing the same task in the target application, and it separates this method from existing ones performing the same task.

3. Quasiglobal Semiparametric Approach

The semiparametric test statistic is used as the primary scoring method for the quasi-global anomaly detection approach. As mentioned in Section 1, the quasi-global anomaly detection approach was designed to tackle the forward looking anomaly detection problem, although the application of the quasi-global algorithm using nadir looking imagery are also considered in Section 4.

We start by describing the background sampling method and its probabilistic model, followed by description of the quasi-global algorithm framework using the semiparametric test statistic.

3.1. Sampling Method and Its Probabilistic Model. Assume that target pixels are present in the $R \times C$ spatial area of a $R \times C \times K$ HS data cube $\mathbf{X}$, denote a the total number of target pixels in $\mathbf{X}$, q the probability of a pixel in $\mathbf{X}$ belonging to the target, and the relationship $q = a/A$, where $A = RC$ (or all pixels in $\mathbf{X}$) (in most applications q is unknown, and if multiple targets are present in the imagery, a will be the total number of all pixels belonging to all targets present in the imagery; also, these targets may or may not have the same material type). In order to represent the unknown clutter background in the imagery, let N blocks of data—all having a fixed small area $(n \times n) \ll (R \times C)$—be randomly selected from the $R \times C$ area, see one of the data cubes in Figure 1. In theory, for $(n \times n) = (1 \times 1)$ and using the assumption that target pixels in $\mathbf{X}$ are disjoint and randomly located across the $R \times C$ imagery area (note that in practice, this assumption is not satisfied when targets are present in the scene, but we will use this assumption to establish a guideline), the probability P that at least one block of data has a target pixel is

$$P(m \geq 1) = 1 - p(m = 0), \tag{24}$$

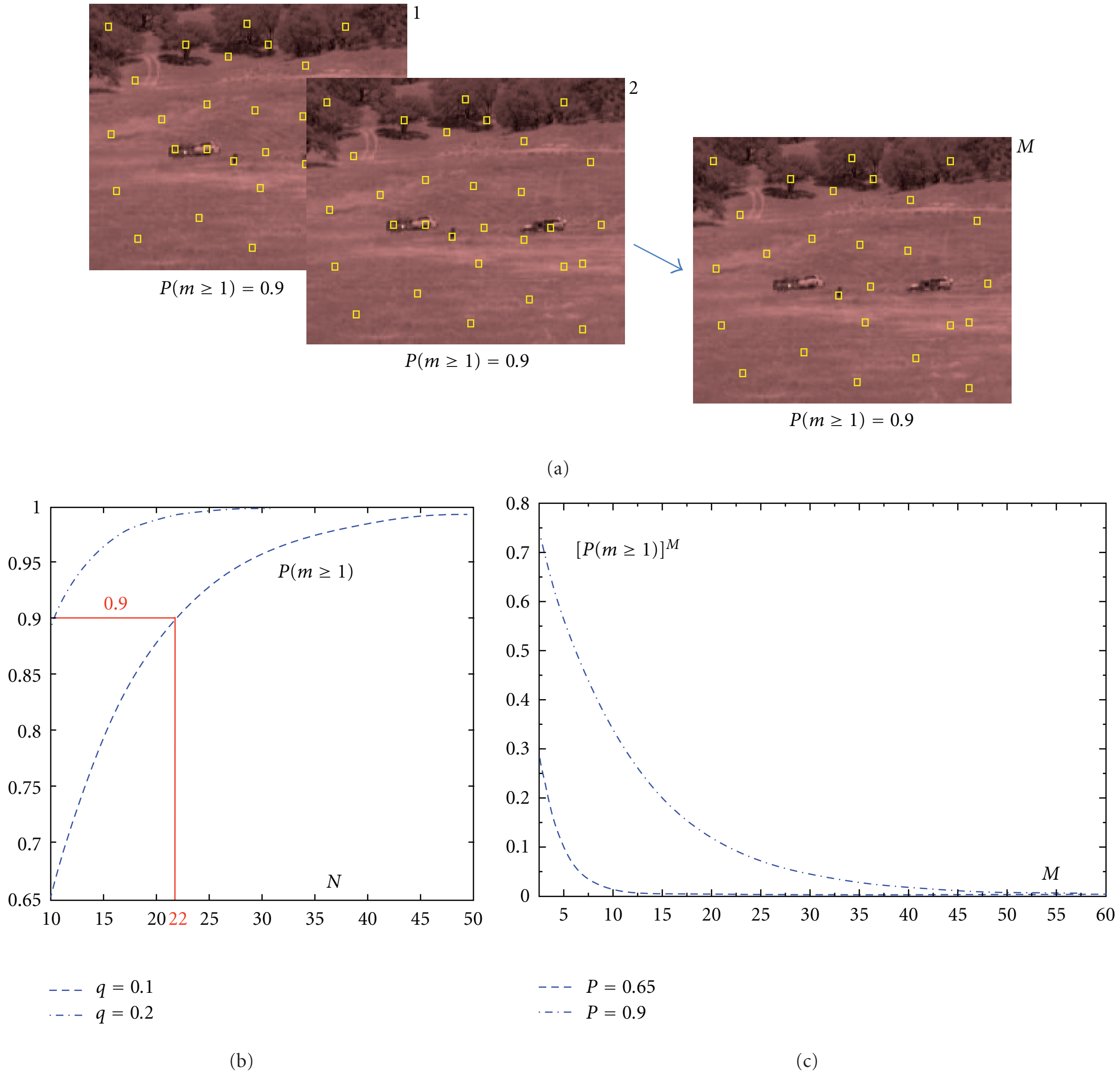
where p is the binomial density function [27], given parameters q and N, and $m \in \{0, 1, \ldots, N\}$ is the number of blocks of data containing a target pixel, or

$$p(m \mid q, N) = \frac{N!}{m!(N-m)!} q^m (1-q)^{N-m}, \tag{25}$$

(symbols | and ! denote *given parameters* and the *factorial operator*, resp.).

For convenience, we will refer to $P(m \geq 1)$ as the *probability of contamination* and m the number of *contaminated* blocks of data.

The implementation of this contamination model to the autonomous background sampling problem requires that each one of the $N(n \times n)$ blocks of data be regarded as an independent reference set $\mathbf{W}_0^{(f)}$ $(f = 1, 2, \ldots, N)$ representing clutter spectra, where $\mathbf{W}_0^{(f)} \in \mathbf{R}^{K \times n_0}$ is a rearranged sequence version of the fth block of data having $n_0 = n^2$ spectra. By necessity, n_0 must be significantly greater than *one*—for statistical purposes—but yet significantly smaller than $A = RC$ (e.g., $n_0/A = 20^2/640^2 = 0.000977$) in order to reasonably satisfy the assumption that a $n \times n$ block of data has an unit area at the $R \times C$ imagery area. A contaminated block of data, then, will be treated qualitatively as a block having target pixels covering a large portion of the block's area (e.g., greater than 0.70). In addition—when targets are present, since pixels representing a single target are expected to be clustered in the imagery, the assumption that each target pixel is randomly located



(a)

(b)

(c)

Figure 7: The relationship between N(the number of randomly selected blocks of data, shown as yellow squares on the imagery) and the contamination probability $P_g(m \geq 1) = P(m \geq 1)$ is shown in (b) for a given q (e.g., $q = 0.10$), which is an upper bound guess representing the maximum ratio between target pixels over the $R \times C$ area. To better characterize the unknown clutter background, a high $N$is most desired, but at a high cost, that is, an undesirably high $P_g(m \geq 1)$. The overall contamination probability, however, can significantly decrease by independently repeating the random sampling process Mnumber of times, as shown (c) of figure, and then fusing results using a suitable method.

across the imagery area will be ignored. Using (24), while ignoring the nonclustered target pixel assumption, implies that the probability of contamination will be overestimated, as blocks of data are less likely to be randomly selected from the same cluster of target pixels (for the autonomous background sampling problem, it is more conservative to overestimate the probability of contamination than to underestimate).

Figure 7(b) shows a plot of the probability of contamination $P(m \geq 1)$ versus N, for two values of q (0.1 and 0.2). It is highlighted in the plot in reference that, for instance, if parameters are set to $(q, N) = (0.10, 22)$ then $P(m \geq 1) = 0.90$. Notice that for $N = 22$, if target pixels are present but cover less than $q = 0.10$ of the imagery area, $P(m \geq 1) = 0.90$ is overestimated by two fronts: (i) pixels from a single target are not randomly spread across the imagery area, but clustered, and (ii) the cumulative number of target pixels covers less than 0.10 of the imagery area. So, (24) provides an upper bound (conservative) approximation of the probability of contamination, given parameters q and N.

Figure 7(b) also shows the tradeoff between having a larger number of spectral sets (increasing N) in order to adequately represent the clutter background, which is desired, and the cost of increasing probability of contamination, which is not desired (in particular, contamination implies that once target pixels are randomly selected by chance from the imagery area, they will be used by a detector as reference set to test the entire imagery, in which case targets would be suppressed).

Since the presence of target pixels in the imagery is unknown a priori, finding a way to decrease the probability of contamination becomes a necessity. In order to decrease this probability, using an adequately large N and a sensible value for q, we propose to independently repeat the random sampling process described in this subsection M number of times. Figure 7(c) illustrates the outcome of M repetitions. If we denote the probability of contamination of the gth random sampling process (or repetition) as $P_g(m \geq 1)$, $1 \leq g \leq M$, for a fixed q and N, note that each $P_g(m \geq 1) = P(m \geq 1)$ and, since $0.0 \leq P(m \geq 1) \leq 1.0$ and these processes will be repeated independently from each other, the overall probability $\widetilde{P}$ that *all* of the processes will be contaminated with at least a contaminated block of data will decrease as a function of increasing M, or

$$\widetilde{P} = P_1(m \geq 1)P_2(m \geq 1)\cdots P_M(m \geq 1) = [P(m \geq 1)]^M. \tag{26}$$

Figure 7(c) plots $\widetilde{P}$ as a function of increasing M, for $P(m \geq 1) = 0.90$ and $P(m \geq 1) = 0.65$. Taking, as an example, the $\widetilde{P}$ curve in Figure 7 corresponding to using $P(m \geq 1) = 0.90$ in (24), notice that for $M > 40$, $[P(m \geq 1)]^M$ decreases to virtually *zero*. This outcome implies that at least one out of the $M > 40$ processes has an extremely high probability of not being contaminated, given that $N = 22$ and target pixels do not cover significantly more than 10% of the imagery area ($q = 0.10$).

3.2. Algorithmic Fusion. A framework for the quasi-global semiparametric anomaly detection algorithm can now be developed using (i) the repeated random sampling model discussed in Section 3.1 (needed to characterize the unknown clutter background in the imagery), (ii) the semiparametric anomaly detector discussed in Section 2.3 (needed to test reference data against the entire imagery), (iii) a way to fuse the results from testing N randomly chosen blocks of data against the entire imagery using small windows (this will produce a 2-dim output surface per repetition), and (iv) a way to fuse M independently produced 2-dim output surfaces into a single 2-dim decision surface.

Start by letting a HS data ($R \times C \times K$) cube $\mathbf{X}$ be available for autonomous testing. Let also N blocks ($n \times n$) of data be randomly selected from the $\mathbf{X}$'s $R \times C$ spatial area and used as a reference library set $\mathbf{W}_0^{(f)}$ ($f = 1, 2, \ldots, N$) representing clutter background spectra, where $\mathbf{W}_0^{(f)} = (\mathbf{y}_{01}^{(f)}, \ldots, \mathbf{y}_{0n_0}^{(f)})$ is a rearranged sequence version of the fth block of data having $n_0 = n^2$ spectra, where$\{\mathbf{y}_{0u}^{(f)}\}_{u=1}^{n_0} \in \mathbf{R}^K$ are K-dim column vectors. Let $\mathbf{W}_1 = (\mathbf{y}_{11}, \ldots, \mathbf{y}_{1n_1})$ be the rearranged version of a ($n \times n$) window of test data at location ij in $\mathbf{X}$—see (1) for column vectors $\{\mathbf{y}_{1h}\}_{h=1}^{n_1} \in \mathbf{R}^K$; first, we would like to automatically test $\mathbf{W}_1$ against all $\{\mathbf{W}_0^{(f)}\}_{f=1}^{N}$, and produce a single output (scalar) value $\widetilde{Z}_{\text{SemiP}}^{(ij)} \geq 0.0$ from these N test results.

For better clarity in this subsection, we repeat the data transformation steps discussed in Section 2.2, but with the inclusion of index $f = 1, \ldots, N$ and letting $\mathbf{y}_{ij} = L_{i1j}, \ldots, L_{iKj}$, where L_{ikj} is the kth radiance value in $\mathbf{y}_{ij}$, $k = 1, \ldots, K$, and

$$\mathbf{W}_0^{(f)} = \left[\mathbf{y}_{01}^{(f)}, \ldots, \mathbf{y}_{0n_0}^{(f)}\right] = \begin{bmatrix} L_{011}^{(f)}, \ldots, L_{01n_0}^{(f)} \\ \vdots \\ L_{0K1}^{(f)}, \ldots, L_{0Kn_0}^{(f)} \end{bmatrix}, \tag{27}$$

$$\nabla_0^{(f)} = \begin{bmatrix} \left(L_{021}^{(f)} - L_{011}^{(f)}\right), \ldots, \left(L_{02n_0}^{(f)} - L_{01n_0}^{(f)}\right) \\ \vdots \\ \left(L_{0K1}^{(f)} - L_{0(K-1)1}^{(f)}\right), \ldots, \left(L_{0Kn_0}^{(f)} - L_{0(K-1)n_0}^{(f)}\right) \end{bmatrix}, \tag{28}$$

$$\overline{\nabla}_0^{(f)} = \frac{1}{n_0} \nabla_0^{(f)} \mathbf{1}_{n_0 \times 1} \tag{29}$$

and, denoting the columns of $\nabla_0^{(f)}$ as $\{\nabla_{0u}^{(f)}\}_{u=1}^{n_0}$,

$$\left\{ x_{0u}^{(f)} = \frac{180}{\pi} \arccos\left(\frac{\left(\nabla_{0u}^{(f)}\right)^t \overline{\nabla}_0^{(f)}}{\left\|\nabla_{0u}^{(f)}\right\| \left\|\overline{\Delta}_0^{(f)}\right\|} \right) \right\}_{u=1}^{n_0}. \tag{30}$$

And equivalently for $\mathbf{W}_1 = (\mathbf{y}_{11}, \ldots, \mathbf{y}_{1n_1})$—the rearranged version of a ($n \times n$) window of test data at location ij in $\mathbf{X}$ and the columns of $\nabla_0^{(f)}$ in (28)—$\{\nabla_{0u}^{(f)}\}_{u=1}^{n_0}$, one has

$$\left\{ x_{1u}^{(f)} = \frac{180}{\pi} \arccos\left(\frac{\left(\nabla_{0u}^{(f)}\right)^t \overline{\nabla}_1}{\left\|\nabla_{0u}^{(f)}\right\| \left\|\overline{\Delta}_1\right\|} \right) \right\}_{u=1}^{n_0}. \tag{31}$$

From (30) and (31), the following two univariate sequences will be used as inputs to the SemiP detector:

$$x_0^{(f)} = \left(x_{01}^{(f)}, x_{02}^{(f)}, \ldots, x_{0n_0}^{(f)}\right), \tag{32}$$

$$x_1^{(f)} = \left(x_{11}^{(f)}, x_{12}^{(f)}, \ldots, x_{1n_0}^{(f)}\right), \tag{33}$$

where $1 \leq f \leq N$.

Following the discussion that led to (33), results from the semiparametric test statistic can be used (or fused) as following:

$$\widetilde{Z}_{\text{SemiP}}^{(ij)} = \min_{1 \leq f \leq N} Z_{\text{SemiP}}^{(ij)(f)}, \tag{34}$$

where

$$Z_{\text{SemiP}}^{(ij)(f)} = n\rho(1+\rho)^{-2}\left(\widehat{\beta}^{(f)}\right)^2 \widehat{\nu}^{2(f)}, \tag{35}$$

$\{Z^{(ij)(f)}_{\text{SemiP}}\}_{f=1}^{N} \geq 0.0$, $n_1 = n_0 = n^2$, $\rho = n_1/n_0$, $(i = 1,\dots,R - n - 1)$ and $(j = 1,\dots,C - n - 1)$ index the left-upper corner pixel of an $n \times n$ window in $\mathbf{X}$; using (32) and (33),

$$\begin{aligned} t^{(f)} &= \left(x_{01}^{(f)},\dots,x_{0n_0}^{(f)},x_{11}^{(f)},\dots,x_{1n_0}^{(f)}\right) \\ &= \left(t_1^{(f)},\dots,t_{n_1+n_0}^{(f)}\right), \end{aligned} \tag{36}$$

$$\hat{v}^{2(f)} = \sum_{u=1}^{n_1+n_o} t_u^{2(f)}\hat{\tilde{g}}_0\left(t_u{}^{(f)}\right) - \left(\sum_{u=1}^{n_1+n_0} t_u{}^{(f)}\hat{\tilde{g}}_0\left(t_u{}^{(f)}\right)\right)^2, \tag{37}$$

$$\hat{\tilde{g}}_0\left(t_u{}^{(f)}\right) = \frac{1}{n_0}\frac{1}{1+\rho\exp\left(\hat{\alpha}^{(f)}+\hat{\beta}^{(f)}t_u{}^{(f)}\right)}, \tag{38}$$

and estimates $\hat{\alpha}^{(f)}$ and $\hat{\beta}^{(f)}$ can be obtained by replacing (α,β) with $(\hat{\alpha},\hat{\beta})$ in (18) and then performing an unconstrained maximization of $l(\alpha,\beta)$; for this paper, a standard unconstrained minimization routines available in Matlab software (i.e., *fminsearch*) was used, setting initial values to $(\hat{\alpha},\hat{\beta}) = (0,0)$.

Notice that if $Z^{(ij)(1)}_{\text{SemiP}}, Z^{(ij)(2)}_{\text{SemiP}},\dots,Z^{(ij)(N)}_{\text{SemiP}}$ are placed in ascending order and denoted by $Z^{(ij)}_{\text{SemiP}(1)}$, $Z^{(ij)}_{\text{SemiP}(2)},\dots$, $Z^{(ij)}_{\text{SemiP}(N)}$, such that $Z^{(ij)}_{\text{SemiP}(1)} \leq Z^{(ij)}_{\text{SemiP}(2)} \leq \dots \leq Z^{(ij)}_{\text{SemiP}(N)}$, then $\tilde{Z}^{(ij)}_{\text{SemiP}} = Z^{(ij)}_{\text{SemiP}(1)}$ according to (34)—the lowest order statistics (see, for instance, [28]). This fact will be used in estimating the asymptotic behavior of the overall quasi-global semiparametric algorithm, shown in the Appendix.

Notice also that if $\mathbf{W}_1$ is significantly different from all $\{\mathbf{W}_0^{(f)}\}_{f=1}^{N}$, then all of the corresponding results $\{Z^{(ij)(f)}_{\text{SemiP}}\}_{f=1}^{N}$ in (35) would yield high values; this outcome would guarantee the lowest order statistics $\tilde{Z}^{(ij)}_{\text{SemiP}}$ in (34) to hold a high value. Otherwise, if $\mathbf{W}_1$ is significantly similar to at least one of the samples in $\{\mathbf{W}_2^{(f)}\}_{f=1}^{N}$, then at least one of the corresponding results in $\{Z^{(ij)(f)}_{\text{SemiP}}\}_{f=1}^{N}$ would yield a low value; this low value would be assigned to $\tilde{Z}^{(ij)}_{\text{SemiP}}$, according to (34).

Since it is unknown a priori whether target spectra are present in $\mathbf{X}$, the entire $\mathbf{X}$ needs to be tested. In order to achieve it, all $\{\tilde{Z}^{(ij)}_{\text{SemiP}}\}_{i=1,j=1}^{R-n-1,C-n-1}$ must be computed according to (34), producing a 2-dim output surface $\tilde{\mathbf{Z}}^{(g)}_{\text{SemiP}}$. The index $g(1 \leq g \leq M)$ has been introduced to results produced by (34) in order to denote the repetition number discussed in Section 3.1, yielding

$$\tilde{\mathbf{Z}}^{(g)}_{\text{SemiP}} = \begin{bmatrix} \tilde{Z}^{(11)(g)}_{\text{SemiP}},\dots,\tilde{Z}^{[1(C-n-1)](g)}_{\text{SemiP}} \\ \vdots \\ \tilde{Z}^{[(R-n-1)1](g)}_{\text{SemiP}},\dots,\tilde{Z}^{[(R-n-1)(C-n-1)](g)}_{\text{SemiP}} \end{bmatrix}. \tag{39}$$

The computation leading to (39) will be independently repeated M number of times in order to significantly reduce the probability of contamination (i.e., samples of targets labeled as clutter background). Applying a cutoff threshold to all pixel values $\tilde{Z}^{(ij)(g)}_{\text{SemiP}}$ in $\tilde{\mathbf{Z}}^{(g)}_{\text{SemiP}}$, such that, pixel values that are above or equal to the threshold become 1 and values below become 0, yielding a binary surface (a probabilistic cutoff threshold for this algorithm is presented in Section 4.3). The M binary surfaces are fused using the logic *OR* operator $\oplus$, leading to the algorithm's final output surface $\mathbf{Z}_{\text{SemiP}}$, or

$$\mathbf{Z}_{\text{SemiP}} = \begin{bmatrix} \left(\tilde{Z}^{(11)(1)}_{\text{SemiP}} \oplus \cdots \oplus \tilde{Z}^{(11)(M)}_{\text{SemiP}}\right),\dots,\left(\tilde{Z}^{[1(C-n-1)](1)}_{\text{SemiP}} \oplus \cdots \oplus \tilde{Z}^{[1(C-n-1)](M)}_{\text{SemiP}}\right) \\ \vdots \\ \tilde{Z}^{[(R-n-1)1](1)}_{\text{SemiP}} \oplus \cdots \oplus \tilde{Z}^{[(R-n-1)1](M)}_{\text{SemiP}},\dots,\tilde{Z}^{[(R-n-1)(C-n-1)](1)}_{\text{SemiP}} \oplus \cdots \oplus \tilde{Z}^{[(R-n-1)(C-n-1)](M)}_{\text{SemiP}} \end{bmatrix}. \tag{40}$$

Figure 8 illustrates $\tilde{\mathbf{Z}}^{(g)}_{\text{SemiP}}$ (39) and $\mathbf{Z}_{\text{SemiP}}$ (40) through a repeated random sampling and result fusion diagram. The diagram shows M independent paths, where, in each path, independent blocks of data are randomly selected from the input HS data cube so that the entire data cube can be tested against these blocks of data, using a testing window of the same block size. Each path, which is indexed by $g\,(1 \leq g \leq M)$, produces a 2-dim output surface $(\tilde{\mathbf{Z}}^{(g)}_{\text{SemiP}})$, where, at the backend of the diagram, all $\{\tilde{\mathbf{Z}}^{(g)}_{\text{SemiP}}\}_{g=1}^{M}$ passes through a logical OR operator on a pixelwise fashion (i.e., only the pixel values at the same pixel location are logically OR'ed), producing a final 2-dim surface $\mathbf{Z}_{\text{SemiP}}$, as shown in (40).

The motivation and functionality shown in Figure 8 are summarized as follows: for a given repetition $g\ (1 \leq g \leq M)$, assume that the realization of $\mathbf{W}_1$ from a window location ij in $\mathbf{X}$ is a spectral sample of a target, and the realizations of $\{\mathbf{W}_0^{(f)}\}_{f=1}^{N}$ are samples of various materials composing the clutter background in $\mathbf{X}$, that is, the randomly selected blocks of data are not contaminated with target spectra. The semiparametric order statistics in (34) is expected to yield a high value at that ij location. Moreover, if the target scale in $\mathbf{X}$ is larger than $n \times n$, then the target will be represented by multiple pixels in $\tilde{\mathbf{Z}}^{(g)}_{\text{SemiP}}$ (39), having high values. These pixels are expected to be clustered, hence, accentuating the target spatial location in $\tilde{\mathbf{Z}}^{(g)}_{\text{SemiP}}$. However, as discussed in Section 3.1, the contamination probability

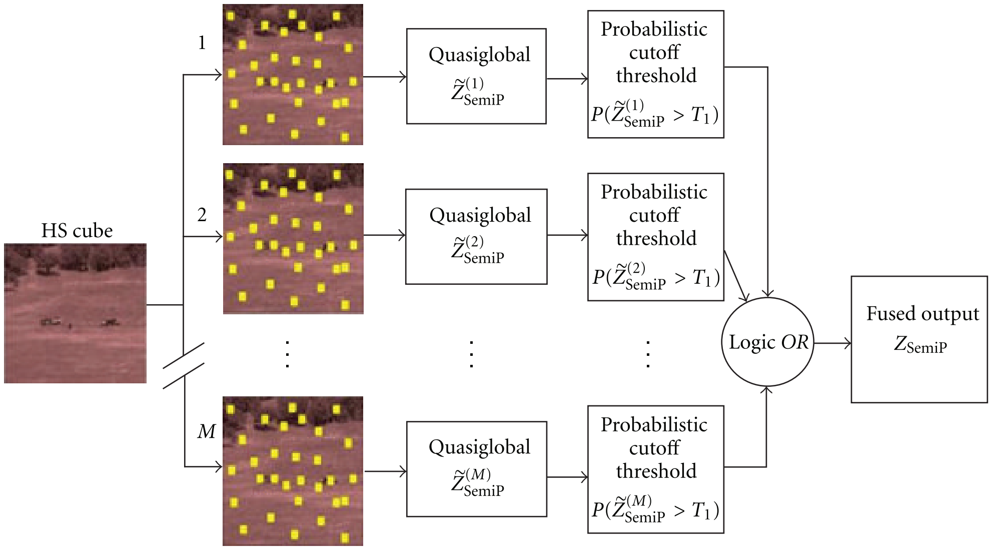

HS cube
1
2
M
Quasiglobal $\tilde{Z}^{(1)}_{\text{SemiP}}$
Quasiglobal $\tilde{Z}^{(2)}_{\text{SemiP}}$
Quasiglobal $\tilde{Z}^{(M)}_{\text{SemiP}}$
Probabilistic cutoff threshold $P(\tilde{Z}^{(1)}_{\text{SemiP}} > T_1)$
Probabilistic cutoff threshold $P(\tilde{Z}^{(2)}_{\text{SemiP}} > T_1)$
Probabilistic cutoff threshold $P(\tilde{Z}^{(M)}_{\text{SemiP}} > T_1)$
Logic *OR*
Fused output Z_{SemiP}

FIGURE 8: Quasiglobal, semiparametric anomaly detection algorithm.

$P(m \geq 1)$, for a given g, increases as a function of increasing N, see Figure 7. Figure 7 shows further that for a fixed q, N and an adequately large M, if (for instance) results $\tilde{Z}^{(22)(1)}_{\text{SemiP}}, \tilde{Z}^{(22)(2)}_{\text{SemiP}}, \ldots, \tilde{Z}^{(22)(M)}_{\text{SemiP}}$ correspond to the same portion of the target at a testing window location ($i = 2$, $j = 2$), then (40) give us the confidence that at least one term in $\tilde{Z}^{(22)(1)}_{\text{SemiP}}, \tilde{Z}^{(22)(2)}_{\text{SemiP}}, \ldots, \tilde{Z}^{(22)(\text{M})}_{\text{Semip}}$ will have a high value with high probability $[1.0 - \tilde{P}(\tilde{m} = M)]$; after application of a cutoff threshold to results in (39), the high value(s) in reference would be captured by the logic *OR* operator, for example, $(\tilde{\text{Z}}^{(22)(1)}_{\text{SemiP}} \oplus \cdots \oplus \tilde{\text{Z}}^{(22)(\text{M})}_{\text{SemiP}})$, as shown in (40) for all ij locations. Notice that a target may also be represented by multiple (clustered) pixel locations in $\mathbf{Z}_{\text{SemiP}}$ (40).

3.3. Setting the Cutoff Threshold and Other Parameters. For autonomous remote sensing applications, properly setting the algorithm's parameters is a critical step. This subsection presents a guideline to address this step. For the quasi-global, semiparametric anomaly detection algorithm, the parameters of main concern are T_1 (the probabilistic cutoff threshold), N (the number of randomly selected blocks of data), M (the number of testing repetitions), and q (the upper bound ratio of target pixels in the data cube over the spatial area of this cube).

Using the asymptotic behavior shown in (A.11) in the Appendix, a cutoff threshold is attained as follow:

$$T_1 = T(a) = \hat{\mu}_{\tilde{g}} + a\hat{\sigma}_{\tilde{g}}, \tag{41}$$

where $a = \tilde{g}^{-1}(1 - \varepsilon_1)$ is the $1 - \varepsilon_1$ quantile of $\tilde{g}(z) = Lg(z)[1 - G(z)]^{L-1}$ (see Appendix),

$$\hat{\mu}_{\tilde{g}} = \sum_{u=1}^{\tilde{n}} z_u \tilde{g}(z_u),$$
$$\hat{\sigma}_{\tilde{g}} = \sqrt{\sum_{u=1}^{\tilde{n}} z_u^2 \tilde{g}(z_u) - \hat{\mu}_{\tilde{g}}^2} \tag{42}$$

are estimates of the mean and standard deviation, respectively, of the known distribution $\tilde{g}(z)$—these estimates can be attained a priori through a simulation experiment using $\tilde{n}$ samples of $\tilde{g}(z)$, and $0 \leq \varepsilon_1 \leq 1$.

For setting parameters N and M, as discussed in Section 3.1, the quasi-global semiparametric algorithm—ideally—requires an adequately large, which undesirably increases the contamination probability $P(m \geq 1)$ per repetition, and an adequately large M, which desirably decreases the overall cumulative contamination probability $\tilde{P}(\tilde{m} = M)$. From (24), (25), and (26) and using the *log* of base 10, a direct transformation leads to

$$N = \frac{\log[1 - P(m \geq 1)]}{\log(1 - q)}, \tag{43}$$

$$\text{M} = \frac{\log\left[\tilde{P}(\tilde{m} = M)\right]}{\log\left[1 - (1 - q)^N\right]}. \tag{44}$$

For a given q, we can fix the values of $P(m \geq 1)$, $\tilde{P}(\tilde{m} = M)$ and obtainand M directly using (43) and (44), respectively. As a guideline, $P(m \geq 1)$ should be set high, but less than 1.0, so that N can also be relatively high and $\tilde{P}(\tilde{m} = M) < 1.0$; $\tilde{P}(\tilde{m} = M)$ should be set very low, near zero. As long as the guideline is followed, interestingly, the actual values of $P(m \geq 1)$ and $\tilde{P}(\tilde{m} = M)$ are unimportant.

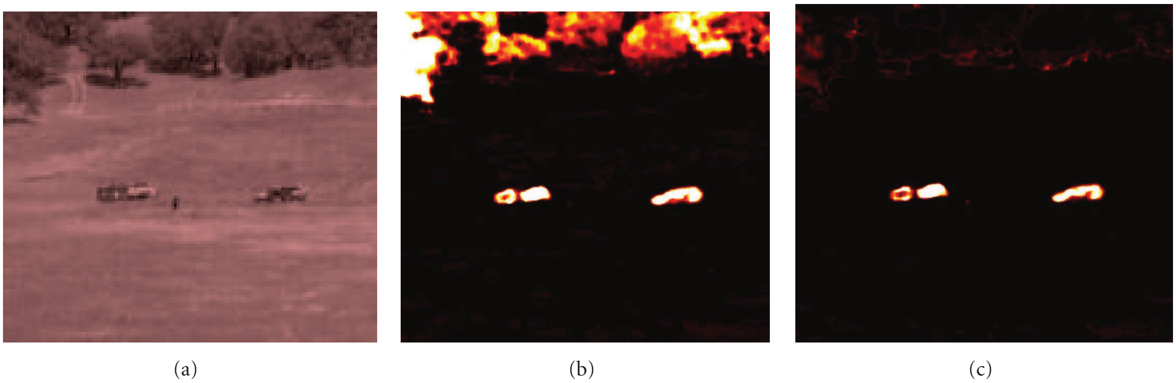


Figure 9: QG-SemiP results testing Cube 2 (a) for scene anomalies; output surface (b) was produced using ($q = 0.10$; $T_1 = T(20)$; $N = 3$; $M = 3$); output surface (c) was produced using ($q = 0.10$; $T_1 = T(20)$; $N = 22$; $M = 40$). Bright pixel values (white) in the output surfaces correspond to values above the probabilistic cutoff threshold T_1—depicting the highest confidence level of anomaly presence in the imagery, relative to N randomly selected blocks of data. Testing procedure was independently repeated M times, as highlighted in Figure 8. Using the available ground truth information of the scene, the white clusters in the far right figure cover about 90% of the motor vehicles (the targets) and no false alarms.

For example, we could fix $P(m \geq 1) = 0.90$ and $\widetilde{P}(\widetilde{m} = M) = 0.01$, and for $q = 0.10$, we obtain directly from (43) and (44) parameter values $N \approx 45$ and $M \approx 44$ (Since N and M are defined as integers, these numbers are rounded off $\approx$). For the results shown in Section 4, we fixed at once $q = 0.10$, $P(m \geq 1) = 0.90$, and $P_R(m_R = M) = 0.015$, which by using (43) and (44) yield $N \approx 22$ and $M \approx 40$.

4. Results

The QG-SemiP algorithm was applied to the HS imagery shown in Figure 1, that is, Cube 1 (nadir looking imagery) and Cube 2 through Cube 4 (forward looking imagery), to test for scene (spectral) anomalies. This subsection presents performance summary using results and guideline discussed in Section 3.3 to set algorithm parameters (q, T_1, N, M). Results using forward looking imagery will be discussed first.

4.1. Forward Looking Imagery. Results testing Cube 2 are shown in Figure 9. For display purposes, the output surface $\mathbf{Z}_{\text{SemiP}}$ (Figure 9, center and right hand side surfaces) is *not* shown as a binary surface; instead, each $\widetilde{\mathbf{Z}}^{(g)}_{\text{SemiP}}$ $(g = 1, \ldots, M)$ output surface is mapped using a pseudocolor map, such that, the brightest pixel values in those surfaces (*white* colored pixels, representing strongest evidence of anomalies) show the locations of results above or equal to the cutoff threshold T_1; while other colors (yellow, red, brown, and black) show lesser evidence of anomalies at the corresponding pixel locations, relative to randomly selected blocks of data. (The color *black* represents no evidence of anomalies.) All of the M surfaces $\widetilde{\mathbf{Z}}^{(g)}_{\text{SemiP}}$, using the threshold based colormap in reference, are then summed to yield the output surface shown in Figure 9 through Figure 11. Additional details follow.

Figures 9(b) and 9(c) show two different outcomes of the quasi-global, semiparametric anomaly detection algorithm, where $n \times n$ was fixed at once to 20×20 (for all data blocks and testing window sizes) and algorithm's parameters were set to ($q = 0.10$; $T_1 = T(20)$; $N = 3$; $M = 3$)—center display—and ($q = 0.10$; $T_1 = T(20)$; $N = 22$; $M = 40$)—right hand side display. The center output surface depicts an example when N is not set sufficiently high, hence, obtaining an inadequate representation of the clutter background. In this case, three blocks of data were randomly selected from the scene (most likely from the open field area, since it is the largest area in the scene), and used by the QG-SemiP detector to suppress (according to $\widetilde{\mathbf{Z}}^{(g)}_{\text{SemiP}}$ $(g = 1, \ldots, M)$ in (39)) the open field in Cube 2, not only once, but most likely $M = 3$ times. As a result, the three motor vehicles and the canopy area on the upper portion of that scene were accentuated relative to the open field. For this initial experiment, we ignored the binomial distribution model and set parameters N and M intentionally low and tested Cube 2 to show the undesired result in Figure 9(b).

For parameters accordingly set to ($q = 0.10$; $T_1 = T(20)$; $N = 22$; $M = 40$) yielded a significantly better result by detecting only the three motor vehicles in the scene, while suppressing the unknown clutter background, see Figure 9(c). Using the available ground truth information of the scene, the white clusters cover about 90 percent of the motor vehicles' spatial area (the targets) and no false alarms. As discussed earlier, for many remote sensing applications, targets (if present in the scene) will not cover more than 10 percent of the imagery spatial area. For instance, the motor vehicle shown at broadside in Cube 2 has 25,000 pixels, covering 6.1% of the imagery area (25000/409600). Note that $q = 0.1$ is robust, because it is independent of targets' aspect or depression angles, relative to the sensor; independent of the number of targets in the scene; and independent of targets' scales, relative to other objects in the scene.

The output surface shown in Figure 9(c) shows three manmade objects (motor vehicles) clearly accentuated (pixel values above T_1) relative to the unknown cluttered environment. It is an achievement, given that no prior information is used about the materials composing the clutter background,

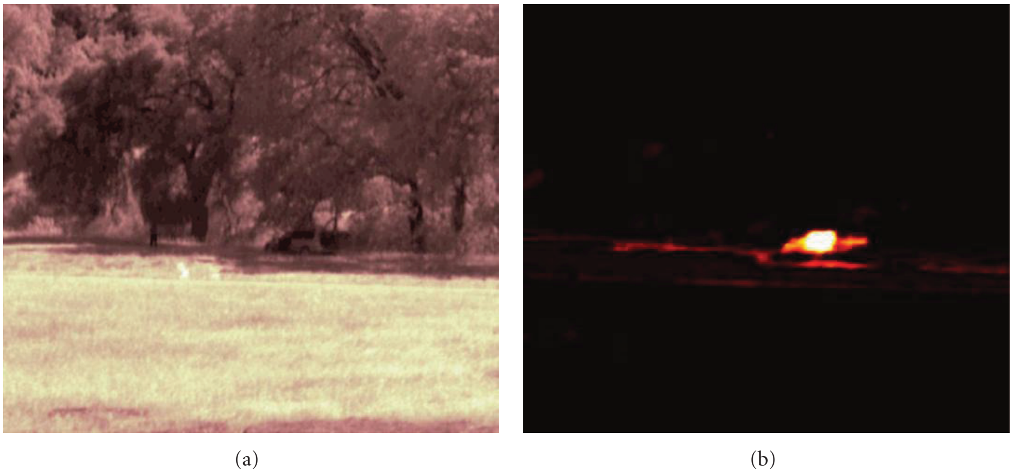


Figure 10: Results testing Cube 3 (a) and corresponding output surface (b). Parameters were set to ($q = 0.10$; $T_1 = T(20)$; $N = 22$; $M = 40$). A motor vehicle is parked in tree shades—around center of the scene. Using the available ground truth information of the scene, the white clusters cover about 73% of the motor vehicle's spatial area and no false alarms.

or about whether targets are present in the scene, or about targets' scales relative to other objects in the imagery. But notice in Figure 9 that the standing person in the scene center is not detected, possibly because the window size might be too large and/or there must have some materials in that background (randomly selected) spectrally similar to the materials representing that person (e.g., pants, shirt, skin). Figures 10 and 11 show additional results.

Figures 10 and 11 show results for Cube 3 and Cube 4, respective; both HS cubes particularly represent difficult cases of clutter suppression. Parameters were also set to ($q = 0.10$; $T_1 = T(20)$; $N = 22$; $M = 40$) for both HS cubes.

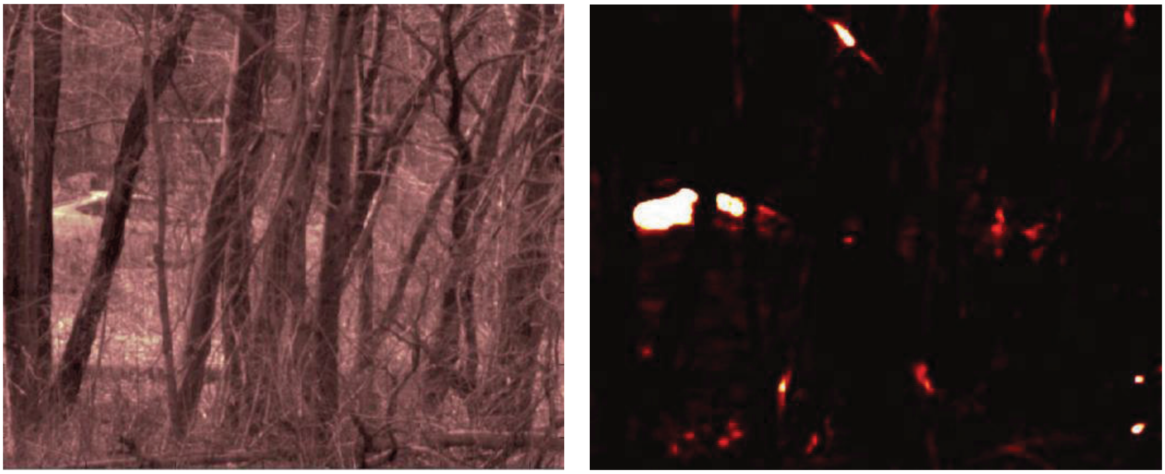
The guideline described in Section 3.3 for setting parameters also worked very well for both complex scenes depicted in Figures 10 and 11. Both output surfaces clearly accentuates the presence of a motor vehicle—one in tree shades (center in Cube 3) and another motor vehicle parked behind a heavily cluttered environment (center left hand side in Cube 4); see *white* pixels, or pixel values greater than or equal to T_1, in both output surfaces in Figures 10 and 11. Setting $T_1 = T(20)$, in essence, means setting the cutoff threshold at 20 standard deviations above the mean of distribution $\tilde{g}(z)$ in (A.10).

Using the available ground truth information of the scenes in Figure 9 through Figure 11, quantitative comparative performances were obtained via receiver's operating characteristic (ROC) curves (vertical axis show Pd for probability of detection, and horizontal axis shows Pfa for probability of false alarms) for some of the anomaly detectors mentioned in Section 1. The data cubes were processed using the global RX, k means, GMM, and QG-SemiP. In essence we used the k mean and GMM in place of SemiP in the context of the parallel random sample, but had to take into consideration some of the inherent constraints of these methods. For instance, for k means and GMM, we recorded ROC curves for parameter N set to $N = 3, 5, 10, 20,$ and 50 and repetition parameter $M = 50$, while for QG-SemiP N was set to 20, 50, 75, and 100 with $M = 50$. Global RX estimates the mean and covariance from the entire data cube, as discussed in Section 1. Figure 12 shows performance of these detectors using ABC to label QG-SemiP, global to label global RX, KM to label k means, and GMM.

Although the parameter values for k means and GMM start at a lower value than QG-SemiP, Figure 12 shows that at lower N the detectors k means and GMM actually perform much better than for N set at higher values, as Nmay be interpreted by these algorithms as the number of distinct classes in the scene. Such performance degradation occurs with large N because the spatial area that corresponds to each individual N block of data is now smaller and samples of the targets, required by the algorithm, need to be included into one of the classes. This outcome contaminated the distribution of the background clutter forcing it to be closer to the distribution of the targets, resulting in performance degradation. The global RX performed reasonably well, as expected since the scene in Figure 9 is relatively less complex. The QG-SemiP detector, also as expected, improved performance as N increased.

Performance of these algorithms on the scenes in Figures 10 and 11 are shown in Figures 13 and 14, respectively. Performance degradation of k means, GMM, and global RX are evident from Figure 13, since the target is on tree shades. QG-SemiP performs well for $N > 20$. In Figure 14, the performance of the k means performed poorly at $N = 20$, since the target is partially blocked by tree trunks, while GMM performed poorly at $N = 3$ and $N = 5$ and was unable to converge at higher N values. The global RX surprisingly worked reasonably well, but completely underperforming the QG-SemiP detector.

4.2. Nadir Looking Imagery. For the nadir looking imagery, Cube 1 in Figure 1 (top), ROC curves are also used as a means to quantitatively compare five anomaly detection approaches: local RX, local KRX, local FLD, local SemiP, and QG-SemiP; see Section 1.



(a)

(b)

Figure 11: Results testing Cube 4 (a) and corresponding output surface (b). Parameters were set to ($q = 0.10$; $T_1 = T(20)$; $N = 22$; $M = 40$). A motor vehicle is shown on the left hand side, between top and bottom, behind tree trunks—a sport car. Cube 4 exemplifies a hard case of autonomous clutter suppression. Using the available ground truth information of the scene, the white clusters cover about 44% of the motor vehicle's spatial area and less than 2% of false alarms.

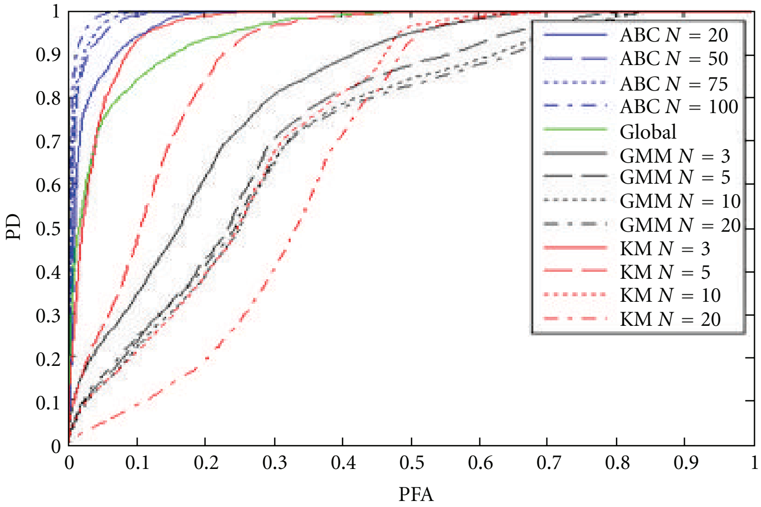


Figure 12: ROC curve performance testing QG-SemiP (ABC), global RX (Global), k means (KM), and GMM detectors on scene shown in Figure 9.

Local anomaly detectors process small ($n \times n$) windows of the HS data cube $\mathbf{X}$, where all the $\mathbf{x}_{rc}$ ($r = 1, \ldots, R$; $c = 1, \ldots, C$) in $\mathbf{X}$ are used; modeling is only done at the level of the $n \times n$ windows, where $n \ll R$ and $n \ll C$ ($\ll$ denoting *many orders of magnitude smaller than*); at the level of the pixel area surrounding these windows. Blocks of data ($n \times n$ windows) that are spectrally different from pixels surrounding them score high using an effective detector in contrast to blocks of data that are not spectrally different from their surrounding pixels. After the detector scores the entire $\mathbf{X}$, it yields a 2-dim surface $\mathbf{Z}$ [$a(R-n-1)\times(C-n-1)$ array of scalars], where a cutoff threshold is then compared to the pixel values in $\mathbf{Z}$. Pixels having values greater than the threshold are labeled local anomalies (notice that the SemiP detector will be used in both local and quasi-global versions).

As described in Section 2.1, the set of *14* ground vehicles near the treeline in Cube 1 (Figure 1) constitutes the target set, but, since anomaly detectors are not designed to detect a particular target set, the meaning of false alarms is not absolutely clear in this context. For instance, a genuine local anomaly not belonging to the target set would be incorrectly labeled as a false alarm. Nevertheless, it does add some value to our analysis to compare detections of targets versus nontargets among the different algorithmic ap-proaches.

Figure 15 shows the ROC curves produced by the output of the five algorithms testing Cube 1 for local or scene

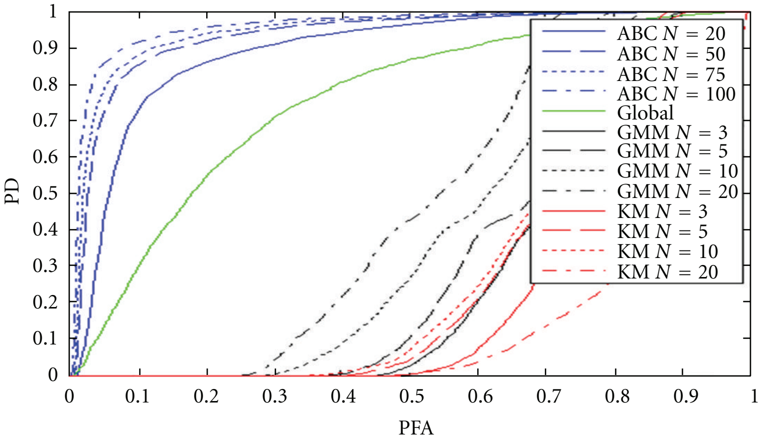


Figure 13: ROC curve performance testing QG-SemiP (ABC), global RX (global), k means (KM), and GMM detectors on scene shown in Figure 10.

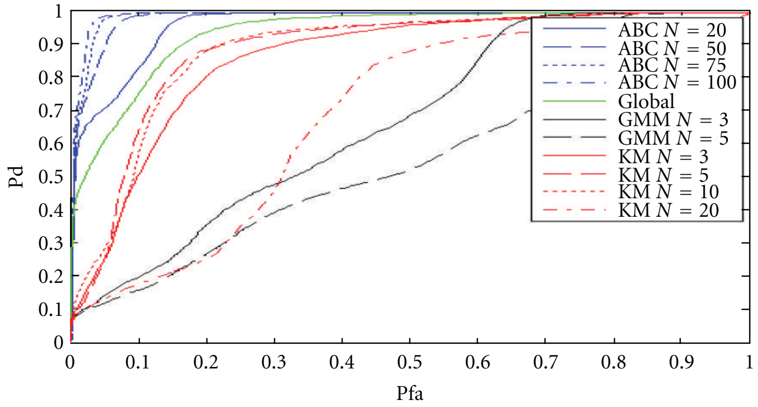


Figure 14: ROC curve performance testing QG-SemiP (ABC), global RX (global), k means (KM), and GMM detectors on scene shown in Figure 11.

anomalies. Detection performance was measured using the ground truth information for the HYDICE imagery. We used the coordinates of all the rectangular target regions and their shadows to represent the ground truth target set. As it can be readily assessed from Figure 15, the local SemiP and QG-SemiP anomaly detection approaches outperform the other three techniques on the tested scene, followed by KRX's performance. The significant display of performance shown in Figure 15 by the semiparamtric algorithms can be further appreciated by taking a closer look at the output surfaces produced by all five detectors, as they show evidences of candidate local and scene anomalies. The intensity of local peaks shown in Figure 16 reflects the strength of the detectors' evidences. Figure 16 shows that the surfaces produced by FLD, RX, and KRX detectors are expected to be significantly more sensitive (producing *false alarms*) to changing cutoff thresholds then the ones produced by the local SemiP and QG-SemiP approaches.

Spatial areas shown in Cube 1 containing the presence of clutter mixtures (e.g., edge of terrain, edge of tree clusters), where FLD, RX, and KRX yield a high number of potential false alarms (false anomalies), are suppressed by the SemiP approach, local, and quasi-global. The reason for this suppression is that, as part of the overall comparison strategy, the semiparametric model combines both the test and reference samples in order to estimate the underlying PDF of the reference sample, as shown by simulation earlier. As such, the semiparametric test statistic ensures that a component of a mixture (e.g., shadow) will not be detected as a local anomaly when it is tested against the mixture itself (e.g., trees and shadows). Performances of such cases are represented in Figure 16 in the form of softer anomalies (significantly less-dominant peaks) in the local SemiP's and QG-SemiP's output surfaces. It is evident from Figure 16 that both versions of the SemiP detectors perform remarkably well accentuating the presence of dominant local or scene anomalies (e.g., targets) from softer anomalies (e.g., transitions of distinct regions). The natural ability of the semiparametric model to manage spectral mixture can best explain the local SemiP and QG-SemiP superior ROC-curve performances shown in Figure 15.

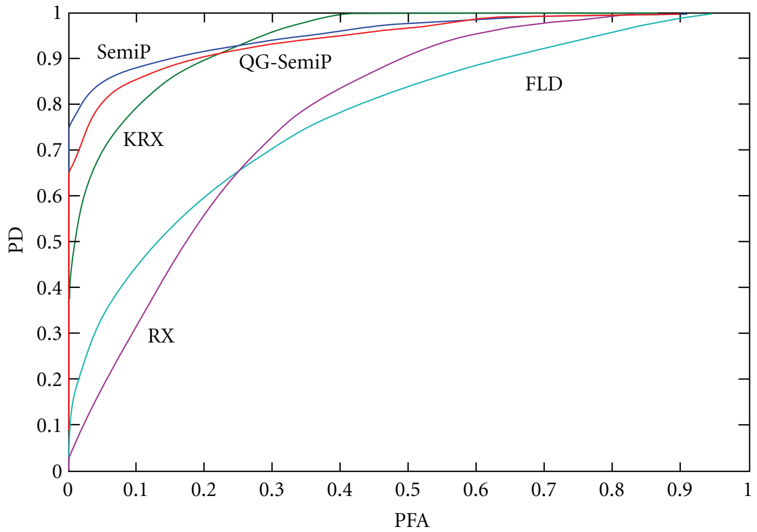


Figure 15: ROC curves for the nadir looking imagery (Cube 1) data scene shown in Figure 1. The semiparametric method, being used in both local and quasiglobal versions, are noticeably less sensitive to different cutoff thresholds; their performances almost achieve an ideal ROC curve for that scene, that is, a step function starting at point (PFA $= 0$, PD $= 1$).

5. Conclusion

This paper introduced an adaptive quasi-global, semiparametric anomaly detection algorithm and evaluated the approach using real HS imagery, where targets (manmade objects) are found in difficult natural clutter backgrounds viewed from two different perspectives—nadir and forward looking. The algorithm features a semiparametric test statistic, which has been recently found to be robust against spectral mixture, and applies random sampling of the imagery to test for anomalies. Random sampling and testing are repeated a number of times in order to mitigate the probability of contamination (spectral samples of candidate targets being sampled and used as clutter reference samples). As such, the algorithm requires no prior information (e.g., a spectral library of the clutter background and/or target, target size, or shape). The algorithm is free from training requirements. We found that the semiparametric model has a natural ability to handle mixtures, although an exhaustive survey of the literature reveals that this fact has never been noticed before by practitioners of the model in other fields of study (e.g., biotechnology).

The repeated sampling and testing procedure was modeled by the binomial family of distributions, where the only target related parameter q (the upper bound proportion of target pixels potentially covering the spatial area of the imagery) is robust—thus invariant—to different sizes and shapes of targets, number of targets present in the scene, target aspect angle, partially obscured targets, or sensor viewing perspective. The paper also discussed how other parameters (N, the total number of sampled data blocks to take from the HS imagery, and M, the number of process repetitions) can be automatically set using a simple guideline.

The algorithm fuses intermediate results through the application of minimum order statistics and logic *OR* operation. The paper presented the algorithm's asymptotic behavior under the null hypothesis, when either the null or the alternative hypothesis is true, for the two-sample test case and the multi-sample test case, where the cumulative probability of the algorithm making mistakes was derived. Using the cumulative probability, a cutoff threshold can be determined from a user specified probability of error. This is a desired feature giving the user some control of predicted errors.

The inherent challenges in adequately modeling spectral variability of targets, while managing spectral variability between targets, have prompted the introduction of a more robust family of algorithms that attempts to detect targets as being anomalous to an unknown natural clutter background. This paper presented an approach that inherently addresses many of the problems and issues pertaining to anomaly detection applications. The advantages of using anomaly detection algorithms have been discussed in this paper; however, it should be emphasized that targets would not be detected as specific manmade objects; they would be detected as anomalies. This is a limitation.

Appendix

Asymptotic Behavior of the Quasiglobal Semiparametric Algorithm

This section shows an analytical asymptotic analysis of the quasi-global semiparametric anomaly detection algorithm. In particular, we would like to investigate the algorithm's cumulative probability of rejecting the null hypothesis, when either the null or the alternative hypothesis is true; that is, the algorithm's probability of making mistakes. We will look first at the asymptotic behavior of the two sample test case, where

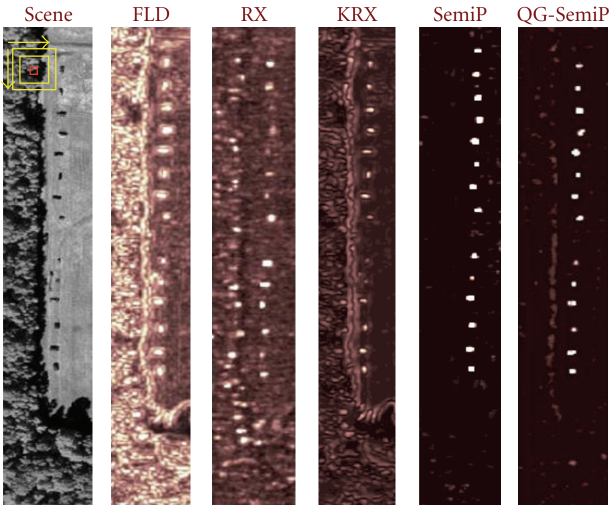


Figure 16: Detection algorithms' output surfaces for Cube 1 (far left). The intensity of local peaks reflects the strength of evidences as *seen* by different anomaly detection approaches. Boundary issues were ignored in this test; surfaces were magnified to about the size of the original image only for the purpose of visual comparison. FLD, RX, KRX, and SemiP performed local anomaly detection by testing spectra within a testing window (red square shown in the scene display—far left, top) to spectra surrounding the testing window (outer window bounded by yellow lines). QG-SemiP performed global anomaly detection, as presented in this paper.

the detector tests a reference sample against a test sample; then we will look at the multisample test case, which uses order statistic to reduce N results to a single result.

Two-Sample Test (2ST). In order to declare an anomaly, a decision threshold T must be chosen; hopefully, separating without errors the null and the alternative hypotheses in some decision space. In the paper's context, values of Z_{SemiP} in (23) greater than T are automatically labeled as anomalies. And since, in real world applications, decision errors are unavoidable, we would like to know whether the asymptotic behaviors of these errors can be determined, and whether they are favorable. The *power function* can provide those answers for the two-sample test (2ST) case. Using (13), the power function of the semiparmetric test statistic for the two-sample test (2ST) case is as follows:

$$\psi(\beta) = \begin{cases} P_{\beta=0}(Z_{\text{SemiP}} > T) \\ P_{\beta\neq 0}(Z_{\text{SemiP}} > T). \end{cases} \tag{A.1}$$

In essence, the power function ψ yields the cumulative probability P of rejecting the null hypothesis H_0 in (13) when either H_0 ($\beta = 0$) or the alternative H_1 ($\beta \neq 0$) is true. This rejection region is $Z_{\text{SemiP}} > T$, where Z_{SemiP} is defined in (23) and T is the decision threshold. Notice in (A.1) that ψ under H_0 corresponds to the well known type I error, or the probability of missing (i.e., the probability of rejecting H_0, given that H_0 is true) and that ψ under H_1 corresponds to the complement of the type II error, or probability of false alarms (i.e., one minus the probability of rejecting H_1, given that H_1 is true). The type I and type II errors constitute the only error types encountered in the context of our discussion. In the ideal case, ψ yields 0.0 when H_0 is true and 1.0 when H_1 is true. Except in trivial situations, this ideal cannot be attained. So, one of our goals is to show that ψ tends in probability to ε (a scalar controlled by the user), when H_0 is true, and that ψ tends in probability to 1.0, when the alternative hypothesis H_1 is true.

If H_0 in (13) is true, the semiparametric detector has the asymptotic behavior shown in (23), and the type-I error is readily obtained by the following:

$$P_{\beta=0}(Z_{\text{SemiP}} > T) \xrightarrow[n\to\infty]{} P(\xi > T) = \varepsilon, \tag{A.2}$$

where ξ is a Chi square distributed random variable with 1 degree of freedom, Z_{SemiP} as defined in (23), ε and T are nonnegative real values.

Setting $\psi(\beta) = P_{\beta=0}(Z_{\text{SemiP}} > T)$, ψ is indeed an asymptotic size ε test, which is controlled by the user.

Now consider an alternative parameter value, such that $\beta \neq 0$, and let ζ^2 be the true variance in (21) and $\hat{\zeta}^2$ be a estimator of ζ^2, or

$$\begin{aligned} \zeta^2 &= \frac{\rho^{-1}(1+\rho)^2}{\nu^2}, \\ \hat{\zeta}^2 &= \frac{\rho^{-1}(1+\rho)^2}{\hat{\nu}^2}, \end{aligned} \tag{A.3}$$

where $\rho = n_1/n_0$ and $\hat{\nu}^2$ is defined in (22).

From (23), we can now write

$$Z_{\text{SemiP}} = \left\{ \left[\underbrace{\left(\frac{\hat{\beta} - \beta}{\sqrt{\zeta^2/n}} \right)}_{A} + \underbrace{\left(\frac{\beta}{\sqrt{\zeta^2/n}} \right)}_{B} \right]^2 \underbrace{\left(\frac{\zeta^2}{\hat{\zeta}^2} \right)}_{C} \right\}. \tag{A.4}$$

Notice in (A.4) that, as n_1 and n_0 go to $+\infty$, $\rho = n_1/n_0$ tends to 1 and ζ^2 tends to ζ^2 for $\nu^2 > 0$. According to (21), the term A in (A.4) converges in distribution to the standard Normal, $N(0,1)$, as n_1 and n_0 (hence, n) go to $+\infty$, no matter what the values of β or ζ^2 are. Note also that the term B converges to $+\infty$ or $-\infty$ in probability, as n goes to $+\infty$, depending on whether β is positive or negative. Since the estimator $\hat{\tilde{g}}_0$ in (38) has been shown [21] to be biased, as n_1 and n_0 go to $+\infty$, $\hat{\zeta}^2$ tends to a constant, see definition of $\hat{\nu}^2$ in (22); leading the term C also to a constant, no matter what values of ζ^2 is. Thus, Z_{SemiP} converges to $+\infty$ in probability and

$$P_{\beta\neq 0}(\text{reject } H_0) = P_{\beta\neq 0}(Z_{\text{SemiP}} > T) \xrightarrow[n\to\infty]{} 1. \tag{A.5}$$

In this way, the semiparametric test statistic shown in (23) also has the properties of asymptotic size ε and asymptotic power 1, which is highly desired.

Multisample Test (mST). The discussions in Sections 2.3 and 4.1 ensure that the output of the semiparametric 2ST has two asymptotic outcomes: $Z_{\text{SemiP}} \xrightarrow[n\to\infty]{} \chi_1^2$ in distribution, if H_0 in (13) is true, or $Z_{\text{SemiP}} \xrightarrow[n\to\infty]{} +\infty$ in probability, if H_1 is true.

Using results leading to (13) and the order statistics $\tilde{Z}_{\text{SemiP}}^{(ij)} = \min_{1\leq f\leq N} Z_{\text{SemiP}}^{(ij)(f)}$ in (34), for the multi-sample test (mST) case, we propose the following null H_2 and alternative H_3 hypotheses

$$\begin{gathered} H_2 : \text{at least one } \left\{\beta^{(f)}\right\}_{f=1}^{N} = 0, \\ H_3 : \text{all } \left\{\beta^{(f)}\right\}_{f=1}^{N} \neq 0, \end{gathered} \tag{A.6}$$

where $\beta^{(f)}$ is the true logistic function parameter—see (12)—corresponding to the fth randomly selected block of data from a HS data cube $\mathbf{X}$.

Now, consider the following: for a given spatial location in $\mathbf{X}$, let $Z_{\text{SemiP}}^{(f)}$ be the semiparametric detector's output for the fth block of data, and assume, without loss of generality, that each one of the first L outputs in the independent sequence of results ($1 \leq L \leq N$, where N is the total number of randomly selected blocks of data in $\mathbf{X}$) has the asymptotic chi-square behavior shown in (23), and that each one of the remainder results has the asymptotic behavior tending to $+\infty$, or

$$\tilde{Z}_{\text{SemiP}}\Big|_{H_2} = \min\left\{\begin{matrix} Z_{\text{SemiP}}^{(1)} \xrightarrow{n\to\infty} \chi_1^2 \\ Z_{\text{SemiP}}^{(2)} \xrightarrow{n\to\infty} \chi_1^2 \\ \vdots \\ Z_{\text{SemiP}}^{(L)} \xrightarrow{n\to\infty} \chi_1^2 \\ Z_{\text{SemiP}}^{(L+1)} \xrightarrow{n\to\infty} +\infty \\ \vdots \\ Z_{\text{SemiP}}^{(N)} \xrightarrow{n\to\infty} +\infty \end{matrix}\right\}. \tag{A.7}$$

Under the null hypothesis H_2 in (A.6), $\tilde{Z}_{\text{SemiP}}$ in (A.7) is bounded because, as $n \to \infty$, $\tilde{Z}_{\text{SemiP}}$ will converge in law to the distribution of the lowest order statistics. (The order statistics of a random sample $Z_1,\ldots,Z_N$ are the sample values placed in ascending order. They are often denoted by $Z_{(1)},\ldots,Z_{(N)}$, where $Z_{(1)} = \min_{f\leq f\leq N} X_f$ and $Z_{(N)} = \max_{f\leq f\leq N} X_f$.) To attain an approximation of the type I error using (A.7), we first ignore all the components in (A.7) that converge in probability to $+\infty$, then we consider only the components that converge in distribution, that is, $(Z_{\text{SemiP}}^{(1)}, Z_{\text{SemiP}}^{(2)},\ldots,Z_{\text{SemiP}}^{(L)})$. The distribution of

$$Z_{\text{SemiP}(1)} = \min_{f\leq f\leq L} Z_{\text{SemiP}}^{(f)} \tag{A.8}$$

from the culled sequence can be attained with the application of Theorem 1.

Theorem 1. *Let $X_{(1)},\ldots,X_{(n)}$ denote the order statistics of a random sample from a continuous population with cumulative distribution function (cdf) $F(x)$ and pdf $f(x)$. Then the pdf of $X_{(j)}$ is*

$$\tilde{f}(x) = \frac{n!}{(j-1)!(n-j)!} f(x)[F(x)]^{j-1}[1-F(x)]^{n-j}, \tag{A.9}$$

where $(\cdot)!$ denotes the factorial operator.

The proof of Theorem 1 can be found in [28].

Using Theorem 1 with $j = 1$ and $n = L$, the pdf of $\tilde{Z}_{\text{SemiP}} = Z_{\text{SemiP}(1)}$ under H_2 in (A.6) is

$$\tilde{g}(z) = Lg(z)[1-G(z)]^{L-1}, \tag{A.10}$$

where $g(z)$ is the Chi square pdf with 1 degree of freedom and $G(z)$ is the corresponding cdf.

Denote the kth logistic function parameter $\beta^{(k)}$ in (A.6) to correspond to the one of the minimum order statistics $Z_{\text{SemiP}(1)}$. As the sample size increases in $Z_{\text{SemiP}(1)}$, that is, $n = n^{(k)} \to \infty$, the probability of rejecting the null hypothesis H_2 in (A.6), when $\beta^{(k)} = 0$, converges to

$$\hat{\psi}\left[\beta^{(k)}\right] = P_{\beta^{(k)}=0}\left(\tilde{Z}_{\text{SemiP}} > T_1\right) \xrightarrow[n\to\infty]{} P(\xi > T_1) = \varepsilon_1, \tag{A.11}$$

where ξ is a random variable distributed by $\tilde{g}(z)$, as defined in (A.10); T_1 a nonnegative real value; ε_1 is a positive real value, controlled by the user.

The variable $\hat{\psi}$ in (A.11) is the type I error under H_2 for the mST case, and it is indeed an asymptotically size ε_1 test.

Now consider the alternative hypothesis H_3 in (A.6), where all $\{\beta^{(f)}\}_{f=1}^{N} \neq 0$. From (A.7) one can write

$$\tilde{Z}_{\text{SemiP}}\Big|_{H_3} = \min\left\{\begin{matrix} Z_{\text{SemiP}}^{(1)} \xrightarrow{n\to\infty} +\infty \\ \vdots \\ Z_{\text{SemiP}}^{(N)} \xrightarrow{n\to\infty} +\infty \end{matrix}\right\}. \tag{A.12}$$

From (A.12), $\tilde{Z}_{\text{SemiP}}$ will converge in probability to $+\infty$, hence, the probability P of rejecting the null hypothesis H_2, given that H_3 is true, tends to 1.0, or

$$P_{\beta^{(k)}\neq 0}(\text{rejecting } H_2) = P_{\beta^{(k)}\neq 0}(Z_{\text{SemiP}(1)} > T_1) \xrightarrow[n\to\infty]{} 1. \tag{A.13}$$

In this way, the quasi-global semiparametric anomaly detection algorithm has the desired properties of asymptotic size ε_1, which is controlled by the user, and asymptotic power 1.0.

Notations

Bold upper case letters may denote a data cube (3 dimensions) or a matrix (e.g., $\mathbf{X}, \mathbf{W}_0$), where the specific case in use is defined in the text.

Lower cases letters denote vectors (bold) or sequences (not bold) (e.g., $\mathbf{x}$, $x_1 = (x_{11}, \dots, x_{1n_1})$).

PDF or pdf: Probability density function.

IID or iid: Independent and identically distributed.

$\mathbf{R}^{d_1 \times d_2}$-$d_1$ by d_2 dimensional set of real numbers.

$\in$ denotes set belonging

HS: Hyperspectral

$\mathbf{X}$: Observed hyperspectral data cube with dimensions of R rows, C columns, and K bands.

$\mathbf{x}_{rc}$: Observed spectrum contained in $\mathbf{X}$ with spatial indexes $(r = 1, \dots, R)$ and $(c = 1, \dots, C)$.

A slideing $n \times n$ window is a 3-dim subset of $\mathbf{X}$, containing $n \cdot n$ spectra.

$\mathbf{W}_1 \in \mathbf{R}^{K \times n_1}$: A matrix representing a hyperspectral sample being observed from a sliding $n \times n$ window in $\mathbf{X}$ (also referred to herein as a test sample); this is a rearranged version of a 3-dim subdata cube, where vertical direction is the dimension K of bands and the horizontal direction is the dimension of countable samples with sample size $n_1 = n^2$.

$\mathbf{y}_{1h} \in \mathbf{R}^K$ $(h = 1, \dots, n_1)$: An observed spectrum of K bands contained in $\mathbf{W}_1$.

$g_1(\mathbf{y} | \boldsymbol{\theta})$: Multivariate joint PDF of $\mathbf{y}_{11}, \dots, \mathbf{y}_{1n_1}$

$\mathbf{W}_0 \in \mathbf{R}^{K \times n_0}$: A matrix representing a hyperspectral sample labled as a reference sample of sample size n_0, having the same specifications of $\mathbf{W}_1$ except perhaps the sample size (n_0 may be different from n_1).

$\mathbf{y}_{0h} \in \mathbf{R}^K (h = 1, \dots, n_0)$: An observed spectrum of K bands contained in $\mathbf{W}_0$.

$g_0(\mathbf{y} | \boldsymbol{\theta})$: Multivariate joint PDF of $\mathbf{y}_{01}, \dots, \mathbf{y}_{0n_0}$.

$\nabla_0 \in \mathbf{R}^{(K-1) \times n_0}$: Output from differentiating $\mathbf{W}_0$.

$\nabla_1 \in \mathbf{R}^{(K-1) \times n_1}$: Output from differentiating $\mathbf{W}_1$.

$x_0 = (x_{01}, x_{02}, \dots, x_{0n_0})$: Univariate sequence used as the reference sample.

$x_1 = (x_{11}, x_{12}, \dots, x_{1n_1})$: Univariate sequence used as the test sample.

$g_0(x)$: Univariate PDF labeled as reference.

$g_1(x)$: Univariate PDF labeled as test

$t = (x_{11}, \dots, x_{1n_1}, x_{01}, \dots, x_{0n_0}) \equiv (t_1, \dots, t_n)$: Sample concatenation, combining samples.

$\widetilde{g}_0(t)$: estimator of $g_0(x)$.

$\widehat{\widetilde{g}}_0(t)$: estimator of $\widetilde{g}_0(t)$.

H_i: Statistical hypothesis i.

Z_{SemiP}: Univariate output of the semiparametric detector.

P: Cumulative probability function, using the binomial family of PDFs as base PDF.

N: The number of randomly selected blocks of data used to represent background objects.

Trial (or process): Take N random blocks of data from the data cube under test, label them as reference background objects, and—using the semiparametric detector and a sliding window across the data cube—test the entire data cube against the same set of N reference blocks of data.

M: The number of trials (repetitions or parallel processes).

$P_g(m \geq 1)$: Cumulative probability of contamination, that is, probability of labeling a randomly selected target sample as a background sample at the gth trial or process.

$\widetilde{P}$: Overall cumulative probability that all of the trials (or processes) are contaminated with at least a contaminated sample from the randomly selected set of reference samples $\{\mathbf{W}_0^{(f)}\}_{f=1}^{N}$.

$\widetilde{Z}_{\text{SemiP}}^{(ij)}$: Retains the lowest order statistics from a set of N semiparametric detector's results.

$\widetilde{\mathbf{Z}}_{\text{SemiP}}^{(g)}$: The 2-dimensional output surface, consisting of $\widetilde{Z}_{\text{SemiP}}^{(ij)}$ values, from the gth trial.

T_1: Adaptive cutoff threshold for $\widetilde{\mathbf{Z}}_{\text{SemiP}}^{(g)}$.

$\mathbf{Z}_{\text{SemiP}}$: A final binary 2-dimensional output surface of the quasi-global semiparametric detector.

References

[1] R. Shcowengerdt, *Remote Sensing, Models and Methods for Image Processing*, Academic, San Diego, Calif, USA, 2nd edition, 1997.

[2] D. G. Manolakis and G. Shaw, "Detection algorithms for hyperspectral imaging applications," *IEEE Signal Processing Magazine*, vol. 19, no. 1, pp. 29–43, 2002.

[3] D. Manolakis, "Taxonomy of detection algorithms for hyperspectral imaging applications," *Optical Engineering*, vol. 44, no. 6, pp. 1–11, 2005.

[4] D. G. Manolakis and G. Shaw, "Hyperspectral image processing for automatic target detection applications," *IEEE Signal Processing Magazine*, vol. 14, pp. 79–116, 2003.

[5] P. H. Suen, G. Healey, and D. Slater, "The impact of viewing geometry on material discriminability in hyperspectral images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 7, pp. 1352–1359, 2001.

[6] G. Healey, "Models and methods for automated material identification in hyperspectral imagery acquired under unknown illumination and atmospheric conditions," *IEEE Transactions*

on Geoscience and Remote Sensing, vol. 37, no. 6, pp. 2706–2717, 1999.

[7] X. Yu, L. E. Hoff, I. S. Reed, A. M. Chen, and L. B. Stotts, "Automatic target detection and recognition in multiband imagery: a unified ML detection and estimation approach," *IEEE Transactions on Image Processing*, vol. 6, no. 1, pp. 143–156, 1997.

[8] H. Kwon and N. M. Nasrabadi, "Kernel RX-algorithm: a nonlinear anomaly detector for hyperspectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 2, pp. 388–397, 2005.

[9] S. Matteoli, M. Diani, and G. Corsini, "Different approaches for improved covariance matrix estimation in hyperspectral anomaly detection," in *Riunione Annuale dell'Associazione Gruppo Nazionale Telecomunicazioni e Tecnologie dell'Informazione*, pp. 1–8, 2009.

[10] A. Banerjee, P. Burlina, and C. Diehl, "A support vector method for anomaly detection in hyperspectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 8, pp. 2282–2291, 2006.

[11] Y. Tan and J. Wang, "A support vector machine with a hybrid kernel and minimal vapnik-chervonenkis dimension," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 4, pp. 385–395, 2004.

[12] P. Gurram and H. Kwon, "Support-vector-based hyperspectral anomaly detection using optimized kernel parameters," *IEEE Geoscience and Remote Sensing Letters*, vol. 8, no. 6, pp. 1060–1064, 2011.

[13] S. Khazai, S. Homayouni, A. Safari, and B. Mojaradi, "Anomaly detection in hyperspectral images based on an adaptive support vector method," *IEEE Geoscience and Remote Sensing Letters*, vol. 8, no. 4, pp. 646–650, 2011.

[14] D. W. J. Stein, S. G. Beaven, L. E. Hoff, E. M. Winter, A. P. Schaum, and A. D. Stocker, "Anomaly detection from hyperspectral imagery," *IEEE Signal Processing Magazine*, vol. 19, no. 1, pp. 58–69, 2002.

[15] P. Hytla, R. C. Hardie, M. T. Eismann, and J. Meola, "Anomaly detection in hyperspectral imagery: a comparison of methods using seasonal data," in *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XIII*, vol. 6565 of *Proceedings of SPIE*, September 2007.

[16] S. Matteoli, F. Carnesecchi, M. Diani, G. Corsini, and L. Chiarantini, "Comparative analysis of hyperspectral anomaly detection strategies on a new high spatial and spectral resolution data set," in *Image and Signal Processing for Remote Sensing XIII*, vol. 6748 of *Proceedings of SPIE*, September 2007.

[17] H. Kwon, S. Z. Der, and N. M. Nasrabadi, "Adaptive anomaly detection using subspace separation for hyperspectral imagery," *Optical Engineering*, vol. 42, no. 11, pp. 3342–3351, 2003.

[18] A. Stocker, "Stochastic expectation maximization (SEM) algorithm," in *Proceedings of the DARPA Adaptive Spectral Reconnaissance Algorithm Workshop*, 1999.

[19] P. Masson and W. Pieczynski, "SEM algorithm and unsupervised statistical segmentation of satellite images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 31, no. 3, pp. 618–633, 1993.

[20] D. Rosario, *Algorithm development for hyperspectral anomaly detection [Ph.D. dissertation]*, The University of Maryland, College Park, Md, USA, 2008.

[21] J. Qin and B. Zhang, "A goodness-of-fit test for logistic regression models based on case-control data," *Biometrika*, vol. 84, no. 3, pp. 609–618, 1997.

[22] K. Fokianos, J. Qin, B. Kedem, and D. A. Short, "Semiparametric approach to the one-way layout," *Technometrics*, vol. 43, no. 1, pp. 56–65, 2001.

[23] D. Rosario, "A semiparametric approach using the discriminant metric SAM (spectral angle mapper)," in *Automatic Target Recognition XIV*, vol. 5426 of *Proceedings of SPIE*, pp. 58–66, September 2004.

[24] B. Kedem, *Time Series Analysis by Higher Order Crossings*, IEEE Press, New York, NY, USA, 1994.

[25] E. L. Lehmann, *Testing Statistical Hypotheses*, Chapman & Hall, New York, NY, USA, 2nd edition, 1993.

[26] J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright, "Convergence properties of the Nelder-Mead simplex method in low dimensions," *SIAM Journal on Optimization*, vol. 9, no. 1, pp. 112–147, 1999.

[27] A. M. Law and W. D. Kelton, *Simulation Modeling and Analysis*, McGraw -Hill, Boston, Mass, USA, 3rd edition, 2000.

[28] G. Casella and R. L. Berger, *Statistical Inference*, Dexbury Press, Belmont, Calif, USA, 1990.

Hindawi Publishing Corporation
Journal of Electrical and Computer Engineering
Volume 2012, Article ID 162106, 16 pages
doi:10.1155/2012/162106

Research Article

Hyperspectral Anomaly Detection: Comparative Evaluation in Scenes with Diverse Complexity

Dirk Borghys,[1] Ingebjørg Kåsen,[2] Véronique Achard,[3] and Christiaan Perneel[4]

[1] *Department CISS, Royal Military Academy, 2007 Brussels, Belgium*
[2] *Land and Air Systems Division, Norwegian Defence Research Establishment (FFI), 2007 Kjeller, Norway*
[3] *Theoretical and Applied Optics Department, French Aerospace Laboratory (ONERA), FR-31055 Toulouse Cedex 4, France*
[4] *Department of Mathematics, Royal Military Academy, Brussels, Belgium*

Correspondence should be addressed to Dirk Borghys, dirk.borghys@gmail.com

Received 24 May 2012; Revised 22 August 2012; Accepted 9 September 2012

Academic Editor: Xiaofei Hu

Anomaly detection (AD) in hyperspectral data has received a lot of attention for various applications. The aim of anomaly detection is to detect pixels in the hyperspectral data cube whose spectra differ significantly from the background spectra. Many anomaly detectors have been proposed in the literature. They differ in the way the background is characterized and in the method used for determining the difference between the current pixel and the background. The most well-known anomaly detector is the RX detector that calculates the Mahalanobis distance between the pixel under test (PUT) and the background. Global RX characterizes the background of the complete scene by a single multivariate normal probability density function. In many cases, this model is not appropriate for describing the background. For that reason a variety of other anomaly detection methods have been developed. This paper examines three classes of anomaly detectors: subspace methods, local methods, and segmentation-based methods. Representative examples of each class are chosen and applied on a set of hyperspectral data with diverse complexity. The results are evaluated and compared.

1. Introduction

Many types of anomaly detectors have been proposed in literature [1, 2]. The most frequently used anomaly detector is the (spectral only version of the) Reed-Xiaoli (RX) detector [3] that is often used as a benchmark to which other methods are compared. The RX detector characterizes the background by its spectral mean vector μ_B and covariance matrix Σ_B. The actual detector calculates the Mahalanobis distance between the pixel under test r and the background as follows:

$$D_{\mathrm{RX}} = (r - \mu_B)^T \Sigma_B^{-1} (r - \mu_B). \tag{1}$$

The global RX detector characterizes the background of the complete scene by a single multivariate normal probability density function (pdf). In many scenes, this model is not adequate. For that reason, several variations of the global RX detector have been proposed in literature [1, 2, 4–12]. They can be sub-divided into three classes: subspace methods, local methods, and segmentation-based methods. In complex scenes the latter category was shown to be very effective and several segmentation-based anomaly detectors (SBAD), not necessarily based on RX, have recently been proposed [13–20]. The aim of the current paper is to compare the results obtained by different types of anomaly detectors in scenes characterized by different types of background. In particular, two rural scenes with subpixel anomalies, a rural scene with some of the targets in shadow, and an urban scene were considered. Representative examples of each of the three previously mentioned classes of anomaly detectors were included in the comparison. In previous work [21], we noted the importance of data reduction and preprocessing on anomaly detection results. The current paper therefore also presents a comparison of results obtained by the different detectors after applying different preprocessing methods.

The evaluation of the detection results is mainly based on receiver-operating characteristic (ROC) curves. For spatially fully resolved targets, the false alarm rate at first detection was also considered. For the two scenes with extended targets, besides an objective evaluation, a more subjective evaluation is also presented. The rest of the paper is organized as follows. Section 2 presents the used datasets; in Section 3 the examined anomaly detection methods are briefly presented; Section 4 presents the different preprocessing methods that have been applied to the data. The last two sections of the paper present the results and the conclusions. The appendix presents a brief exploratory data analysis that mainly aims at verifying to what extent the different used datacubes comply with the assumption of global or local unimodal multivariate normality.

2. Overview of the Dataset

The analysis was performed on a set of hypercubes of scenes with various backgrounds and representative of three scenarios as follows:

(i) a rural environment with subpixel targets (CAM and OSLO1),

(ii) a rural environment with some of the targets in shadow (BJO),

(iii) an urban environment (OSLO2).

Table 1 presents an overview of the used dataset. The first two datacubes are real hyperspectral images in which a matrix of anomalies was inserted artificially. Figure 1 shows RGB composites of these images on which the targets have been superimposed.

The results shown in this paper were obtained with 10% mixing ratio subpixel anomalies for the CAM scene. For OSLO1, the mixing ratio was varied from 100% to 10%. The inserted anomalies are spectra of a green paint (CAM) and a green fabric (OSLO1).

The BJO image (Figure 2(a)) was acquired over a natural scene with an agricultural region and a small forest near the village of Bjoerkelangen in Norway. The figure shows the target locations with light blue colored rectangles representing the target sizes superimposed on the RGB composite of the scene. Fourteen targets composed of different types of material and with different colors were laid in the scene during the image acquisition. Targets T3–T7 were in shadow. T3 was in deep shadow between the trees, and the four others were in the shadow at the edge of the forest. Table 2 presents the dimensions and material types of the different targets.

The OSLO2 scene (Figure 2(b)) is part of the center of Oslo. In this scene, four targets (T1–T4) were laid out. Their respective dimensions in the image are T1: 5 × 10, T2: 5 × 9, T3: 2 × 6, and T4: 6 × 7 pixels. Targets T2 and T3 are pieces of green fabric and the other two of a blue plastic. T1–T3 were laid out on the grass in a park, and T4 was put on an asphalt background in the shadow from a building.

The CAM image was rectified and atmospherically corrected. The images BJO, OSLO1, and OSLO2 were not rectified before processing and all processing on these scenes was applied to radiance data, that is, without applying any atmospheric correction.

3. Anomaly Detection Methods

Besides global RX, representative examples of three categories of anomaly detectors are examined in the paper. Figure 3 presents an overview of the selected detectors in the three classes. As can be seen from the figure, many of the investigated methods are RX-based, but for the subspace detection methods and in particular for the SBAD methods, some anomaly detectors that are not related to RX have also been included. The different detectors are briefly described below.

3.1. Subspace Methods. The subspace methods are global and have in common that they apply principal component analysis (PCA) or singular value decomposition (SVD) to the datacube. The first PCA/SVD bands are supposed to represent the background and they are eliminated in different ways by the various subspace methods. Subspace anomaly detectors are thus global anomaly detectors applied on a spectral subset (subspace). For all of the subspace methods, the only parameter is the number of PCA or SVD bands (n_b) that is considered to represent the background. If this number is set too high, targets will disappear in the background, if it is too low, too many false alarms will remain. Automatically determining an optimal value for the dimension of the background subspace remains a current research topic.

3.1.1. Subspace RX (SSRX). In SSRX, the global RX is applied on a limited number of PCA bands. The first PCs are discarded in SSRX.

3.1.2. RX after Orthogonal Subspace Projection (OSPRX). In OSPRX, the first PCA/SVD components define the background subspace and the data are projected onto the orthogonal subspace before applying the RX detector [2, 22].

In the current paper, the SVD of the global spectral covariance matrix Σ is used. Because Σ is positive definite, the SVD is equivalent to the following eigenvector/eigenvalue decomposition:

$$\Sigma = U\Lambda U^T, \tag{2}$$

where U is the matrix of eigenvectors of the decomposition and Λ the diagonal matrix with decreasing eigenvalues. The projection operator P_{SVD} is defined as a function of the first n_b eigenvectors (columns of U), corresponding to the highest eigenvalues, $W = U(1 \dots n, 1 \dots n_b)$ as follows:

$$D_{\text{OSPRX}}(r) = r^T(I - P_{\text{SVD}})r = r^T\left(I - WW^T\right)r, \tag{3}$$

with I the $n \times n$ identity matrix. n is the number of channels in the datacube, and n_b the number of channels used to model the background subspace ($1 \le n_b < n$).

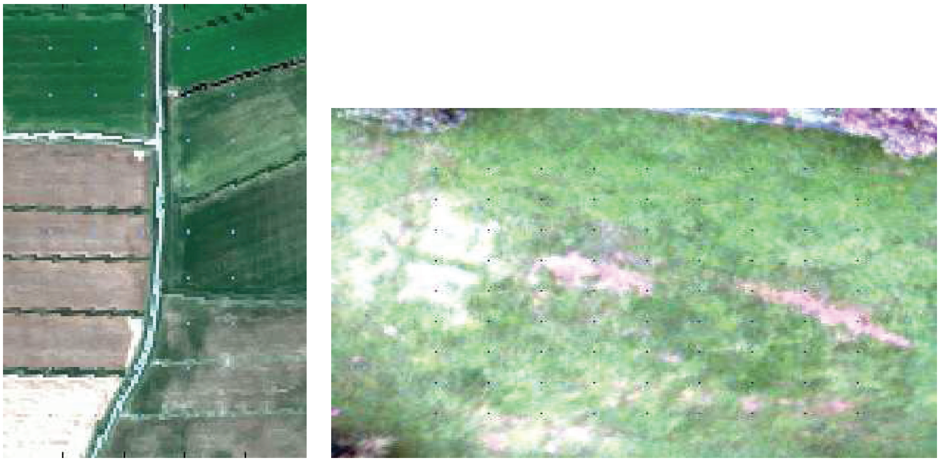
3.1.3. RX after "Partialling Out" the Clutter Subspace (PORX). In this method, the effect of the clutter in a pixel is removed (partialled out) component-wise by predicting each of its

Table 1: Overview of the used dataset.

Name	Site	Sensor name	Number of bands n	Spectral range (in μm)	Spatial resolution (in m)	Image size (in pixels)	Number of targets	Total target size (in pixels)
CAM	Camargue (Fr)	Hymap	126	0.44–2.45	4	150 × 100	45	45
OSLO1	Oslo (No)	HySpex	80	0.410–0.984	0.25	286 × 287	81	81
BJO	Bjoerkelangen (No)	HySpex	80	0.410–0.984	0.25	700 × 1600	14	574
OSLO2	Oslo (No)	HySpex	80	0.410–0.984	0.25	700 × 1600	4	45

Table 2: Material types and sizes (in pixels) of the different targets in the BJO image.

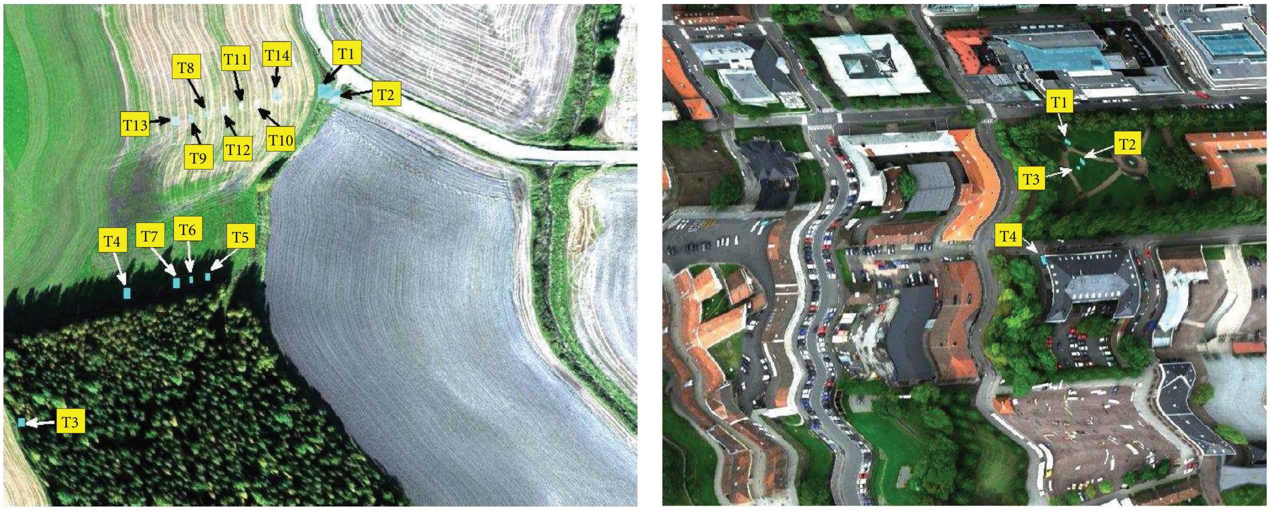
T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14
Green car	Red car	Cloths	Cloths	Painted boards			Cloths	Cloths	Painted boards				People
				Paint 1	Paint 2	Paint 3			Paint 1	Paint 2	Paint 3	Paint 4	
8 × 22	5 × 11	4 × 3	2 × 3	5 × 7	4 × 4	3 × 3	2 × 7	3 × 4	8 × 10	5 × 7	4 × 7	5 × 8	7 × 16



(a)

(b)

Figure 1: RGB color composite of the CAM (a) and OSLO1 (b) datacubes with targets superimposed, respectively, in cyan and black.



(a)

(b)

Figure 2: RGB composite of the BJO (a) and OSLO2 (b) datacubes with target locations indicated.

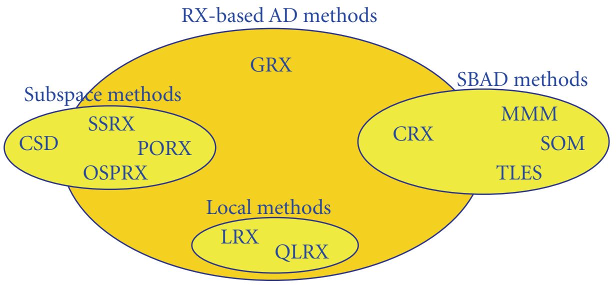


Figure 3: Schematic overview of the examined anomaly detection methods.

spectral components as a linear combination of its high-variance principal components. The detector applies the RX detector on the residual. Details of the method can be found in [10].

3.1.4. Complimentary Subspace Detector (CSD). The CSD is not an RX-based method. In the CSD, the highest variance principal components are again used to define the background subspace and the other PCs, to define the target subspace (the complimentary subspace) [7]. The PUT is then projected on the two subspaces and the anomaly detector is the difference of the projection onto the target subspace and the background subspace as follows:

$$D_{\text{CSD}}(r) = r^T P_t r - r^T P_b r, \tag{4}$$

where

$$\begin{aligned} P_b &= U(1\ldots n, 1\ldots n_b)U^T(1\ldots n, 1\ldots n_b), \\ P_t &= U(1\ldots n, (n_b+1)\ldots n)U^T(1\ldots n, (n_b+1)\ldots n). \end{aligned} \tag{5}$$

3.2. Local Methods. In the local anomaly detection methods, the statistics of the background are estimated locally in a window around the PUT. A double sliding window is used: a guard window and an outer window are defined, and the background statistics are determined using the pixels between the two (see Figure 4). Sometimes a triple window is used where the covariance matrix of the background is estimated in a larger window than the average local background spectrum.

3.2.1. Local RX (LRX). In LRX, the covariance matrix Σ_B and mean spectrum μ_B of the background are estimated locally in a triple window around the PUT. In the used implementation, the size of the guard window is a parameter from which the size of the two other windows is determined as a function of the number of bands in the image as follow:

$$\begin{aligned} W_\mu &= \min_{k\,\text{odd}}\left(k^2 - W_G^2 \geq \sqrt{10n}\right), \\ W_\Sigma &= \min_{k\,\text{odd}}(k^2 - W_G^2 \geq 10n). \end{aligned} \tag{6}$$

3.2.2. Quasi-local RX (QLRX). Quasi-local RX (QLRX) [9] offers a compromise between the global and local RX approach. In QLRX, the global covariance matrix Σ is decomposed using eigenvector/eigenvalue decomposition (2). The eigenvectors are kept in the RX, but the eigenvalues are replaced by the maximum of the local variance and the global eigenvalue as

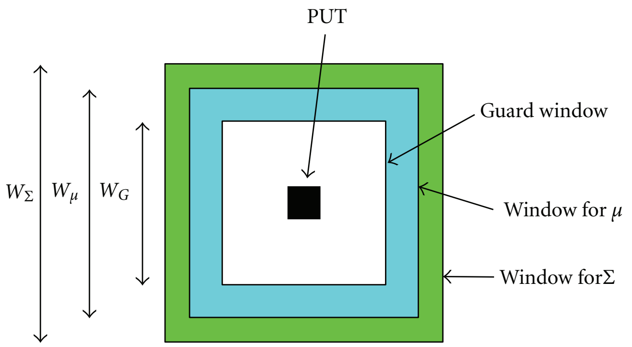


Figure 4: Sliding triple window used in the local AD methods.

$$\lambda_{\text{QL}}^i = \left[\max\left(\lambda_{\text{loc}}^i, \lambda_{\text{glob}}^i\right)\right], \tag{7}$$

where i is an index denoting the pixel in the image. This means that the score of the detector will be lower at locations of the image with high variance (e.g., edges) than in more homogeneous areas. Spectral statistical standardization (see Section 4.4) is applied as a preprocessing step. The local variance is determined in a double sliding window.

3.3. Segmentation-Based Methods. In complex scenes, the hypothesis of a single multivariate normal distribution of background spectra is usually not verified, not even locally. For that reason several segmentation-based anomaly detectors have been proposed in literature. In this paper, four of these methods have been included in the analysis.

3.3.1. Class-Conditional RX (CRX). In CRX, the image is first segmented, the covariance matrix and mean within each class i (i.e., Σ_i and μ_i) are determined. The Mahalanobis distance between the PUT and each of the classes is calculated. The final result is the minimum of these distances:

$$D_{\text{CRX}} = \min_i\left[(r-\mu_i)^T\Sigma_i^{-1}(r-\mu_i)\right]. \tag{8}$$

In the current paper, K-means clustering is used and the parameters of the method are the minimum number of pixels allowed in each class and the maximum number of classes used in the clustering. The number of classes follows from these parameters.

3.3.2. Method Based on Multivariate Normal Mixture Models (MMM) [13]. A Stochastic Expectation Maximization (SEM) algorithm [23] is used for fitting a multivariate normal mixture model to the image for describing the background. The anomaly detector detects pixels that have a low probability according to the fitted model.

The parameters of the method are the maximum number of mixture components and the termination threshold for the iterative parameters estimation method. The idea behind the mixture model is to use the mixture components (the multivariate normal pdfs) as base pdfs in a row expansion of the true pdf, which can then, in principle, have any shape.

3.3.3. Two-Level Endmember Selection Method (TLES). The principle of the TLES method [19] is as follows: a small scanning window (50 × 50 pixels) runs over the image and at each position of the window the principal background spectra are determined using a segmentation method based on endmember selection. Endmembers that correspond to at least a given percentage (MP) of the image tile are stored. At the end of the process, an endmember selection is again applied on the stored endmembers and linear unmixing is applied on the image. Anomalies correspond to pixels with a large residue after unmixing. In [19], N-FINDR was used as the endmember selection method. In the current paper, the minimum volume simplex analysis algorithm (MVSA) [24] was used because it was found to give better results. Parameters of the method are the threshold MP and the number of endmembers kept in the two stages of the algorithm.

3.3.4. Method Based on Self-Organizing Maps (SOM). A trained SOM is considered as a representation of the background classes in the scene. Anomalies are determined by computing the spectral distances of the pixels from the SOM units [16, 17]. The SOM was applied on the first PCA components and run using a square map consisting of NsxNs hexagonal cells. The SOM was optimized sequentially. The parameters of the method are Ns and the number of PCA bands used.

4. Preprocessing Methods

Before applying the actual anomaly detectors, some preprocessing methods were applied to the data. Three different types of preprocessing were applied.

The first type is data dimension reduction, which has two objectives. The first objective is to describe the background better and to obtain more reliable statistical estimation, especially when applying local methods where the number of samples to compute statistics from is low. Moreover, reducing data dimension allows to reduce the size of the windows for the local methods. This reinforces the local aspect of the method and reduces the risk that nearby targets overlap with the window used to compute background statistics. The second objective is to project the data on axes where the anomalies are enhanced, that is, the most separated from the background pixels. In this paper, we focus on two different methods, spectral binning which fulfills the first objective, and kurtosis-based dimension reduction, that attempts to fulfill both.

The second type of preprocessing aims to account for the effects of shadow. In this paper, a simplified approach consisting in square root transforming the data is used (Section 4.5).

Finally, some AD methods need some specific preprocessing that is described in Sections 4.3 and 4.4.

4.1. Dimension Reduction by Spectral Binning (SB). As noted in [25, 26], dimension reduction can improve hyperspectral anomaly detection performance substantially. We have applied a dimension reduction method based on spectral binning, similar to the method applied in [26]. The binning consists in averaging over groups of neighboring bands, down to a spectral resolution of about 30 nm. The binning tends to improve the signal-to-noise ratio by reducing the relative contribution of photon noise. When this can be done while preserving the relevant spectral features, the result is improved detection performance.

4.2. Kurtosis-Based Dimension Reduction (KDR). As anomaly detection aims to search for outliers, a projection that enhances outliers applied as a preprocessing can improve detection performance. It has been shown that kurtosis is very sensitive to outliers. In [27] data are projected on the (first) eigenvectors of the kurtosis matrix K:

$$K = \Sigma^{-1}\frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^t\Sigma^{-1}(X_i - \mu)(X_i - \mu)(X_i - \mu)^t, \quad (9)$$

where X_i is the ith element of X, the matrix of observations (the spectrum of the ith pixel), and μ and Σ are the spectral mean and covariance matrix of the datacube. N is the total number of pixels in the image. This method is mainly useful if the data are unimodally distributed, that is, in scenes characterized by a relatively homogeneous background (in this work, the CAM and OSLO1 images). Usually, only the first 3 to 5 kurtosis components are kept for further processing. This is the case for GRX, LRX and QLRX, MMM, and SOM. For the subspace methods and TLES, all kurtosis components were used.

4.3. Spectral Whitening. If the eigenvalues and eigenvectors of the covariance matrix of the complete image are, respectively, Λ and U, and μ is the average spectral vector of the image, then the spectral whitening of the pixel r is given by [6, 7]

$$r_W = U\Lambda^{-1}U^T(r - \mu). \quad (10)$$

After spectral whitening, a Gaussian distributed variable becomes spherically symmetric and this is sometimes beneficial for detection [6]. Whether whitening is beneficial for the anomaly detector depends on the AD method and the datacube. For CSD, spectral whitening is always applied. The other subspace methods were applied with and without whitening and the best result obtained is reported in this paper.

4.4. Spectral Statistical Standardization. The spectral statistical standardization converts each spectral band to have a zero mean and a standard deviation of one. This is necessary for the QLRX in order to make sure that the global eigenvalues and the local variances can be interchanged in the algorithm.

4.5. Square Root Transform. Detection performance is generally degraded in shadow, due, among other things, to low signal-to-noise ratio and distortions of the spectral signatures caused by secondary illumination, the adjacency effect, and path-scattered skylight. In addition, the large dynamic range of the data from scenes containing both sunlit and shadowed areas makes the data modeling task more difficult. In order to improve detection performance in shadow, different strategies can be applied: de-shadowing and illumination suppression for estimation of sunlit-like radiance in shadowed areas [28–31], sun/shadow segmentation and application of adapted modeling in the respective areas [32], or transformation of the data to account for the effects of shadow. We have square root transformed the data $r_{\text{sqrt}} = \sqrt{r}$. This reduces the dynamic range of the data, and, perhaps more importantly, makes the noise signal level independent [33], with benefits for data modeling through a suppression of the influence of noisy low-level signals. At low signal levels for homogeneous backgrounds, the dominating source of variation in the signal is Poisson distributed counting (photon) noise, and square root transforming the data yields approximate normality [34]. Square root transforming the data of course also affects the distribution in the, vast majority of, cases where scene clutter is the main source of signal variation, but in unpredictable ways for complex backgrounds.

5. Evaluation Method

Experimental ROC (receiver operating characteristic) curves, showing the detection rate (DR) versus the false alarm rate (FAR), are used to evaluate the results obtained with the various detectors. For the images with resolved targets, a pixel-based ROC curve is calculated for each target, whereas for the images with subpixel targets, an ROC curve is calculated based on all the targets in the image. DR is plotted versus the logarithm of the FAR (the resulting curve is referred to as a logROC), and the area under the logROC curve (the logAUC) is calculated and used as the measure of performance. The reason for using a logarithmic FAR scale is that it ensures equal weight across the range of FAR values.

For extended targets (in BJO and OSLO2), ROC curves give the detection performance for each pixel of the target with respect to the false alarm rate. For defense and security applications, it is also of interest to assess the performance at the first detection of a target. In this paper, for extended targets, we therefore also determined the false alarm rate at the first detection for each of the targets.

Besides these objective evaluation metrics, it is also interesting to look at the type of false alarms that the various detectors produce in the different scenes. Therefore for the "best" detectors, a detection image is shown corresponding to the threshold for which at least one pixel of the most difficult target is detected. This subjective result is shown for scenes with extended targets (BJO and OSLO2).

6. Results and Discussion

6.1. Implementation Issue: Parameter Selection. The different examined AD methods depend on different parameters. For some of the applied methods, the parameters were set according to experience, whereas for others, where we lack this experience and where no consensus exists in literature for setting the parameters, we chose the optimal parameter setting through an optimization process in order to make the comparison between the detectors fair.

For the local methods, the parameters are the dimensions of the guard window and the outer window(s). The guard window should be set to be larger than the largest target of interest expected to be present in the scene, and size of the outer window(s) is derived from it as explained in Section 3.2.1. For the two datasets with subpixel anomalies, the guard window was set to 1, and for the two other datasets, it was set to 15.

GRX has no parameters.

For subspace methods, the only parameter is the dimensionality attributed to the background subspace. Several methods have been proposed for estimating this "signal subspace," mainly for unmixing purposes [35–39]. The two latter focus on finding the signal subspace dimension in the presence of "rare signals." They are thus likely to add the signal components containing the target to the signal subspace and are therefore less relevant to the choice of the signal subspace in subspace anomaly detection. The remaining methods [35–37] give different results and none of the signal subspace dimensions estimated by these methods correlate in a consistent way with the optimal number of bands in the subspace detectors for the different scenes. A consistent way for identifying the proper dimensionality to use in modeling the background clutter for subspace anomaly detection has yet to be found, as mentioned in the conclusion of [10]. Because the aim of the paper is to compare the different algorithms in the different scenes, the dimensionality parameter is optimized for each detector/scene combination. The complete range of possible background dimensionality (1 to $n-1$) was explored and the results shown are the best results obtained by that algorithm.

Each of the SBAD methods has its own set of parameters. For CRX, two parameters are set: a maximum number of classes and a minimum number of pixels per class. This last parameter can be used to reduce the risk that anomalies form their own classes. Then the maximum number of classes can be set higher than the actual number of background classes. The minimum number of pixels in each of the classes is set to a low percentage (for instance 0.5%) of the total number of pixels. For MMM, the parameters are similar to those of CRX. For SOM and TLES, the parameters, described in Section 3.3, were varied in a reasonable range and the results shown are the best obtained for the examined range of parameters.

6.2. Results for Subpixel Detection in a Rural Environment

6.2.1. Results for CAM. Table 3 shows the logAUC results for the different detectors obtained in the CAM dataset with a 10% mixing ratio. Results are shown without prior data reduction and with two types of data reduction: spectral binning and kurtosis-based data reduction. We see from the

Table 3: LogAUC results for CAM scene with a mixing ratio of 10% for different types of data reduction.

	Data reduction method		
AD method	No data reduction	Spectral binning	Kurtosis data reduction
GRX	0.595	0.637	*0.732*
SSRX	0.742	0.778	**1.000**
OSPRX	*0.931*	0.754	0.893
PORX	0.568	0.205	*0.957*
CSD	*0.868*	0.826	0.763
LRX	0.616	0.950	**1.000**
QLRX	0.720	0.828	**1.000**
CRX	0.683	0.830	**1.000**
TLES	0.128	0.268	*0.698*
SOM	0.116	*0.529*	0.120
MMM		**1.000**	**1.000**

Table 4: LogAUC results for OSLO1 scene with a mixing ratio of 33% for different types of data reduction.

AD	No data reduction	Spectral binning	Kurtosis data reduction
GRX	0.163	0.357	*0.390*
SSRX	0.292	*0.486*	0.394
OSPRX	0.383	*0.486*	0.391
PORX	0.322	*0.493*	0.391
CSD	0.246	0.391	*0.396*
LRX	0.821	0.983	**0.995**
QLRX	0.294	0.395	*0.473*
CRX	0.163	0.357	*0.449*
TLES	0.108	*0.466*	0.356
SOM	0.232	0.385	*0.463*
MMM	0.128	0.544	*0.597*

table that all SBAD methods, local methods, and GRX benefit from dimension reduction, more from kurtosis dimension reduction (KDR) than from spectral binning (SB) (with exception for SOM, which performs much better after SB than after KDR). For the subspace methods, the results are more diverse: OSPRX and CSD perform best on data with no dimension reduction, whereas for SSRX and PORX a large improvement in performance is observed after KDR.

Figure 5 shows a scatter plot of the CAM data with 10% mixing ratio subpixel targets in the 2D space defined by the first two kurtosis components. The figure shows that the targets are very well separated from the background after KDR transform in this scene. This clear separation is not observed on any of the PCA components or on spectral bands before or after spectral binning.

The best results for this scene are indeed obtained with SSRX, LRX, QLRX, CRX, and MMM after KDR. These detectors all achieve a logAUC of 1.0, which means that all targets have been detected with a false alarm rate that is smaller than 1/image size, that is, FAR $< 6.6 * 10^{-5}$. Since the results are saturated, we cannot properly distinguish between the methods.

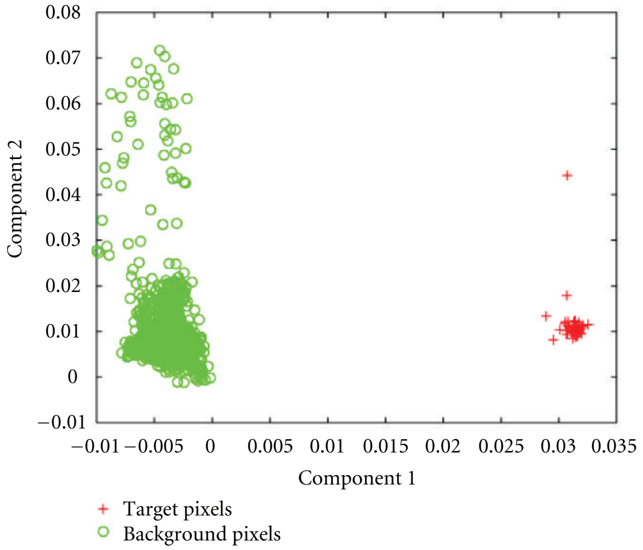


Figure 5: Projection of background (green) and target (red) pixels on the first two kurtosis component axes for CAM10.

6.2.2. Results for OSLO1. Table 4 shows the logAUC results for the different detectors obtained in the OSLO1 dataset with a 33% mixing ratio. Results are shown without prior data reduction and with spectral binning and kurtosis-based data reduction.

LRX clearly outperforms all other detectors. The assumption of local normality is very well met in this image (cf. Table 6 in the Appendix). LRX is also, in contrast to all other methods, able to model the background without influence of targets. For subpixel targets this is particularly true, since the guard window will always contain the whole target. As regards data reduction, for SBAD and local methods and GRX, the results are the same here as for CAM10: the methods all benefit from dimension reduction, and more from KDR than from SB (with exception for TLES which benefits more from SB than from KDR). In this dataset, we also observe an improvement in performance with dimension reduction for the subspace methods, but for these methods SB is generally more beneficial than KDR.

For OSLO1, the behavior of each detector as a function of the mixing ratio was also investigated. Figure 6 shows the logAUC results of the different detectors versus the mixing ratio for the OSLO1 datacube. In the experiment, the mixing ratio was varied from 100% (full pixel anomaly) to 10%. For creating the figure, the data reduction method that gives the best results was selected for each of the detectors. Results of global RX-based methods and CSD are shown as solid lines, LRX and QLRX results as dashed lines, and results of segmentation-based methods as dot-dashed lines. LRX clearly gives the overall best results, followed by SSRX for larger target portions and MMM for smaller. The result of

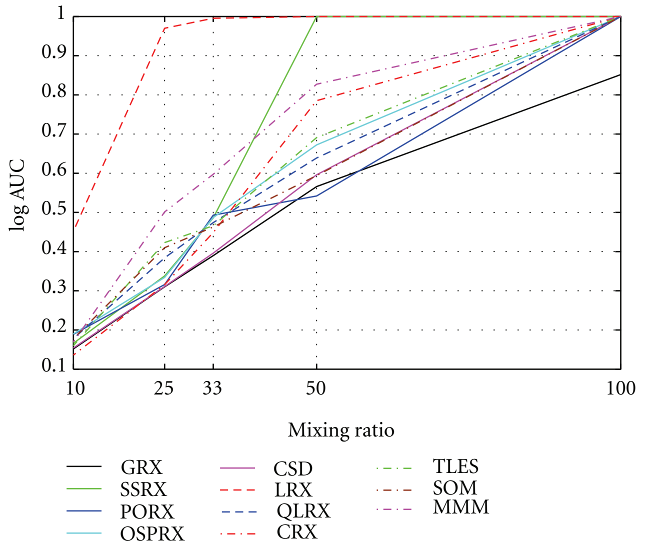


Figure 6: LogAUC versus mixing ratio for the different detectors in OSLO1.

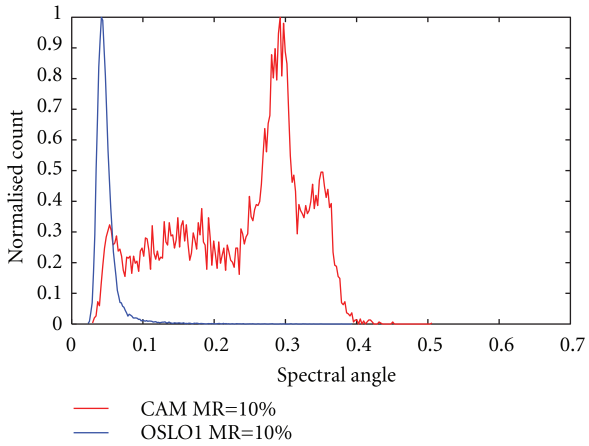


Figure 7: Normalized histograms of the spectral angle of the background pixels with respect to the average target spectrum in the CAM and OSLO1 scenes, at 10% mixing ratio.

SSRX for the 50% mixing ratio is very deviant, suggesting perhaps that subspace fit is somewhat random.

Contrary to the CAM scene, in OSLO1 the performance of the detectors at 10% mixing ratio is very low. The targets are more difficult to detect in the OSLO1 scene than in the CAM scene although the OSLO1 scene has a more homogeneous background and conforms well to the multivariate Gaussian assumption (see discussion of Table 5 in the Appendix). The OSLO1 scene is more difficult than the CAM scene because both the spectral angle and the Euclidean distance between the targets and the different background spectra are much smaller in the former. Figure 7 illustrates this point by means of the normalized histogram of the spectral angle of all background pixels with respect to the average target spectrum for both scenes at a mixing ratio of 10%. The figure shows that the spectral angle is indeed larger in the CAM scene than in the OSLO1 scene. The spread of the spectral angle in the CAM scene also illustrates the heterogeneity of the background.

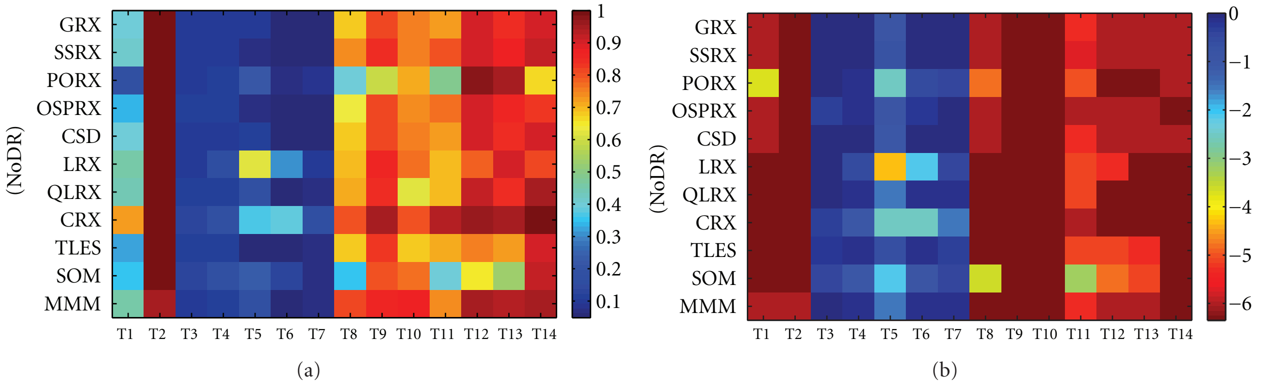
6.3. Results for a Rural Environment with Some of the Targets in Shadow (BJO). Figures 8 to 10 show a graphical representation of the logAUC (a) and the logarithm of the false alarm rate at first detection (logFARAt1stDet) (b) for each of the detectors and for each target for the BJO scene. The colors represent the value of the respective performance metric. The color map is such that red corresponds to the best performance. The three different figures represent results after different types of preprocessing: Figure 8 shows results obtained without any preprocessing, Figure 9 results after spectral binning, and Figure 10 results after spectral binning and square root transform.

From the figures it is immediately clear that the targets in shadow (T3–T7) are more difficult to detect than the others. T3, hidden in the forest is the most difficult to detect. T2 (a red car) is the most easily detectable target.

We observe a slight improvement of performance from spectral binning for all detectors except TLES and SOM, which have their detection performance for targets in the sun substantially degraded by spectral binning. The reason why we—as opposed to what we saw in the previous datasets—only observe a slight improvement in performance from spectral binning, could be that the targets in sun are so easy to detect that we detect them anyhow, whereas the targets in shadow need some form of compensation for shadow in order to be detectable. The levels of the performance values indicate that this could be the case. For LRX, the size of the windows used to calculate the background statistics depends on the number of bands, and the results hence are not comparable across different numbers of bands.

Square root transforming the data generally improves the detection performance for targets in shadow substantially. It also improves the performance for targets in the sun for some detectors, but for others, mainly the subspace detectors and GRX, it reduces the performance for some sunlit targets—notably targets that are intensity anomalies (T8, T10, and T11), and that hence have their degree of anomality reduced when the data are square root transformed. The best results on this dataset are obtained with MMM, LRX, and CRX. MMM gives the overall best results, whereas LRX gives the best results for the targets in shadow: for targets T5 and T6, LRX gives significantly better results than MMM. LRX gives also slightly better results than MMM for T10 and T11, which are painted with the same paints as, respectively, T5 and T6. The logAUC and logFARAt1stDet results are globally inter-consistent, but they do show supplementary information. The results are consistent with the complexity of the scene and with the compliance with the multinormal distribution assumption locally shown in Table 6 of the Appendix.

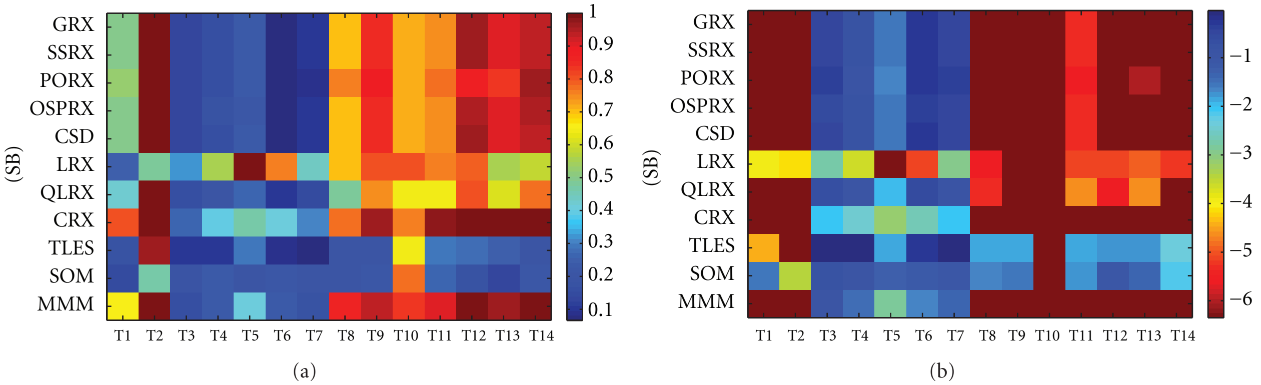
In order to give an idea of the type of false alarms produced by the three best detectors, in Figure 11 the results of the three best detectors (MMM, LRX, and CRX after spectral binning and sqrt transform) are superimposed on a grayscale image of the BJO scene. The shown results are



(a)

(b)

Figure 8: logAUC and logFARAt1stDet results per target without any data reduction for BJO.



(a)

(b)

Figure 9: logAUC and logFARAt1stDet results per target after spectral binning for BJO.

thresholded detection results with the threshold set to the lowest first detection level for the true targets (i.e., the threshold for which at least one pixel of the most difficult target is detected). The figure shows that the false alarms produced by MMM mainly consist of isolated pixels in the forest and also some more extended false alarms at the top right of the image. LRX produces some small false alarms in the forest while CRX detects part of the stream as well as some detections in the forest. Most of the false alarms are detected by only one detector. On the other hand, for each target, except T7, an overlap in the detected zone for the three detectors is seen. The detectors are thus complimentary and fusing their results may be of interest. Likely causes of the complementarity of the results are that LRX are able to account more correctly for local illumination, whereas MMM/CRX are able to model locally heterogeneous background (forest) more precisely. The results for CRX indicate that too few classes are used: a large background structure like the stream is poorly modelled.

6.4. Results for the Urban Scene (OSLO2). Figure 12(a) shows the logAUC results for the OSLO2 scene. Figure 12(b) shows the logFARAt1stDet results. For creating the figure, the data reduction method that gives the best results was selected for each of the detectors. The best data reduction method was spectral binning for most detectors, but for OSPRX and QLRX the best results were obtained without data reduction, and for LRX, KDR gave the best results. The superiority of spectral binning over KDR for most detectors is to be expected because of the complexity of the scene (cf. Table 5 in the Appendix), and similarly the good performance of KDR for LRX—it can be attributed to local normality, see Table 6 in the Appendix. The mixed results for subspace detectors are consistent with the results for CAM and OSLO1. The result for QLRX, on the other hand, is not, but we should probably not read too much into this, since the detector more or less fails to detect the targets.

It can be seen that the values of logAUC are much lower than for the other datacubes. The maximum value obtained here is around 0.5. This is due to the complexity of the scene: the targets inserted into the scene are not the only anomalies. In an urban environment many objects can present an anomalous spectrum, for example, cars, special roof materials, and so forth. The comparison therefore only shows how well the different anomaly detection methods cope with urban “clutter.”

From the two figures it can be seen that MMM gives the overall best results, followed by TLES and CRX. Of the two latter CRX gives the best results according to the logAUC metric and TLES according to the logFARAt1stDet metric. As could be expected the SBAD methods thus obtain the best results in the urban scene.

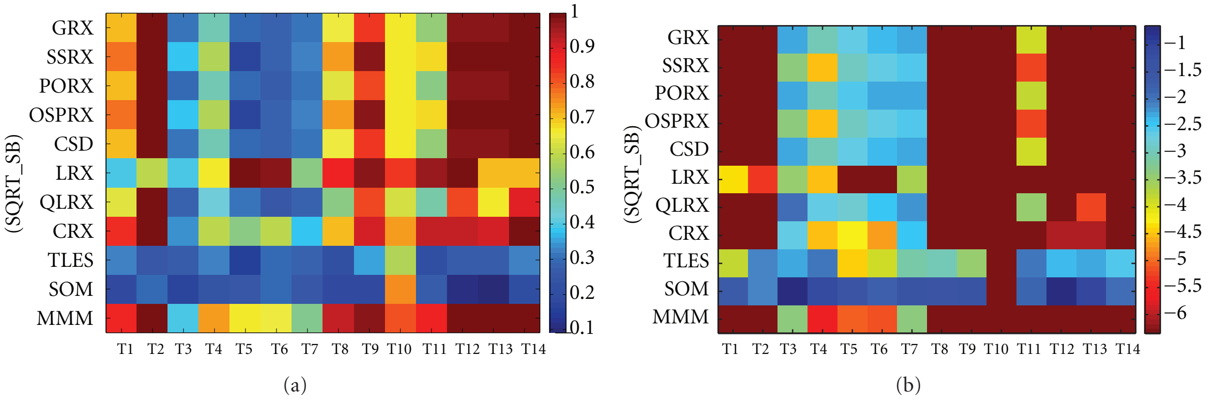


Figure 10: logAUC and logFARAt1stDet results per target after SQRT transform and spectral binning for BJO.

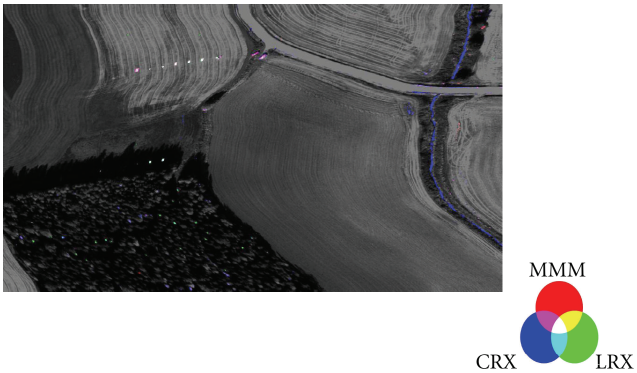


Figure 11: Color composite of the detection results for MMM, LRX, and CRX at lowest 1st detection threshold, superimposed on B/W image of BJO (R: MMM results, G: LRX results, B: CRX results).

GRX and all subspace methods perform quite bad, except for OSPRX that gives quite good results for targets T1, T2, and T4. Contrary to what we have seen in the other datasets, LRX does not perform particularly well in this dataset—despite good compliance with the multivariate normal distribution assumption locally near the targets, cf. Table 6 in the Appendix. The reason for this is probably that the remaining (non-target) parts of the scene are not well described by a local multivariate normal distribution, and hence we get lots of false alarms. This assumption is partly verified by comparing with the results of OSLO1. The target material of targets T2 and T3 in OSLO2 is the same as the material of the OSLO1 targets, and the local background is very similar (grass at two different places in Oslo), so a difference in performance between the images ought to be due to a difference in the number of false alarms.

As mentioned above, an urban scene presents many objects that may have an anomalous spectrum and that thus will be considered a false alarm in the above evaluation. It is therefore of interest to give an idea of these false alarms for each of the detectors. Figure 13 presents the results of the three best detectors (MMM, CRX, and TLES), superimposed on a grayscale image of the OSLO2 scene. The results shown are thresholded detection results with the threshold set to the lowest first detection level for the true targets.

The figure shows that target 1 is the most easy, and that it was detected completely by the three detectors. T2 and T3 have been completely detected by MMM, while the two other detectors (at the selected thresholds) detect only a part of the interior. On the contrary, T4 has been completely detected by TLES and CRX, while MMM only detects a part of its interior. The figure also shows that many of the "false alarms" are quite different for the three detectors. TLES detects parts of the vegetation. CRX detects a set of small objects next to the building on the lower left in the image. Some cars and small structures on roofs have been detected by a combination of detectors. This subjective evaluation shows that the three detectors that perform best according to the "objective" evaluation are quite complimentary. Examining further the properties of their results may lead to interesting ideas on fusion of anomaly detectors.

7. Conclusions

This paper evaluates the performance of anomaly detection methods in scenes with different backgrounds and types of targets: agricultural scenes with subpixel targets, an agricultural scene with some of the targets in shadow, and an urban scene. Three classes of anomaly detectors were considered besides the global RX: subspace methods, local methods, and segmentation-based anomaly detection (SBAD) methods.

For subpixel anomaly detection in scenes of low complexity (rural and non shadow), LRX gives the best results, followed by MMM. From the investigated global RX-based methods SSRX and OSPRX give the best results. For the SBAD and local methods and GRX, detection results are improved by data reduction, and (with minor exceptions) more by kurtosis dimension reduction than by spectral binning. The improvement of results after kurtosis-based data reduction for most of the detectors illustrates the potential of customizing the data reduction method.

For the rural scene with some of the targets in shadow the results show that it is important to account for the effects

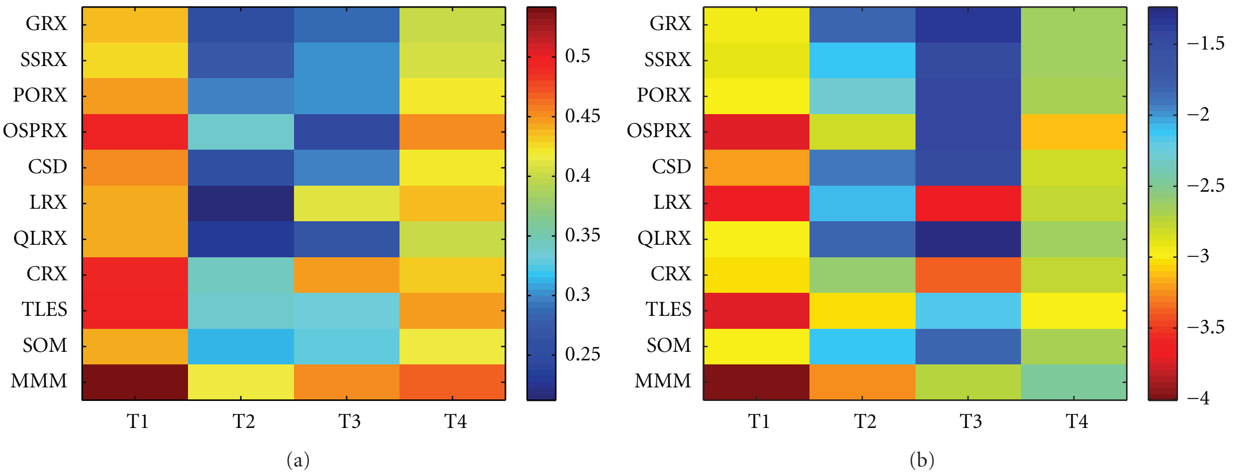


FIGURE 12: logAUC and logFARAt1stDet results per target for OSLO2 using the best data reduction method for each detector.

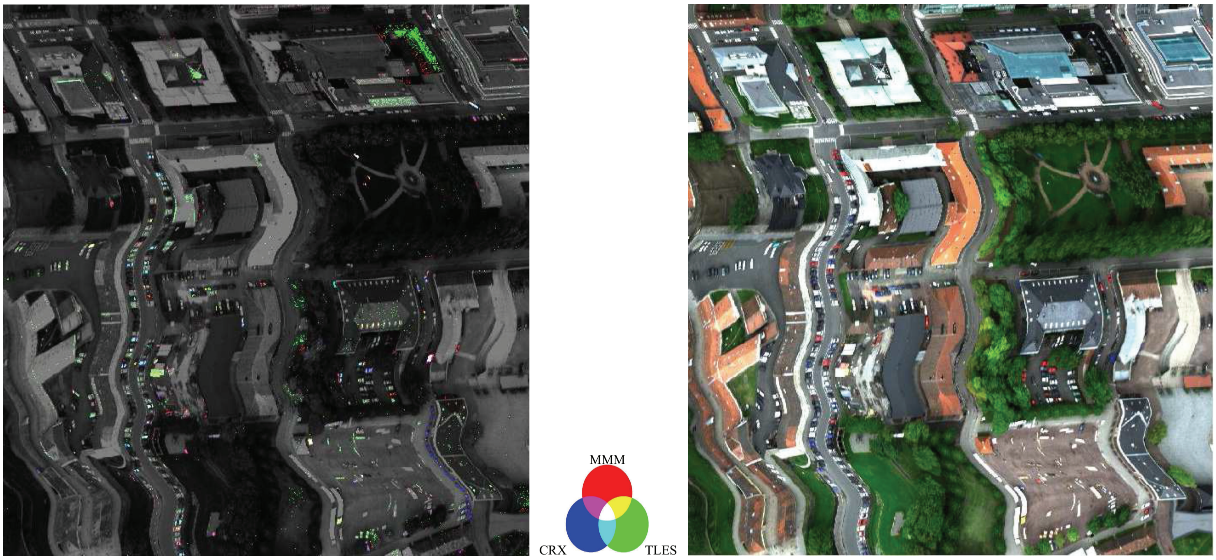


FIGURE 13: Left: color composite of the detection results for MMM, CRX, and TLES at lowest 1st detection threshold, superimposed on B/W image of OSLO2 (R: MMM results, G: CRX results, B: TLES results).

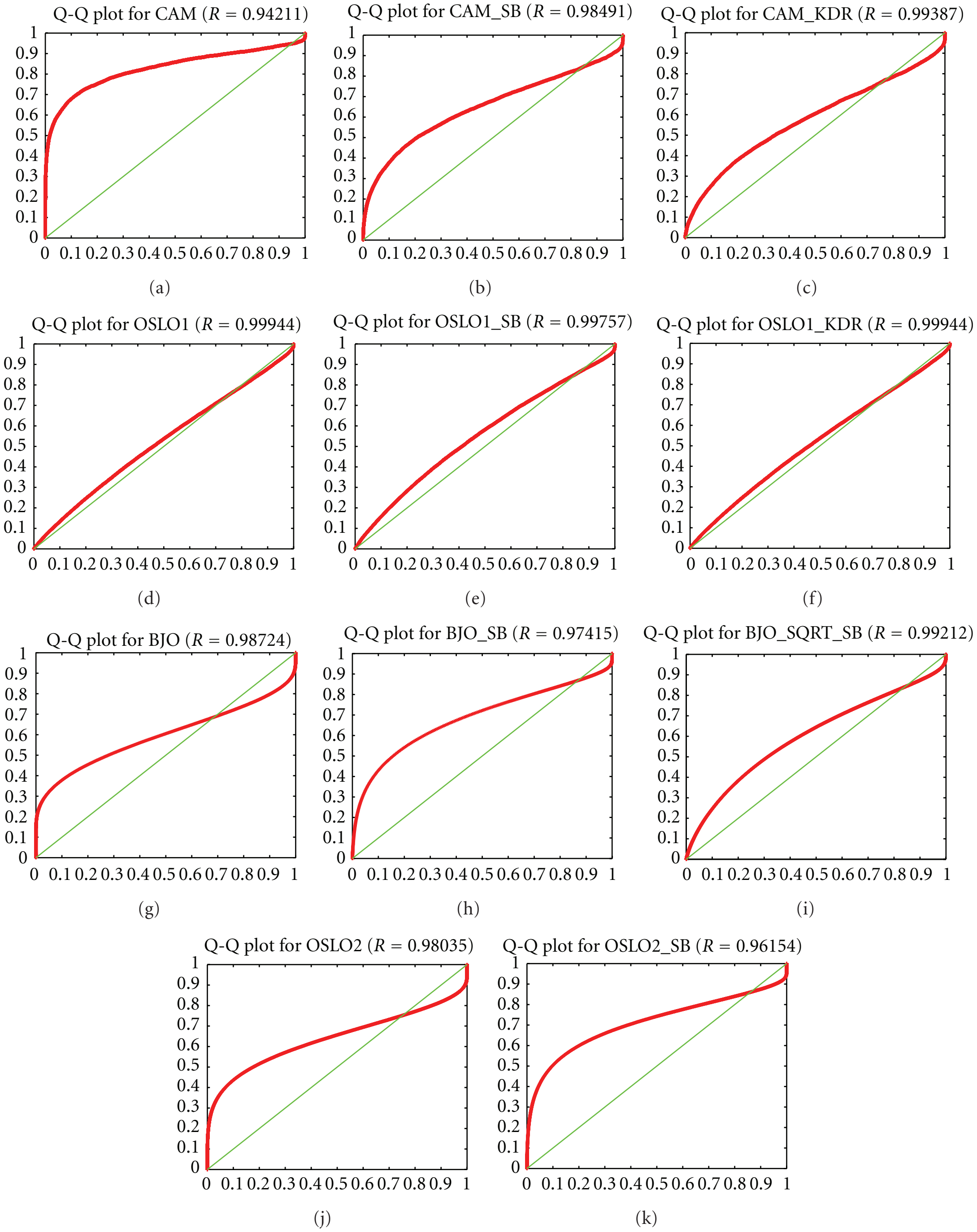
of shadow. In this paper this was done using a simplified approach consisting in square root transforming the data. After this transform MMM gives the best overall results, followed by LRX and CRX. For some targets LRX gives significantly better results than MMM. These three best detectors produce different false alarms while producing a common detection for all but one of the targets. They are thus complementary to each other and fusion of their results should be beneficial.

In the urban environment the SBAD methods perform best. The overall best result for the urban scene is obtained by MMM, TLES, and CRX. Of the globl RX-based methods OSPRX gives the best results in this dataset. Subjective evaluation of the detection results shows that the best performing detectors give complimentary results, and that "false alarms" are mainly due to objects with anomalous spectra in the scene such as cars and parts of buildings. Further investigation of this complementarity may lead to efficient detector fusion.

Appendix

Exploratory Data Analysis

The main aim of the exploratory data analysis is to investigate how well the different datacubes comply with the assumption of unimodal multivariate normality. If a distribution is multivariate normal, the square of the Mahalanobis distances of its samples follows a χ^2 distribution with degrees of freedom equal to the dimension of the multivariate variable [40]. The compliance can then be investigated visually using a Q-Q plot of the empirical cumulative distribution function (CDF) of the Mahalanobis distance and the CDF of the theoretical χ^2 distribution. Figure 14 shows the Q-Q plots



Q-Q plot for CAM (R = 0.94211)

(a)

Q-Q plot for CAM_SB (R = 0.98491)

(b)

Q-Q plot for CAM_KDR (R = 0.99387)

(c)

Q-Q plot for OSLO1 (R = 0.99944)

(d)

Q-Q plot for OSLO1_SB (R = 0.99757)

(e)

Q-Q plot for OSLO1_KDR (R = 0.99944)

(f)

Q-Q plot for BJO (R = 0.98724)

(g)

Q-Q plot for BJO_SB (R = 0.97415)

(h)

Q-Q plot for BJO_SQRT_SB (R = 0.99212)

(i)

Q-Q plot for OSLO2 (R = 0.98035)

(j)

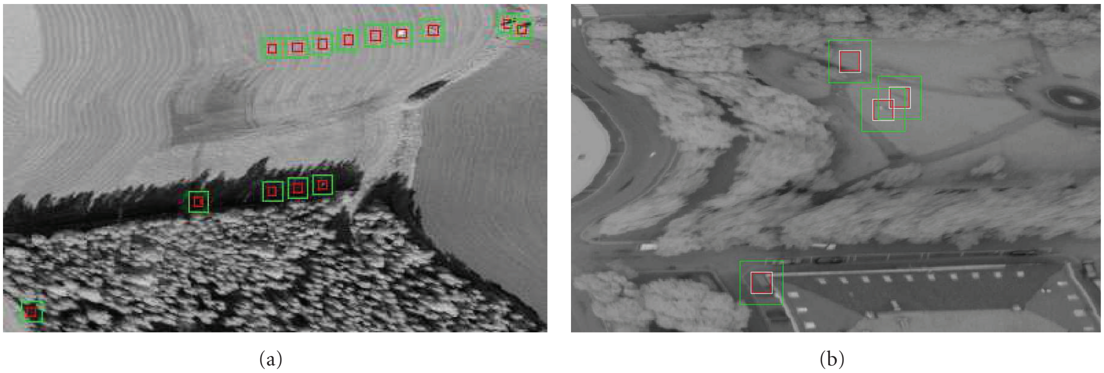
Q-Q plot for OSLO2_SB (R = 0.96154)

(k)

Figure 14: Q-Q plots for the different data cubes used in this paper.

for the four scenes and the different used preprocessing methods. Table 5 shows the correlation coefficient between the empirical and theoretical CDF of the Mahalanobis distance as well as the maximal deviation between the two (the Kolmogorov-Smirnov test statistic).

As can be expected the OSLO1 scene, consisting of a very homogeneous background, conforms best to the assumption of global multivariate normality. Both the original data and the data after kurtosis data reduction have a high correlation coefficient between the two CDFs and a low value for the KS-statistic. None of the other datasets comply with the global normality assumption. For the CAM scene the multivariate normality improves by preprocessing and the best normality is achieved after kurtosis data reduction when all kurtosis components are considered. When only the first five components are considered, the global normality


(a)

(b)

Figure 15: Guard window (red) and outer window used for calculating local Σ centered at each target location for BJO (a) and OSLO2 (b).

Table 5: Correlation coefficient and Kolmogorov-Smirnov test statistic between empirical and theoretical CDF of the Mahalanobis distance and the condition number of global Σ.

Name	Preprocessing	Correlation	KS statistic	Condition number of Σ
CAM	None	0.937	0.579	$1.11e + 10$
	SB	0.984	0.295	$7.23e + 08$
	KDR	0.993	0.178	$2.91e + 08$
	KDR (5 kcp)	0.892	0.540	$2.95e + 00$
OSLO1	None	0.999	0.049	$9.11e + 03$
	SB	0.997	0.093	$8.32e + 03$
	KDR	0.999	0.049	$6.73e + 00$
	KDR (5 kcp)	0.973	0.251	$4.91e + 00$
BJO	None	0.986	0.278	$1.63e + 05$
	SB	0.973	0.343	$1.03e + 05$
	SQRT + SB	0.992	0.188	$1.06e + 05$
OSLO2	None	0.979	0.336	$3.38e + 04$
	SB	0.961	0.409	$2.27e + 04$

assumption is not met. For BJO the combination of the square root transform and spectral binning improves the normality of the data.

Several of the investigated anomaly detection methods (the global RX-based methods and LRX) rely on the estimation and inversion of the spectral covariance matrix Σ. It is known that the sample covariance matrix in many cases needs to be regularized before inversion [41, 42]. The regularization makes the problem of finding the inverse mathematically stable, but if the initial matrix does not have a stable inverse, that is, is well-conditioned, the obtained inverse might not lead to a good detection result for the detector. In Table 5 the condition numbers of the covariance matrices of the complete images are also given. The condition number is the ratio between the highest and lowest singular value and it provides an indication of the accuracy of the results of matrix inversion. In CAM and OSLO1 the various preprocessing methods appear to decrease the condition number. This decrease is particularly significant for KDR in the OSLO1 scene. For KDR, contrary to the normality assumption, considering only the first five components reduces the condition number substantially.

The local anomaly detection methods estimate the characteristics of the background in a local window around the current pixel. For LRX the data in that local window are supposed to follow a multivariate normal distribution. In order to assess the validity of this assumption, the normality was checked in the neighborhood of each target in the different scenes. The neighborhood is defined in the same way as for the actual LRX detector. Table 6 shows the same estimators of normality as well as the condition number as in Table 5 but based on a local background estimation. As there are different targets in each of the scenes, the average and standard deviation over all targets is given. One can notice that for the BJO and OSLO2 scene the local normality assumption is much better met than the global one. For OSLO1 and BJO the condition number of local Σ is also better than that of the global Σ. The large standard deviation of the condition numbers are due to some targets for which local covariance matrix has a very high condition number. For CAM, this is the case for three targets and the median of the local condition number in that scene is 5.3×10^9 (no preprocessing). For the scenes with a limited number of targets this is explored in more detail in Table 7. The table shows the values obtained for the three estimators in the local window around each of the targets. Figure 15 shows the guard window (red) and the outer window (green) used for estimating local Σ superimposed on a grayscale representation of the two scenes. The green window is the outer window used when no data reduction is applied. For BJO and OSLO2, where the guard window is 15 × 15 and the number of bands $n = 80$, this means that the outer window has a size of 33 × 33. In OSLO2, LRX is applied after KDR and only 5 kurtosis components are kept. This results in an outer window of 17 × 17 represented by the white squares in Figure 15(b).

From Table 7 it appears that the normality assumption in the BJO scene is best obeyed for the local neighborhood of targets T8–T13. T3 has the most heterogeneous background, as can be also be seen in Figure 15(a). T1 and T2 deviate from the normality assumption because of target contamination: the outer windows overlap part of the adjacent target. It is

Table 6: Correlation coefficient and Kolmogorov-Smirnov test statistic between empirical and theoretical CDF of the Mahalanobis distance around each of the targets and the condition number of local Σ.

Name	Preprocessing	Correlation mean ± std	KS statistic mean ± std	Condition number of local Σ mean ± std
CAM	None	0.95 ± 0.03	0.48 ± 0.17	1.15*e* + 10 ± 2.3*e* + 10
	SB	0.98 ± 0.01	0.28 ± 0.11	3.97*e* + 09 ± 8.4*e* + 09
	KDR	0.992 ± 0.004	0.15 ± 0.04	8.25*e* + 08 ± 2.6*e* + 08
	KDR (5 kcp)	0.991 ± 0.011	0.128 ± 0.07	6.08*e* + 01 ± 1.04*e* + 02
OSLO1	None	0.999 ± 0.0003	0.04 ± 0.009	6.33*e* + 03 ± 5.6*e* + 03
	SB	0.998 ± 0.0008	0.068 ± 0.01	3.02*e* + 03 ± 2.5*e* + 03
	KDR	0.9995 ± 0.0003	0.04 ± 0.009	3.90*e* + 01 ± 3.3*e* + 01
	KDR (5 kcp)	0.993 ± 0.003	0.11 ± 0.03	1.20*e* + 01 ± 2.0*e* + 01
BJO	None	0.994 ± 0.02	0.076 ± 0.10	7.34*e* + 04 ± 1.4*e* + 05
	SB	0.996 ± 0.006	0.089 ± 0.06	5.38*e* + 04 ± 8.6*e* + 04
	SQRT + SB	0.996 ± 0.006	0.089 ± 0.06	5.38*e* + 04 ± 8.6*e* + 04
OSLO2	None	0.996 ± 0.006	0.097 ± 0.06	3.27*e* + 04 ± 1.0*e* + 04
	SB	0.996 ± 0.005	0.102 ± 0.05	2.08*e* + 04 ± 1.8*e* + 04

Table 7: Correlation coefficient and Kolmogorov-Smirnov test statistic between empirical and theoretical CDF of the Mahalanobis distance calculated in a local window around each target and the condition number of local Σ for BJO and OSLO2, both without preprocessing.

TGT	Correlation	KS statistic	Condition number of local Σ
		BJO	
T1	0.9976	0.121	1.746*e* + 05
T2	0.9968	0.104	1.884*e* + 05
T3	0.9329	0.420	5.405*e* + 05
T4	0.9991	0.055	6.775*e* + 03
T5	0.9995	0.043	6.864*e* + 03
T6	0.9994	0.052	8.845*e* + 03
T7	0.9992	0.057	6.654*e* + 03
T8	0.9999	0.027	1.345*e* + 04
T9	0.9999	0.022	1.264*e* + 04
T10	0.9998	0.033	1.144*e* + 04
T11	0.9997	0.031	1.134*e* + 04
T12	0.9997	0.032	1.426*e* + 04
T13	0.9999	0.021	1.123*e* + 04
		OSLO2	
T1	0.9995	0.052	3.780*e* + 04
T2	0.9986	0.096	4.352*e* + 04
T3	0.9991	0.052	3.021*e* + 04
T4	0.9874	0.190	1.935*e* + 04

well known that target contamination degrades the results of detectors [42, 43]. For the targets in the shadow at the edge of the forest (T4–T7) the normality assumption is reasonably well met and the condition number of the local covariance matrix is lower than for the other targets. The outer windows for these targets fall entirely in the shadow zone.

Table 7 shows that in OSLO2 the assumption of local normality is best met for targets T1 and T3 while for T4 the assumption is less valid. The corresponding figure shows that T4 has indeed the most heterogeneous local background. There is some target contamination between T2 and T3 when no data reduction is applied, while this is not the case when KDR is applied and only five kurtosis components are used.

Acknowledgments

The authors would like to acknowledge Salvatore Resta at the University of Pisa for providing the ground truth for the OSLO2 data set. The OSLO1, OSLO2, and BJO datasets were provided by the Norwegian Defence Research Establishment (FFI). The CAM scene was provided by ESA (DAISEX experiment) through a collaboration with ONERA. The MVSA matlab code, used in TLES, is kindly provided by Professor J. Bioucas on http://www.lx.it.pt/~bioucas/code.htm.

References

[1] D. W. J. Stein, S. G. Beaven, L. E. Hoff, E. M. Winter, A. P. Schaum, and A. D. Stocker, "Anomaly detection from hyperspectral imagery," *IEEE Signal Processing Magazine*, vol. 19, no. 1, pp. 58–69, 2002.

[2] S. Matteoli, M. Diani, and G. Corsini, "A tutorial overview of anomaly detection in hyperspectral images," *IEEE Aerospace and Electronic Systems Magazine*, vol. 25, no. 7, pp. 5–27, 2010.

[3] I. S. Reed and X. Yu, "Adaptive multiple-band CFAR detection of an optical pattern with unknown spectral distribution," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 10, pp. 1760–1770, 1990.

[4] C. I. Chang and S. S. Chiang, "Anomaly detection and classification for hyperspectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 40, no. 6, pp. 1314–1325, 2002.

[5] H. Kwon and N. M. Nasrabadi, "Kernel RX-algorithm: a nonlinear anomaly detector for hyperspectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 2, pp. 388–397, 2005.

[6] A. Schaum, "Advanced methods of multivariate anomaly detection," in *Proceedings of the IEEE Aerospace Conference*, pp. 1–7, March 2007.

[7] A. P. Schaum, "Hyperspectral anomaly detection beyond RX," in *Proceedings of the SPIE Algorithms and Technologies for Multispectral, Hyperspectral and Ultraspectral Imagery XII*, vol. 6565, 2007.

[8] E. Lo and L. T. C. John Ingram, "Hyperspectral anomaly detection based on minimum generalized variance method," in *Proceedings of the Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XIV*, vol. 6966, March 2008.

[9] C.E. Caefer, J. Silverman, O. Orthal, D. Antonelli, Y. Sharoni, and S.R. Rotman, "Improved covariance matrices for point target detection in hyperspectral data," *Optical Engineering*, vol. 47, no. 7, Article ID 076402, 2008.

[10] E. Lo and A. Schaum, "A hyperspectral anomaly detector based on partialling out a clutter subspace," in *Proceedings of the Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XV*, vol. 7334, April 2009.

[11] E. Lo, "Maximized subspace model for hyperspectral anomaly detection," *Pattern Analysis and Applications*, pp. 1–11, 2011.

[12] E. Lo, "Variable subspace model for hyperspectral anomaly detection," *Pattern Analysis and Applications*, pp. 1–13, 2011.

[13] I. Kåsen, P. E. Goa, and T. Skauli, "Target detection in hyperspectral images based on multi-component statistical models for representation of background clutter," in *Proceedings of the SPIE on Electro-Optical and Infrared Systems: Technology and Applications*, vol. 5612, pp. 258–264, October 2004.

[14] D. Blumberg, E. Ohel, and S. Rotman, "Anomaly detection in noisy multi- and hyperspectral images of urban environments," in *Proceedings of the 3rd GRSS/ISPRS Symposium*, Tempe, Ariz, USA, March 2005.

[15] C. Willis, "Anomaly detection in hyperspectral imagery using statistical mixture models," in *Proceedings of the 2nd EMRS DTC Technical Conference*, Edinburgh, UK, 2005.

[16] O. Duran and M. Petrou, "A time-efficient method for anomaly detection in hyperspectral images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 12, pp. 3894–3904, 2007.

[17] O. Duran and M. Petrou, "Spectral unmixing with negative and superunity abundances for subpixel anomaly detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 6, no. 1, pp. 152–156, 2009.

[18] T. Veracini, S. Matteoli, M. Diani, and G. Corsini, "Fully unsupervised learning of Gaussian mixtures for anomaly detection in hyperspectral imagery," in *Proceedings of the 9th International Conference on Intelligent Systems Design and Applications (ISDA '09)*, pp. 596–601, December 2009.

[19] D. Borghys, E. Truyen, M. Shimoni, and C. Perneel, "Anomaly detection in hyperspectral images of complex scenes," in *Proceedings of the 29th Symposium of the European Association of Remote Sensing Laboratories*, Chania, Greece, June 2009.

[20] Y. Tarabalka, T. V. Haavardsholm, I. Kåsen, and T. Skauli, "Real-time anomaly detection in hyperspectral images using multivariate normal mixture models and GPU processing," *Journal of Real-Time Image Processing*, vol. 4, no. 3, pp. 287–300, 2009.

[21] D. Borghys, I. Kasen, V. Achard, and C. Perneel, "Comparative evaluation of hyperspectral anomaly detectors in different types of background," in *Proceedings of the SPIE Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XVIII*, vol. 8390, Baltimore, April 2012.

[22] C. I. Chang, "Orthogonal subspace projection (OSP) revisited: a comprehensive study and analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 3, pp. 502–518, 2005.

[23] P. Masson and W. Pieczynski, "SEM algorithm and unsupervised statistical segmentation of satellite images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 31, no. 3, pp. 618–633, 1993.

[24] J. Li and J. M. Bioucas-Dias, "Minimum volume simplex analysis: a fast algorithm to unmix hyperspectral data," in *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*, pp. III250–III253, July 2008.

[25] S. S. Shen and E. M. Bassett, "Information theory based band selection and utility evaluation for reflective spectral systems," in *Proceedings of the SPIE Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery VIII*, vol. 4725, pp. 18–29, April 2002.

[26] I. Kåsen, A. Rødningsby, T. V. Haavardsholm, and T. Skauli, "Band selection for hyperspectral target detection based on a multinormal mixture anomaly detection algorithm," in *Proceedings of the SPIE Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XIV*, vol. 6966, March 2008.

[27] D. Peña, F. J. Prieto, and J. Viladomat, "Eigenvectors of a kurtosis matrix as interesting directions to reveal cluster structure," *Journal of Multivariate Analysis*, vol. 101, no. 9, pp. 1995–2007, 2010.

[28] E. A. Ashton, B. D. Wemett, R. A. Leathers, and T. V. Downes, "A novel method for illumination suppression in hyperspectral images," in *Proceedings of the SPIE Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XIV*, vol. 6966, March 2008.

[29] R. Richter and A. Müller, "De-shadowing of satellite/airborne imagery," *International Journal of Remote Sensing*, vol. 26, no. 15, pp. 3137–3148, 2005.

[30] S. M. Adler-Golden, M. W. Matthew, G. P. Anderson, G. W. Felde, and J. A. Gardner, "An algorithm for de-shadowing spectral imagery," in *Proceedings of the 11th JPL Airborne Earth Science Workshop*, pp. 3–4, March 2002.

[31] M. Shimoni, G. Tolt, C. Perneel, and J. Ahlberg, "Detection of vehicles in shadow areas using combined hyperspectral and lidar data," in *Proceedings of the Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 4427–4430, Vancouver, Canada, July 2011.

[32] B. D. Wemett, J. K. Riek, and R. A. Leathers, "Dynamic thresholding for hyperspectral shadow detection using Levenberg-Marquardt minimization on multiple gaussian illumination distributions," in *Proceedings of the SPIE Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XV*, vol. 7334, April 2009.

[33] T. Skauli, "Sensor noise informed representation of hyperspectral data, with benefits for image storage and processing," *Optics Express*, vol. 19, no. 14, pp. 13031–13046, 2011.

[34] P. McCullagh and J. Nelder, *Generalized Linear Models*, chapter 6, CRC Press, 1999.

[35] J. Harsanyi, W. Farrand, and C. Chang, "Determining the number and identity of spectral endmembers: an integrated approach using neyman pearson eigenthresholding and iterative constrained rms error minimization," in *Proceedings of the 9th Thematic Conference on Geologic Remote Sensing*, February 1993.

[36] C. I. Chang and Q. Du, "Estimation of number of spectrally distinct signal sources in hyperspectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 3, pp. 608–619, 2004.

[37] J. Bioucas-Diaz and M. Nascimiento, "Hyperspectral subspace identification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 8, pp. 2435–2445, 2008.

[38] N. Acito, M. Diani, and G. Corsini, "A new algorithm for robust estimation of the signal subspace in hyperspectral images in the presence of rare signal components," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 11, pp. 3844–3856, 2009.

[39] N. Acito, "Hyperspectral signal subspace identification in the presence of rare signal components," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 4, pp. 1940–1954, 2010.

[40] B. Manly, Ed., *Multivariate Statistical Methods*, Chapman and Hall, Boca Raton, Fla, USA, 1995.

[41] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of Multivariate Analysis*, vol. 88, no. 2, pp. 365–411, 2004.

[42] S. Matteoli, M. Diani, and G. Corsini, "Different approaches for improved covariance matrix estimation in hyperspectral anomaly detection," in *Proceedings of the Annual Meeting of the Italian National Telecommunications and Information Theory Group (GTTI '09)*, 2009.

[43] J. Theiler and B. R. Foy, "Effect of signal contamination in matched-filter detection of the signal on a cluttered background," *IEEE Geoscience and Remote Sensing Letters*, vol. 3, no. 1, pp. 98–102, 2006.

Hindawi Publishing Corporation
Journal of Electrical and Computer Engineering
Volume 2012, Article ID 818175, 8 pages
doi:10.1155/2012/818175

Research Article

Non-Gaussian Linear Mixing Models for Hyperspectral Images

Peter Bajorski

Graduate Statistics Department and Center for Imaging Science, Rochester Institute of Technology, Rochester, NY, USA

Correspondence should be addressed to Peter Bajorski, pxbeqa@rit.edu

Received 17 June 2012; Accepted 10 September 2012

Academic Editor: Prudhvi Gurram

Modeling of hyperspectral data with non-Gaussian distributions is gaining popularity in recent years. Such modeling mostly concentrates on attempts to describe a distribution, or its tails, of all image spectra. In this paper, we recognize that the presence of major materials in the image scene is likely to exhibit nonrandomness and only the remaining variability due to noise, or other factors, would exhibit random behavior. Hence, we assume a linear mixing model with a structured background, and we investigate various distributional models for the error term in that model. We propose one model based on the multivariate t-distribution and another one based on independent components following an exponential power distribution. The former model does not perform well in the context of the two images investigated in this paper, one AVIRIS and one HyMap image. On the other hand, the latter model works reasonably well with the AVIRIS image and very well with the HyMap image. This paper provides the tools that researchers can use for verifying a given model to be used with a given image.

1. Introduction

The following linear mixing model (with a structured background) is often used in hyperspectral imaging literature [1–4]:

$$\mathbf{r} = \mathbf{B} \cdot \boldsymbol{\alpha} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{r}$ is a p-dimensional vector (e.g., of reflectance or radiance,) of a pixel spectrum, $\mathbf{B}$ is a fixed matrix of spectra of m materials present in the image (as columns $\mathbf{b}_j$, $j = 1, \ldots, m$), and $\boldsymbol{\alpha}$ is an unknown vector of material abundances. The error term $\boldsymbol{\varepsilon}$ is often assumed to follow the multivariate normal (Gaussian) distribution $N(\mathbf{0}, \sigma^2\mathbf{I})$ or some other distributional assumptions can be made here.

In this paper, we want to address the question whether model (1) is realistic for hyperspectral images and what type of a distribution should be used for the error term $\boldsymbol{\varepsilon}$. In previous research [5], we provided a preliminary investigation of the marginal distributions of $\boldsymbol{\varepsilon}$ to be modeled by the exponential power distribution. We expand this research here to model the multivariate structure of $\boldsymbol{\varepsilon}$ and to propose some other models such as the multivariate t-distribution.

The multivariate t-distribution is a popular distribution for modeling hyperspectral data (see [6–8]). In the current literature, this distribution is mostly used for modeling the variability of background materials in a purely stochastic model without a deterministic part such as the one defined in model (1). Here we try to use that distribution for modeling the error term $\boldsymbol{\varepsilon}$, while the major part of the image variability is explained by the deterministic part $\mathbf{B} \cdot \boldsymbol{\alpha}$ of the model.

In Section 2, we introduce our notation and show how the deterministic and the stochastic parts of the model are constructed based on the singular value decomposition (SVD). We also provide details on an AVIRIS hyperspectral image used for the numerical results performed in this paper. In Section 3, we explore potential models for the marginal distributions of the error term. In Section 4, we explore potential models for the joint multivariate distributions of the error term. The AVIRIS image is used for numerical examples in Sections 3 and 4. In Section 5, we provide an additional example using a subset of the HyMap Cooke City image. Conclusions are formulated in Section 6.

2. Preliminary Considerations

In this paper, we are going to assume that the abundances of all materials sum up to 1, that is,

$$\sum_{j=1}^{m} \alpha_j = 1, \tag{2}$$

where α_i are coordinates of the vector $\boldsymbol{\alpha}$. In practice, this assumption may not necessarily be strictly fulfilled due to imperfections in estimation of background signatures and in the linearity of the spectral mixing process. However, it is reasonable to assume that (2) is fulfilled at least approximately. We do not specifically address another reasonable assumptions that $0 \leq \alpha_j \leq 1$, but an appropriate choice of the spectral signatures in $\mathbf{B}$ should result in positive (or almost positive) α_j.

We also assume that no a priori information is available about the spectral signatures in $\mathbf{B}$, that is, we are trying to "guess" all terms on the right-hand side of (1). Let us now assume that we have a hyperspectral image with n pixels, and the model (1) takes the form

$$\mathbf{r}_i = \mathbf{B} \cdot \boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i, \quad (3)$$

for $i = 1, \ldots, n$. Denote the average of all pixel spectra by $\bar{\mathbf{r}} = \sum_{i=1}^{n} \mathbf{r}_i$. Let $\mathbf{B}^*$ be a matrix of vectors $(\mathbf{b}_j - \bar{\mathbf{r}})$, $j = 1, \ldots, m$ as columns. Because of the property (2), the model (3) is equivalent to the following model centered at $\bar{\mathbf{r}}$:

$$\mathbf{r}_i - \bar{\mathbf{r}} = \mathbf{B}^* \cdot \boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i. \quad (4)$$

Let us denote by $\mathbf{X}$ an n by p matrix of vectors $(\mathbf{r}_i - \bar{\mathbf{r}})$, $i = 1, \ldots, n$ and write its SVD as

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T, \quad (5)$$

where $\mathbf{U}$ is an n by p matrix (with columns $\mathbf{u}_j$, $j = 1, \ldots, p$), $\mathbf{D}$ is a diagonal p by p matrix of singular values d_j, $j = 1, \ldots, p$, and $\mathbf{V}$ is an orthogonal p by p matrix (with columns $\mathbf{v}_j$, $j = 1, \ldots, p$). We assume that $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$. The SVD in (5) can also be written as

$$\mathbf{X} = \sum_{j=1}^{p} d_j \mathbf{u}_j \mathbf{v}_j^T. \quad (6)$$

In order to build a model of the form (4), we now want to identify m out of the total of p basis vectors $\mathbf{v}_j$ such that $\sum_{j=1}^{m} d_j \mathbf{u}_j \mathbf{v}_j^T$ represents the deterministic part $\mathbf{B}^* \cdot \boldsymbol{\alpha}_i$ of the model (4). This deterministic part will be selected based on the idea that the major materials present in the image exhibit nonrandom behavior, while the remaining variability due to noise, or due to some undetected small amounts of materials, or due to other imperfections in the model, should exhibit more random behavior. We want the remaining sum $\sum_{j=m+1}^{p} d_j \mathbf{u}_j \mathbf{v}_j^T$ to represent realizations of the error term $\boldsymbol{\varepsilon}$. We will investigate that sum in the basis system defined by the vectors $\mathbf{v}_j$, that is, realizations of a random vector that we denote by $\mathbf{Y}_{m+1} = (Y_{m+1}, \ldots, Y_p)$. Note that each n dimensional vector $d_j \mathbf{u}_j$, $j = (m+1), \ldots, p$ represents a sample from the distribution of Y_j. We want to study those samples so that an appropriate distributional assumption about the error terms can be made. We can define the standardized form $Z_j = Y_j/\text{St.Dev.}(Y_j)$ with the sample of realizations given by $\sqrt{n-1}\,\mathbf{u}_j$. The vector $\mathbf{Z}_{m+1} = (Z_{m+1}, \ldots, Z_p)$ represents the error term $\boldsymbol{\varepsilon}$ in the sphered, or whitened, coordinates. The marginal (uncorrelated) distributions of $\mathbf{Z}_{m+1}$ will be studied in the next section followed by modeling the joint distribution in Section 4.

Figure 1: Color rendering of the cluttered AVIRIS urban scene in Rochester, NY, USA, used in this paper.

Numerical results in Sections 3 and 4 use a 100 by 100 pixel (so $n = 10{,}000$) AVIRIS urban image in Rochester, NY, USA, near the Lake Ontario shoreline (see Figure 1). The scene has a wide range of natural and man-made clutter including a mixture of commercial/warehouse and residential neighborhoods to add a wide range of spectral diversity. Prior to processing, invalid bands, due to atmospheric water absorption, were removed reducing the overall dimensionality to $p = 152$ spectral bands. This image was used earlier in [9, 10].

3. Modeling Marginal Distributions

Here we investigate two families of distributions as models for the distributions of Z_j's. The first one is the exponential power distribution (also called general error or general Gaussian distribution) with the location parameter μ, the scale parameter $\beta > 0$, and the shape parameter $\lambda > 0$. Its density is defined by

$$f(x) = \frac{\lambda}{2\beta\Gamma(1/\lambda)} \exp\left[-\left(\frac{|x-\mu|}{\beta}\right)^{\lambda}\right] \quad \text{for } x \in \mathbb{R}, \quad (7)$$

where $\Gamma(s) = \int_0^\infty t^{s-1} e^{-t} dt$ for $s > 0$ is the gamma function. This is a flexible family of distributions. Its special cases are the Gaussian distribution ($\lambda = 2$) and the Laplace distribution ($\lambda = 1$). We assume that $\mu = 0$ since the distribution of Z_j is already centered. For each Z_j, $j = 1, \ldots, p$, the remaining parameters, λ and β, were estimated using the maximum likelihood principle. The fit of data to the resulting exponential power distribution was then checked with the chi-square test based on the statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(N_i - n \cdot p_i)^2}{n \cdot p_i}, \quad (8)$$

where N_i is the count of observations in the ith interval ($i = 1, \ldots, k = 25$), and p_i is the interval probability based on the testing distribution. The intervals were chosen so that $p_i = 1/k$, $i = 1, \ldots, k$. The tested distribution should be rejected when $\chi^2 \geq \chi^2_{k-1-m}(\alpha/p)$, where $\chi^2_{k-1-m}\ (\alpha)$ is the upper $100 \cdot$

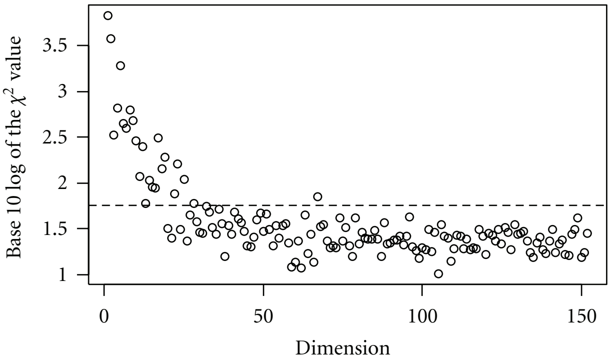


FIGURE 2: Base 10 log values of the chi-square statistic (for testing the fit to the exponential power distribution) plotted versus the dimension number j (for the AVIRIS image).

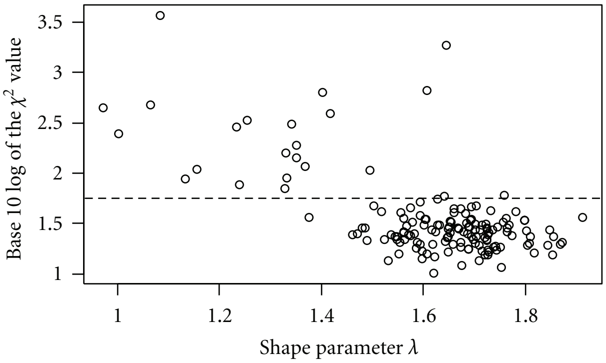


FIGURE 3: Base 10 log values of the chi-square statistic (for testing the fit to the exponential power distribution) plotted versus the shape parameter λ (for the AVIRIS image).

αth percentile from the chi-squared distribution with $(k - 1 - m)$ degrees of freedom, and m is the number of estimated parameters. The value α/p is chosen based on the Bonferroni principle since p inferences are performed here. Figure 2 shows the base 10 log values of the chi-square statistic plotted versus the dimension number j. The horizontal line is at the threshold level of $\chi^2_{25-1-2}(0.01/152) = 56.79$ (1.75 on the log scale). The points above the threshold represent the first 19 dimensions and then 22, 23, 25, 28, and 67. Since not all dimension numbers are consecutive, we are faced with a difficulty of choosing a suitable value for m. We propose two potential strategies.

(1) Select as m the largest dimension number (here 28) that is not an "outlier" (such as dimension 67 here). With this approach, we would accept the imperfect modeling along the 67th dimension, and some of the dimensions (20, 21, 24, 26, and 27) would be represented in the deterministic part of the model, even though they could be modeled stochastically by the exponential power distribution. This approach is consistent with the principle of keeping the dimensions with the highest explained variability.

(2) Select all dimensions above the threshold for the deterministic part of the model (possibly including dimension 67) and use the remaining dimensions for the error term. From the point of view of notation in Section 2, we would reorder the dimensions, so that the "non-random" dimensions 22, 23, 25, 28, and 67 would be numbered from 20 to 24, and m would be chosen as 24.

Figure 3 shows the base 10 log values of the chi-square statistic plotted versus λ. For the values below the threshold, where the exponential power distribution model would be used, the λ values range from 1.37 to 1.91. Hence, the distributions have tails heavier than the Gaussian distribution, but not as heavy as those of the Laplace distribution.

The second potential family of distributions for modeling the distributions of Z_j's is the t-distribution with ν degrees of freedom given by the density

$$f(x) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\pi\nu}\,\Gamma(\nu/2)}\left(1+\frac{x^2}{\nu}\right)^{-(\nu+1)/2} \quad \text{for } x \in \mathbb{R}. \tag{9}$$

Since Z_j is scaled to have variance 1, we also need to scale the t-distribution. Hence, we fit a distribution with the density $g(z) = \gamma f(\gamma z)$, where f is defined in (9) and $\gamma = \sqrt{\nu/(\nu-2)}$. We assume that ν is larger than 2 but is not necessarily an integer. The scaling depends on the unknown parameter ν, so it needs to be taken into account in the maximum likelihood estimation of ν. The fit of data to the scaled t-distribution was again checked with the chi-square test. Figure 4 is analogous to Figure 2, but here the chi-square statistic measures the fit to the scaled t-distribution. The horizontal line is at the threshold level of $\chi^2_{25-1-1}(0.01/152) = 58.36$ (1.77 on the log scale). The points above the threshold represent the first 10 dimensions and then 12, 14, 17, 19, and 23. Here again, we can use one of the proposed two strategies for identifying the dimensionality m of the deterministic part of the model. We can see that the scaled t-distribution gives a significantly better fit than the exponential power distribution.

Figure 5 shows the base 10 log values of the chi-square statistic plotted versus the number of degrees of freedom ν. (The parameter ν had a very large value for the first dimension (Z_1 with the highest chi-squared value), and it is not shown in the graph for clarity of presentation.) For the values below the threshold, where the scaled t-distribution model would be used, the ν values range from 4.3 to 43.3. Hence again, the distributions have tails heavier than the Gaussian distribution, but not heavier than those of the scaled t-distribution with 4 degrees of freedom.

4. Modeling Joint Multivariate Distributions

The model fitting discussed in the previous section accounted only for the marginal distributions of Z_j's. A more challenging task is to ensure a fit of the joint multivariate

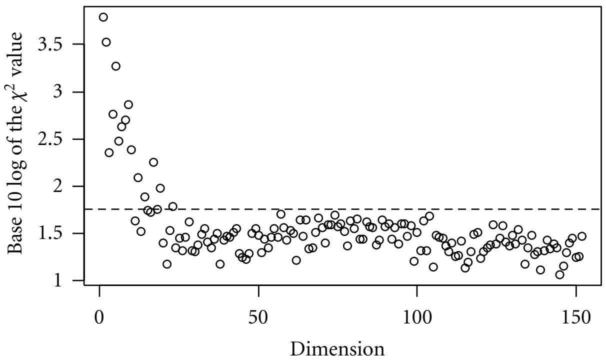


Figure 4: Base 10 log values of the chi-square statistic (for testing the fit to the scaled t-distribution) plotted versus the dimension number j (for the AVIRIS image).

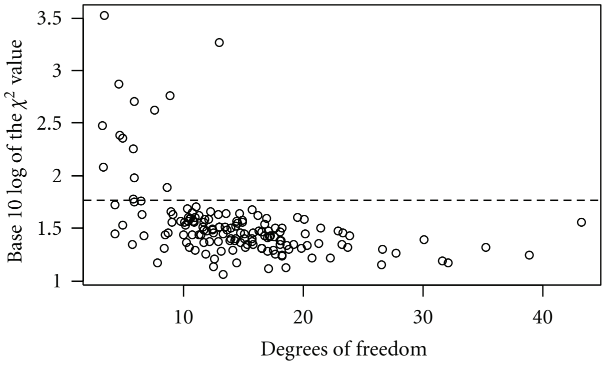


Figure 5: Base 10 log values of the chi-square statistic (for testing the fit to the scaled t-distribution) plotted versus the number of degrees of freedom ν (for the AVIRIS image).

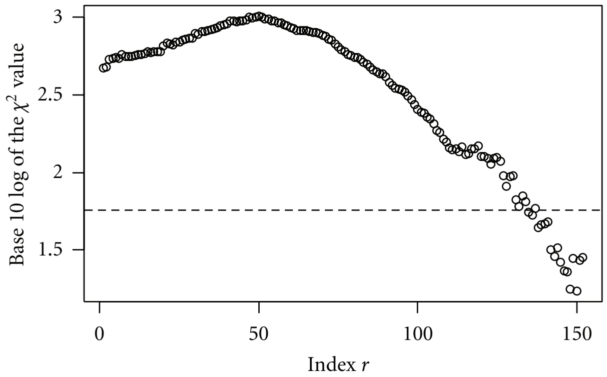


Figure 6: Base 10 log values of the chi-square statistic (for testing the multivariate fit to $G_r(x)$) plotted versus r (for the AVIRIS image).

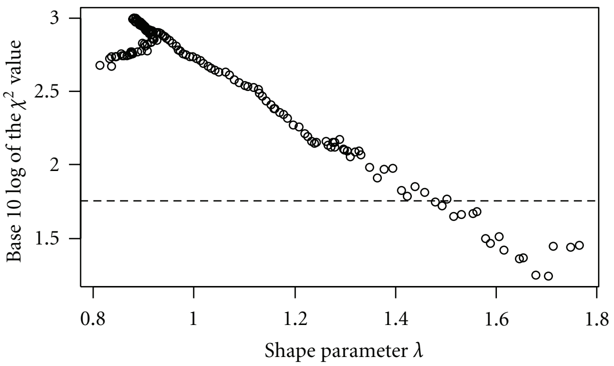


Figure 7: Base 10 log values of the chi-square statistic (for testing the multivariate fit to $G_r(x)$) plotted versus the shape parameter λ (for the AVIRIS image).

distribution of data to a suitable model. In the whitened version, the components of $\mathbf{Z}_r = (Z_r, \dots, Z_p)$ are uncorrelated, but they might be either stochastically independent or dependent, which depends on a specific assumed model. Again, we will use two competing models. The first one is based on the assumption of independent Z_j components following an exponential power distribution. In order to verify this model, we investigate the joint distribution of $\mathbf{Z}_r = (Z_r, \dots, Z_p)$ for $r = 1, \dots, p$. More specifically, we investigate the fit of the theoretical cumulative distribution function (CDF) $G_r(x) = P\{Z_r \leq x, Z_{r+1} \leq x, \dots, Z_p \leq x\}$ to its empirical equivalent. Note that $G_r(x)$ can be regarded as a univariate simplification of the multivariate CDF of $\mathbf{Z}_r$ with all coordinates being equal. Based on the assumption of independence, $G_r(x) = F(x)^{p-r+1}$, where $F(x)$ is the CDF of the exponential power distribution defined in (7). The empirical equivalent of $G_r(x)$ is the fraction of observations (realizations) such that $Z_j \leq x$, $j = r, \dots, p$, or equivalently $\max_{r \leq j \leq p} Z_j \leq x$. The CDF $G_r(x)$ depends on the parameters $\beta > 0$ and $\lambda > 0$ that were estimated using the maximum likelihood principle. The chi-square test based on the statistic (8) was then used to assess the fit of the observations of $\max_{r \leq j \leq p} Z_j$ to $G_r(x)$. The resulting base 10 log values of the chi-squared statistic are shown in Figure 6. As before, the horizontal line is at the threshold level of $\chi^2_{25-1-2}(0.01/152) = 56.79$ (1.75 on the log scale). The points below the threshold represent the dimensions 135 and 136, and then 138 through 152 ($= p$). The value for $G_{137}(x)$ is only slightly above the threshold. We can say that the model of independent components with an exponential power distribution is quite successful in modeling the multivariate distribution of the "last" 18 dimensions from 135 until the last one (152). However, the model is not very successful in modeling further dimensions, where we apparently observe significant dependencies among Z_j's. This is consistent with low-dimensional components of $\mathbf{Z}_r$ being independent or having very weak dependence structure that does not show up as significant. That dependence structure gets stronger with higher dimensions.

Figure 7 shows the base 10 log values of the chi-square statistic plotted versus λ. For the values below the threshold, where a model with independent exponential power distributions would be used, the λ values range from 1.48 to 1.77. Hence again, the distributions have tails heavier than the Gaussian distribution, but not as heavy as those of the Laplace distribution.

The second multivariate model that we want to discuss is based on the standard d-dimensional multivariate t-distribution with the density function

$$f(\mathbf{x}) = \frac{\Gamma((\nu+d)/2)}{(\pi\nu)^{d/2}\Gamma(\nu/2)}\left(1+\frac{\mathbf{x}^T\mathbf{x}}{\nu}\right)^{-(\nu+d)/2} \quad \text{for } \mathbf{x} \in \mathbb{R}^d. \tag{10}$$

The variance-covariance matrix of this distribution is equal to $\gamma^2 \cdot \mathbf{I}_d$, where $\mathbf{I}_d$ is a d-dimensional identity matrix and $\gamma = \sqrt{\nu/(\nu-2)}$. Since we deal here with sphered data modeled by $\mathbf{Z}_r = (Z_r, \dots, Z_p)$ (with Var $(\mathbf{Z}_r) = \mathbf{I}_d$ and $d = p-r+1$), we want $\gamma\mathbf{Z}_r$ to follow the standard multivariate t-distribution. Hence, an appropriate candidate distribution for $\mathbf{Z}_r$ is a scaled multivariate t-distribution with the density function given by $g(\mathbf{z}) = \gamma^d f(\gamma\mathbf{z})$, where f is defined in (10). We can write

$$g(\mathbf{z}) = \frac{\Gamma((\nu+d)/2)}{[\pi(\nu-2)]^{d/2}\Gamma(\nu/2)}\left(1+\frac{\mathbf{z}^T\mathbf{z}}{\nu-2}\right)^{-(\nu+d)/2} \quad \text{for } \mathbf{z} \in \mathbb{R}^d. \tag{11}$$

This distribution is spherically contoured in the sphered coordinates and elliptically contoured in the original coordinates. All marginal distributions of the scaled multivariate t-distribution are the scaled t-distributions discussed in Section 3 that were already confirmed as reasonable models for the AVIRIS data. Here, we want to verify if the multivariate t-distribution provides an adequate multivariate structure for those data.

If $\gamma\mathbf{Z}_r$ follows the standard multivariate t-distribution, then $\|\gamma\mathbf{Z}_r\|^2/d$ follows the F-distribution with d and ν degrees of freedom (see [11]). Hence, in order to check the assumption of the multivariate t-distribution, we verify if $\|\mathbf{Z}_r\|^2/d$ follows the F-distribution scaled by $1/\gamma^2$. As before, the degrees of freedom parameter ν is estimated based on the maximum likelihood principle, and the fit is assessed based on the chi-square statistic. Figure 8 shows the base 10 log values of the chi-square statistic plotted versus r. The horizontal line is at the threshold level of $\chi^2_{25-1-1}(0.01/152) = 58.36$ (1.77 on the log scale). We can see only one value below the threshold, suggesting the scaled F-distribution is suitable only for $\|\mathbf{Z}_{152}\|^2$, which is consistent with the marginal scaled t-distribution for Z_{152} (discussed in Section 3) since the vector $\mathbf{Z}_{152}$ is one-dimensional. All remaining $\mathbf{Z}_r$, $r < p = 152$, are apparently not modeled well by the scaled multivariate t-distribution. We note that the components of the multivariate t-distribution are not independent, and the dependency structure proposed by this distribution is apparently not consistent with the AVIRIS hyperspectral data investigated here.

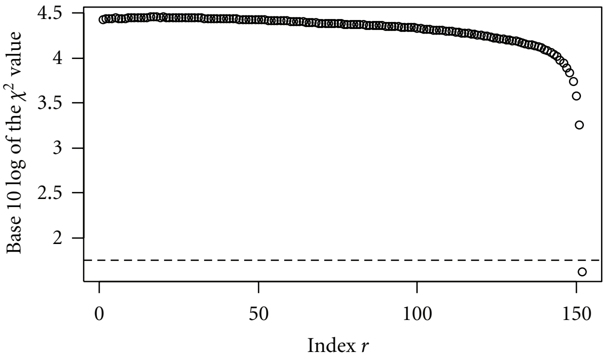


Figure 8: Base 10 log values of the chi-square statistic (for testing the multivariate t-distribution) plotted versus r (for the AVIRIS image).

5. Cooke City Image

In the two previous sections, we used an AVIRIS image as an example to demonstrate how to fit a linear mixing model with the $\boldsymbol{\varepsilon}$ term being potentially non-Gaussian. Specific numerical results might be different, of course, for other images. Hence, it is interesting to perform the same calculations on another image. The AVIRIS image has a wide range of spectral diversity due to the presence of various natural and man-made materials. Hence, it would be interesting to try a more homogenous dataset with less variety. One way to do this could be to classify an image into various types of cover material, and then use spectra from one class as our dataset. A disadvantage of such an approach is the possibility of removing tails of distributions, which might have a tendency of being assigned to a different class. Hence, we chose a relatively uniform subimage of forest in the HyMap Cooke City (see [12]) image as marked by a red rectangle shown in Figure 9. Four spectral bands were also removed due to some suspicious values. The spectral dimensionality of the dataset used is then $p = 122$, and the spatial dimensionality is 50 by 300 pixels for a total of $n = 15{,}000$ pixels.

In order to investigate the marginal distributions, we now follow the ideas and notation of Section 3. The fit with the exponential power distribution is verified with Figure 10, which is analogous to Figure 2. Using Strategy 1 explained in Section 3, we would identify 26 as the dimensionality of the deterministic part of the model.

Figure 11 (analogous to Figure 3) shows the base 10 log values of the chi-square statistic plotted versus λ. The highest value of λ is almost perfectly equal to 2, suggesting the Gaussian distribution. Most of the λ values are around 1.8 and higher, suggesting slightly lighter tails than those for the AVIRIS image (compare with Figure 3).

Figure 12 (analogous to Figure 4), shows a slightly better fit of the scaled t-distribution than that of the exponential power distribution. However, the resulting dimensionality would be identified as the same as before (26).

Figure 13 (analogous to Figure 5) shows the base 10 log values of the chi-square statistic plotted versus the number of degrees of freedom ν. The high values of ν again point to more Gaussian-like marginal distributions than those of the AVIRIS image.

So far, we discussed only the marginal distributions of Z_j's as a precursor to checking the model fit. We now want to consider a more challenging task to ensure a fit of the joint multivariate distribution of data to a suitable model

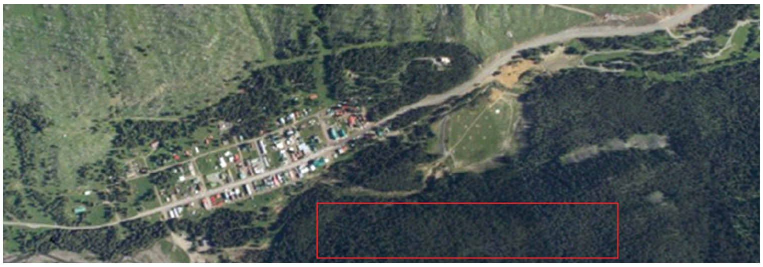


Figure 9: Color rendering of the 280 by 800 pixel HyMap Cooke City image (see [12] for details about the image).

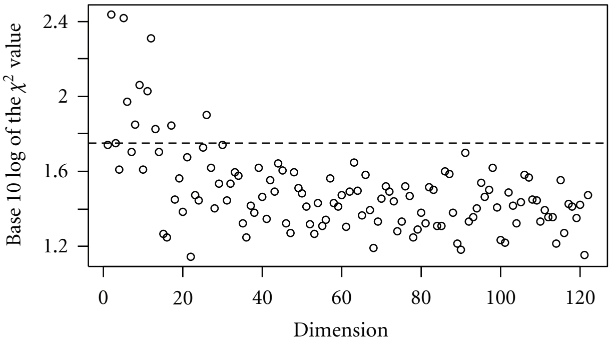


Figure 10: Base 10 log values of the chi-square statistic (for testing the fit to the exponential power distribution) plotted versus the dimension number j (for the Cooke City image).

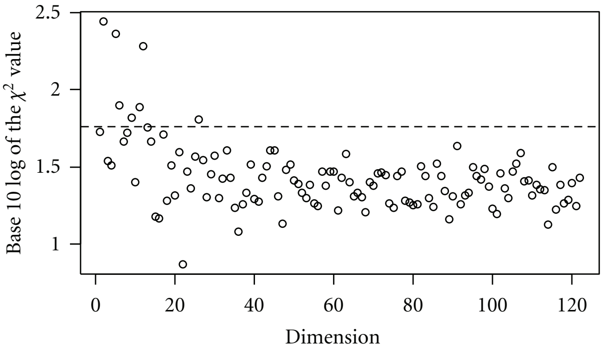


Figure 12: Base 10 log values of the chi-square statistic (for testing the fit to the scaled t-distribution) plotted versus the dimension number j (for the Cooke City image).

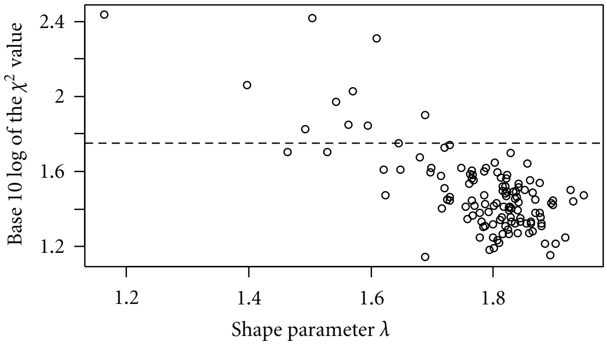


Figure 11: Base 10 log values of the chi-square statistic (for testing the fit to the exponential power distribution) plotted versus the shape parameter λ (for the Cooke City image).

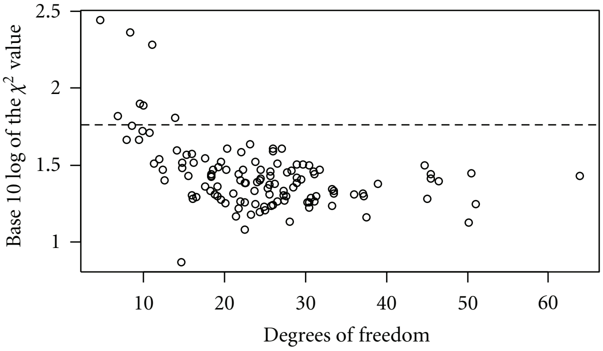


Figure 13: Base 10 log values of the chi-square statistic (for testing the fit to the scaled t-distribution) plotted versus the number of degrees of freedom ν (for the Cooke City image).

as discussed in Section 4. Figure 14 (analogous to Figure 6) shows the base 10 log values of the chi-square statistic (for testing the multivariate fit to $G_r(x)$ representing independent exponential power distributions) plotted versus r. This figure is so strikingly different from the analogous Figure 6, that the author double checked the code (in fact the same code was used in both cases). Note that even for r as small as 1 (which represents all dimensions from 1 until $p = 122$), we obtain an acceptable fit with independent exponential power distributions. The only unacceptable fit is for $k = 78$, which points to some minor issues with the model.

Regarding the final conclusion about the model fit, we need to keep in mind that this is just one test of multivariate fit, and it needs to be considered together with the results shown in Figure 12 telling us that some of the marginal distributions do not have the satisfactory fit. Hence, we can accept the previous conclusion of the dimensionality of the deterministic part of the model as 26, and the remaining dimensions are remarkably well modeled by independent exponential power distributions. This successful modeling might be largely due to this dataset being a fairly homogenous set of spectra (forest area in the Cooke City image).

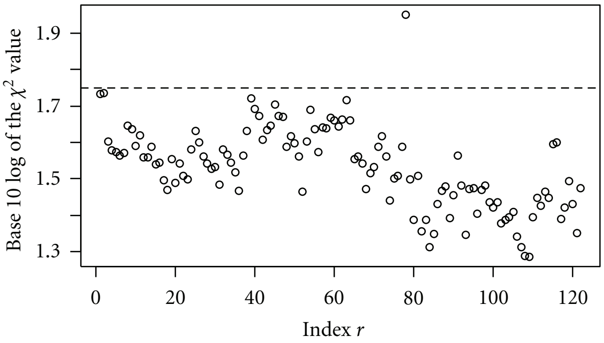


Figure 14: Base 10 log values of the chi-square statistic (for testing the multivariate fit to $G_r(x)$) plotted versus r (for the Cooke City image).

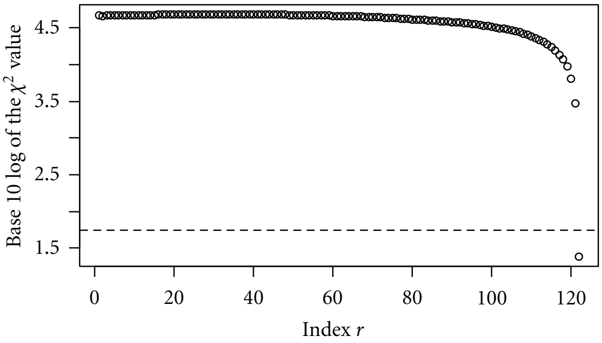


Figure 16: Base 10 log values of the chi-square statistic (for testing the multivariate t-distribution) plotted versus r (for the Cooke City image).

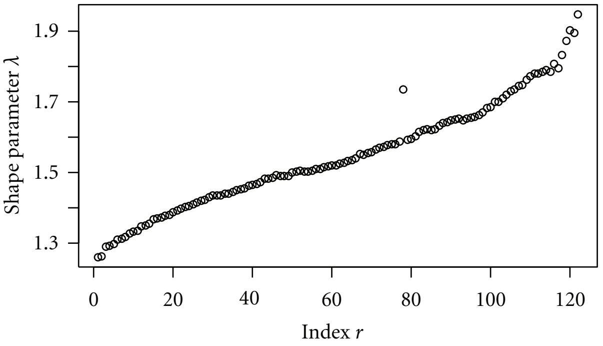


Figure 15: The shape parameter λ (from the multivariate fit to $G_r(x)$) plotted versus r (for the Cooke City image).

Since almost all chi-square values in Figure 14 are not significant, it would not be interesting to show an analog of Figure 7. Hence, in Figure 15, we created a plot of the shape parameter λ values (from the multivariate fit to $G_r(x)$) versus r. We can see that for large r, the fit is fairly close to the Gaussian distribution (λ close to 2). Then for smaller r, λ's are getting smaller, which represents much heavier tails of the distribution, almost up to the Laplace distribution (λ close to 1).

We have also checked the fit to the multivariate t-distribution as shown in Figure 16, which is an analog of Figure 8. We can observe the only good fit at $r = p = 122$ (which is really a univariate fit of the last dimension), which means that none of the multivariate t-distributions fits well to the data.

6. Conclusions

In this paper, we investigated various distributional models for the error term in the linear mixing model with a structured background. The models were tested on two datasets. One was an AVIRIS image and the other one was a subimage of a forest area in a HyMap Cooke City image. The first proposed model was based on independent components following an exponential power distribution. The model fitted reasonably well to both datasets in terms of modeling marginal distributions. For the AVIRIS image, the fit of the joint distribution worked quite well for a small number of components, but not for a larger number. For the forest area in the Cooke City image, the fit with the joint distribution of independent exponential power distributions worked very well. This successful modeling might be largely due to this dataset being a fairly homogenous set of spectra.

The second model was based on the multivariate t-distribution. It performed well in terms of the resulting marginal distributions in both datasets, but the dependency structure imposed by this distribution was entirely inconsistent with both datasets. More research is needed to investigate the two models on other hyperspectral images. However, the multivariate t-distribution model does not look promising at this point, while the exponential power distribution model seems to have more potential.

References

[1] D. G. Manolakis and G. Shaw, "Detection algorithms for hyperspectral imaging applications," *IEEE Signal Processing Magazine*, vol. 19, no. 1, pp. 29–43, 2002.

[2] L. L. Scharf and B. Friedlander, "Matched subspace detectors," *IEEE Transactions on Signal Processing*, vol. 42, no. 8, pp. 2146–2156, 1994.

[3] S. Johnson, "The relationship between the matched-filter operator and the target signature space-orthogonal projection classifier," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 38, no. 1, pp. 283–286, 2000.

[4] P. Bajorski, "Analytical comparison of the matched filter and orthogonal subspace projection detectors for hyperspectral images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 7, part 2, pp. 2394–2402, 2007.

[5] P. Bajorski, "A family of distributions for the error term in linear mixing models for hyperspectral images," in *Imaging Spectrometry XIII*, S. S. Shen and P. E. Lewi, Eds., vol. 7086 of *Proceedings of SPIE*, August 2008.

[6] D. B. Marden and D. Manolakis, "Using elliptically contoured distributions to model hyperspectral imaging data and generate statistically similar synthetic data," in *Algorithms and Technologies for MultiSpectral, Hyperspectral, and Ultraspectral Imagery X*, S. S. Shen and P. E. Lewi, Eds., vol. 5425 of *Proceedings of SPIE*, pp. 558–572, April 2004.

[7] D. Manolakis, M. Rossacci, J. Cipar, R. Lockwood, T. Cooley, and J. Jacobson, "Statistical characterization of natural hyperspectral backgrounds using t-elliptically contoured distributions," in *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XI*, S. S. Shen and P. E. Lewi, Eds., vol. 5806 of *Proceedings of SPIE*, pp. 56–65, April 2005.

[8] J. Theiler and C. Scovel, "Uncorrelated versus independent Elliptically-contoured distributions for anomalous change detection in hyperspectral imagery," in *Computational Imaging VII*, C. A. Bouman, E. L. Miller, and I. Pollak, Eds., vol. 7246 of *Proceedings of SPIE-IS & T Electronic Imaging*, p. 72460T, January 2009.

[9] P. Bajorski, "Generalized detection fusion for hyperspectral images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 4, pp. 1199–11205, 2012.

[10] P. Bajorski, *Statistics For Imaging, Optics, and Photonics*, John Wiley & Sons, New York, NY, USA, 2011.

[11] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, John Wiley & Sons, New York, NY, USA, 3rd edition, 2003.

[12] D. Snyder, J. Kerekes, I. Fairweather, R. Crabtree, J. Shive, and S. Hager, "Development of a web-based application to evaluate target finding algorithms," in *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS '08)*, vol. 2, pp. 915–918, Boston, Mass, USA, July 2008.

Hindawi Publishing Corporation
Journal of Electrical and Computer Engineering
Volume 2012, Article ID 409357, 15 pages
doi:10.1155/2012/409357

Research Article

Randomized SVD Methods in Hyperspectral Imaging

Jiani Zhang,[1] Jennifer Erway,[1] Xiaofei Hu,[1] Qiang Zhang,[2] and Robert Plemmons[3]

[1] *Department of Mathematics, Wake Forest University, Winston-Salem, NC 27109, USA*
[2] *Department of Biostatistical Sciences, Wake Forest School of Medicine, Winston-Salem, NC 27157, USA*
[3] *Departments of Mathematics and Computer Science, Wake Forest University, Winston-Salem, NC 27109, USA*

Correspondence should be addressed to Robert Plemmons, plemmons@wfu.edu

Received 16 May 2012; Accepted 2 August 2012

Academic Editor: Heesung Kwon

We present a randomized singular value decomposition (rSVD) method for the purposes of lossless compression, reconstruction, classification, and target detection with hyperspectral (HSI) data. Recent work in low-rank matrix approximations obtained from random projections suggests that these approximations are well suited for randomized dimensionality reduction. Approximation errors for the rSVD are evaluated on HSI, and comparisons are made to deterministic techniques and as well as to other randomized low-rank matrix approximation methods involving compressive principal component analysis. Numerical tests on real HSI data suggest that the method is promising and is particularly effective for HSI data interrogation.

1. Introduction

Hyperspectral imagery (HSI) data are measurements of the electromagnetic radiation reflected from an object or scene (i.e., materials in the image) at many narrow wavelength bands. Often, this is represented visually as a cube, where each slice of the cube represents the image at a different wavelength. Spectral information is important in many fields such as environmental remote sensing, monitoring chemical/oil spills, and military target discrimination. For comprehensive discussions, please see, for example, [1–3]. Hyperspectral image data is often represented as a matrix $A \in \mathbb{R}^{m\times n}$, where each entry A_{ij} is the reflection of ith pixel at the jth wavelength. Thus, a column of A contains the entire image at a given wavelength; each row contains the reflection of one pixel at all given wavelengths—often referred to as the *spectral signature* of a pixel.

HSI data can be collected over hundreds of wavelengths —creating truly massive data sets. The transmission, storing, and processing of these large data sets often present significant difficulties in practical situations [1]. Dimensionality reduction methods provide means to deal with the computational difficulties of the hyperspectral data. These methods often use projections to compress a high-dimensional data space represented by a matrix A into a lower-dimensional space represented by a matrix B, which is then factorized. Such factorizations are referred to as *low-rank* matrix factorizations, resulting in a low-rank matrix approximation to the original HSI data matrix A. See, for example, [2, 4–6].

Dimensionality reduction techniques are generally regarded as lossy compression; that is, the original data is not exactly represented or reconstructed by the lower-dimensional space. For lossless compression of HSI data, there have been efforts to exploit the correlation structure within HSI data plus coding the residuals after stripping off the correlated parts; see, for example, [7, 8]. However, given the large number of pixels, these correlations are often restricted to the spatially or spectrally local areas, while the dimension reduction techniques essentially explore the global correlation structure. By coding the residuals after subtracting the original matrix by its low-dimensional representation, one can compress the original data in a lossless manner, as in [8]. The success of lossless compression requires low entropy of the data distribution, and, as we shall see in the experiments section, generally the entropy of residuals for our method will be much lower than the entropy of the original data.

Low-rank matrix factorizations can be computed using two general types of algorithms: deterministic and probabilistic. The most popular methods for deterministic low-rank factorizations include the singular value decomposition

(SVD) [9] and principal component analysis (PCA) [10]. Advantages of these methods include the following: first, often a small number of singular vectors (or principal components) sufficiently capture the action of a matrix; second, the singular vectors are orthonormal; third, the truncated SVD (TSVD) is the optimal low-rank representation of the original matrix in terms of Frobenius norm by the Eckart-Young theorem [9]. This last advantage is especially suited for compression with the TSVD method, since the Frobenius norm of the residual matrix is the smallest among all rank-k representations of the original matrix, and hence we should expect a much lower entropy in its distributions—making it suitable for compressive coding schemes. Both decompositions offer truncated versions so that these decompositions can be used to represent an n-band hyperspectral image with the data-size-equivalent of only k images, where $k \ll n$. For applications of the SVD and PCA in hyperspectral imaging see, for example, [11, 12].

The traditional deterministic way of computing the SVD of a matrix $A \in \mathbb{R}^{m\times n}$ is typically a two-step procedure. In the first step, the matrix is reduced to a bidiagonal matrix using householder reflections or sometimes combined with a QR decomposition if $m \gg n$. This takes $O(mn^2)$ floating-point operations (flops), assuming that $m \geq n$. The second step is to compute the SVD of the bidiagonal matrix by an iterative method in $O(n)$ iterations, each costing $O(n)$ flops. Thus, the overall cost is still $O(mn^2)$ flops [13, Lecture 31]. In HSI applications, the datasets can easily break into the million-pixel or even giga-pixel level, which renders this operation impossible on typical desktop computers.

One solution is to apply probabilistic methods which give closely approximated singular vectors and singular values, while the complexity is at a much lower level. These methods begin by randomly projecting the original matrix to obtain a lower-dimensional matrix, while the range of the original matrix is asymptotically kept intact. The much-smaller projected matrix is then factorized using a full-matrix decomposition such as SVD or PCA, after which the resulting singular vectors are backprojected to the original space. Compared to deterministic methods, probabilistic methods often offer the lower cost and more robustness in computation, while achieving high-accuracy results. See the seminal paper [14], and the references therein.

Knowing the redundancy of HSI data, especially in the spectral dimension, recently we have observed studies on the compressive HSI sensing, either algorithmic [11, 15] or experimental [6, 16, 17], and all of them involve a random projection of the data onto a lower-dimensional space. For example, in [11] Fowler proposed an approach that exploits the use of compressive projections in sensors that integrate dimensionality reduction and signal acquisition to effectively shift the computational burden of PCA from the encoder platform to the decoder site. This technique, termed compressive-projection PCA (CPPCA), couples random projections at the encoder with a Rayleigh-Ritz process for approximating eigenvectors at the decoder. In its use of random projections, this technique can be considered to possess a certain duality with our approach to randomized SVD methods in HSI. However, CPPCA recovers coefficients of a known sparsity pattern in an unknown basis. Accordingly, CPPCA requires the additional step of eigenvector recovery.

In this paper, we present a *randomized singular value decomposition* (rSVD) method for the purposes of lossless compression, reconstruction, classification, and target detection. On a large HSI dataset we apply the rSVD method to demonstrate its efficiency and effectiveness of the proposed method. On another HSI dataset, we will show the effectiveness of the proposed algorithm in detecting targets, especially small targets, through singular vectors. In terms of reconstruction quality, we will compare our algorithm with CPPCA [11] by using the signal-to-noise ratio (SNR).

We note that Chen et al. [18] have recently provided an extensive study on the effects of linear projections on the performance of target detection and classification of hyperspectral imagery. In their tests they found that the dimensionality of hyperspectral data can typically be reduced to 1/5 ∼ 1/3 that of the original data without severely affecting performance of commonly used target detection and classification algorithms.

The structure of the remainder of the paper is as follows. In Section 2, we give a detailed overview of rSVD in Section 2.1, the connections between this work and CPPCA in Section 2.2, and the compression and reconstruction of HSI data in Section 2.3. In Section 3, we present numerical results of the rSVD method on two publicly available real data sets. Finally, we draw some conclusions and identify some topics for future work in Section 4.

2. Review of Randomized Singular Value Decomposition

We start by defining terms and notations. The singular value decomposition (SVD) of a matrix $A \in \mathbb{R}^{m\times n}$ is defined as

$$A = U\Sigma V^T, \tag{1}$$

where U and V are orthonormal, and Σ is a rectangular diagonal matrix whose entries on the diagonal are the singular values denoted as σ_i. The column vectors of U and V are left and right singular vectors, respectively, denoted as u_i and v_i. Define the truncated SVD (TSVD) approximation of A as a matrix A_k such that

$$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T. \tag{2}$$

We define the randomized SVD (rSVD) of A as follows:

$$\widehat{A}_k = \widehat{U}\widehat{\Sigma}\widehat{V}^T, \tag{3}$$

where $\widehat{U}$ and $\widehat{V}$ are both orthonormal, and $\widehat{\Sigma}$ is diagonal with diagonal entries denoted as $\widehat{\sigma}_i$. Denote the column vectors of $\widehat{U}$ and $\widehat{V}$ as $\widehat{u}_i$ and $\widehat{v}_i$, respectively, and call them randomized singular vectors. Here, u_i, v_i, and σ_i are related to $\widehat{u}_i$, $\widehat{v}_i$, and $\widehat{\sigma}_i$, respectively. Define the residual matrix of a TSVD approximation as follows:

$$R_k = A - A_k, \tag{4}$$

```
Input: An $m \times n$ matrix $A$ and $a$ precision measure $\epsilon$.
Output: An $m \times k$ matrix $Q$ and rank $k$.
Initialize $Q$ as an empty matrix, $e = 1$ and $k = 0$.
while $e > \epsilon$ do
    (1) $k = k + 1$.
    (2) $y_i = A\omega_i$, where $\omega_i$ is a Gaussian random vector.
    (3) $q_i = (I - QQ^T)y_i$.
    (4) $q_i = q_i/\|q_i\|$.
    (5) $Q \leftarrow [Q\ q_i]$.
    (6) $\Omega \leftarrow [\Omega\ \omega_i]$.
    (7) Compute error $e = \|A - QQ^T A\|_F/\|A\|_F$.
end
```

Algorithm 1: Construct an orthonormal matrix Q that approximates the range of matrix $A\Omega$.

and the residual matrix of a rSVD approximation as follows:

$$\widehat{R}_k = A - \widehat{A}_k. \tag{5}$$

Define the random projection of a matrix as follows:

$$Y = \Omega^T A, \quad \text{or} \quad Y = A\Omega, \tag{6}$$

where Ω is a random matrix with independent and identically distributed (i.i.d.) entries.

2.1. Randomized SVD Algorithm. The rSVD algorithm as considered by [14] explores approximate matrix factorizations using random projections, separating the process into two stages. In the first stage, random sampling is used to obtain a reduced matrix whose range approximates the range of A; in the second stage, the reduced matrix is factorized. In this paper, we use this framework for computing the rSVD of a matrix A.

The first stage of the method is common to many approximate matrix factorization methods. For a given $\epsilon > 0$, we wish to find a matrix Q with orthonormal columns such that

$$\|A - QQ^T A\|_F^2 \leq \epsilon. \tag{7}$$

Algorithm 1 [14] can be used to compute Q.

Because in practice we may not know the target rank k of A, Algorithm 1 allows us to look for an appropriate target rank based upon given ϵ such that (7) holds. However if the target rank is known, one can avoid computing the error term e at each iteration by replacing the *While* loop with a *For* loop. In practice, the number of columns of Q is usually chosen to be slightly larger than the numerical rank of A [14]. Without loss of generality, we assume that $Q \in \mathbb{R}^{m\times l}$, where $l \ll n$. The columns of Q form an orthogonal basis for the range of $A\Omega$, where Ω is a matrix composed of the random vectors $\{w_i\}$, typically with a standard normal distribution [14]. The range of the product $A\Omega$ is an approximation to the range of A.

The second stage of the rSVD method is to compute the SVD of the reduced matrix $Q^T A \in \mathbb{R}^{l\times m}$. Since $l \ll n$, it is generally computationally feasible to compute the SVD of the reduced matrix. Letting $\widetilde{U}\widehat{\Sigma}\widehat{V}^T$ denote the SVD of $Q^T A$, we obtain that

$$A \approx \left(Q\widetilde{U}\right)\widehat{\Sigma}\widehat{V}^T = \widehat{U}\widehat{\Sigma}\widehat{V}^T, \tag{8}$$

where $\widehat{U} = Q\widetilde{U}$ and $\widehat{V}$ are orthogonal matrices, and thus by (8), $\widehat{U}\widehat{\Sigma}\widehat{V}^T$ is an approximate SVD of A, and the range of $\widehat{U}$ is an approximation to range of A. Algorithm 2 summarizes the discussion above. See [14] for details on the choice of l, along with extensive numerical experiments using randomized SVD methods and a detailed error analysis of the two-stage method described above.

Next we discuss several variations of Algorithm 2 depending on the properties of A. We will test all cases in the numerical results section.

Case 1. If knowing the target rank k, and if the singular values of A decay rapidly, we can skip Algorithm 1 by simply using the rank revealing QR factorization, $Y = QR$, where Q is an orthogonal basis of the range of Y. Figure 1 from [19] compares the approximation error e_k and the theoretical error σ_{k+1} of a matrix A, and clearly when the singular values of A decay rapidly, e_k is close to the theoretical error σ_{k+1} with high probability.

Case 2. If the singular values of A decay gradually, or σ_k/σ_1 is not small, we may lose the accuracy of estimates. Consider introducing a power q and forming Y as $Y = (AA^T)^q AR$. Since $(AA^T)^q\Omega$ has the same singular vectors as A, while its singular values, $\{\sigma_i^{2q-1}, i = 1, \ldots, n\}$, decay more rapidly. Hence the error will be smaller by Theorems 2.3 and 2.5 in [14]. From Figure 2, we see that the e_k is not always close to σ_{k+1}, especially when $q = 0$, but, by increasing the power q, we observe the reduction of errors.

Case 3. Algorithm 2 requires us to revisit the input matrix, while this may be not feasible for large matrices. For example, in ultraspectral imaging [20], one could have thousands of spectral bands, and PCA on such datasets would require computing the eigenvectors and eigenvalues of a covariance matrix with a huge dimension. Another example is in the atmospheric correction model called MODTRAN5 [21], that utilizes large lookup tables (LUTs), and the compression

```
Input: An $m \times n$ matrix $A$ and rank $k$ with $k \leq n \leq m$.
Output: The rSVD of $A$: $\widehat{U}, \widehat{\Sigma}, \widehat{V}$.
(1) Generate a Gaussian random matrix $\Omega \in \mathbb{R}^{n\times k}$.
(2) Form the projection of $A$: $Y = A\Omega$.
(3) Construct $Q \in \mathbb{R}^{m\times k}$ by Algorithm 1.
(4) Set $B = Q^T A \in \mathbb{R}^{k\times n}$.
(5) Compute the SVD of $B$, $B = \widetilde{U}\widehat{\Sigma}\widehat{V}^T$.
(6) $\widehat{U} = Q\widetilde{U}$.
```

Algorithm 2: The basic rSVD algorithm.

```
Input: An $m \times n$ matrix $A$ and an integer $J$.
Output: $\{B_j, Q_j, \widehat{r}_j, j = 1, \ldots, J\}$.
$j = 1$
while flight continues do
    (1) Acquire the HSI data, $A_j$, scanned in the last few seconds.
    (2) Apply Algorithm 1 for $Q_j$ and $B_j$.
    (3) Compute the residual, $r_j = A_j - Q_j B_j$.
    (4) Code $r_j$ as $\widehat{r}_j$ with a parallel floating point coding algorithm.
    (5) Store $Q_j$, $B_j$ and compressed $r_j$.
    (6) $j = j + 1$.
end
```

Algorithm 3: rSVD encoder.

of LUTs by the PCA technique would again require the eigen decomposition of large covariance matrices. Here we introduce a variation of Algorithm 2 that only requires one pass over a large symmetric matrix. Now we define matrix B as follows:

$$B = Q^T A Q, \tag{9}$$

and we multiply by $Q^T\Omega$, that is,

$$BQ^T\Omega = Q^T A Q Q^T \Omega. \tag{10}$$

Since $A \approx AQQ^T$, we have the following approximation:

$$BQ^T\Omega \approx Q^T A\Omega = Q^T Y, \tag{11}$$

and hence by a least-square solution we have

$$B \approx Q^T Y \left(Q^T\Omega\right)^{\dagger}, \tag{12}$$

where the superscript † represents the pseudoinverse. Notice the absence of A in the approximate formula of B. Thus, for a large symmetric A, we will use (12) rather than $Q^T A$ to compute B, while the rest of Algorithm 2 would remain the same.

2.2. Connections to CPPCA. A significant difference between the compressive-projection PCA (CPPCA) approach and our work is that CPPCA uses a random orthonormal matrix P to compress the data matrix A. In comparison, though we also use random projections, the orthonormal matrix Q is constructed from, and directly related to, the data matrix A. In particular, we compute an orthogonal Q such that $\|A - QQ^T A\|_F \leq \epsilon$. Also, because the projection onto convex sets (POCSs) algorithm is used for reconstruction, the projection matrix P of CPPCA has to be different for different blocks of the scene, which have to be independently drawn and orthogonalized; meanwhile, one random projection matrix Ω in rSVD is sufficient and can be applied to the whole dataset. Another restriction of CPPCA lies in the fact that the Rayleigh-Ritz method requires well-separated eigenvalues [22], which might be true for the first few largest eigenvalues, but usually not true for the smaller eigenvalues. In a later section we present our approach for matrices with slowly decaying singular values in Case 2.

2.3. Compression and Reconstruction of HSI Data by rSVD. The flight times of airplanes carrying hyperspectral scanning imagers are usually limited by the data capacity, since within 5 to 10 seconds hundreds of thousands of pixels of hyperspectral data are collected [1]. Hence for real-time onboard processing, it would be desirable to design algorithms capable for processing this amount of data within 5 to 10 seconds before the next section of the scene is scanned. Here we use the proposed rSVD algorithm to losslessly compress blocks of HSI data, each within a frame of 10 second flight time, which is equivalent to dividing the HSI data cube along the flight direction, either the x or y direction, with the number of rows (y direction) or columns (x direction) determined by the ground sample distance (GSD) and the flight speed. Algorithm 3 describes the rSVD encoder, which outputs $\{B_j, Q_j, \widehat{r}_j, j = 1, \ldots, J\}$ to be stored onboard, where B_j and Q_j are the outputs of Algorithm 2 for

```
Input: $\{B_j, Q_j, \hat{r}_j, j = 1, \ldots, J\}$.
Output: Reconstructed matrix $\hat{A}$.
For $j = 1 : J$ do
    (1) Decode $r_j$ from $\hat{r}_j$ with a parallel floating point decoding algorithm.
    (2) $A_j = Q_j B_j + r_j$.
    (3) $j = j + 1$.
end
Group all $\{A_j, j = 1, \ldots, J\}$ together in $A$.
```

Algorithm 4: rSVD decoder.

the jth block of data, while $\hat{r}_j$ is the coded residual. These are then used by Algorithm 4 to reconstruct the original data losslessly, and we can see it only involves a one-pass matrix-matrix multiplication and is without iterative algorithms. Compared to CPPCA, the number of bytes used for storing the Bs and Qs is smaller, and the reconstruction only involves matrix-matrix multiplication. The only possible bottleneck might be the residual coding, but the recent development in floating point coding has seen throughputs reaching as much as 75 *Gb/s* [23] on a graphic processing unit (GPU), while even on an eight-Xeon-core computer we have seen throughput at 20 *Gb/s*, and both would be sufficiently fast to code the required amount of HSI data within 10 seconds.

3. Numerical Experiments

3.1. Accuracy of the rSVD Estimates. In this section, we will first compare the results from rSVD and from the exact TSVD by the MATLAB function, "*svds,*" which computes the largest k singular values and the associated singular vectors of a large matrix. It is considered to be an efficient and accurate method to obtain the TSVD. To simulate large HSI datasets, we generate random test matrices $A \in \mathbb{R}^{m \times n}$, with n fixed at 100 representing 100 spectral channels, while $m = 100,000; 200,000; \ldots, 2,000,000$, representing the number of pixels. The singular values of A are simulated as following a power decay with the power set as -1, that is, $\sigma_k/\sigma_1 = 1/k$. We will use Algorithm 2 to compute the rSVD. The comparison of computation time is shown in Figure 3(a), from which we find that "*svds*" is almost as effective as rSVD when n is relatively small. However, when m increases, the computation times of "*svds*" increase at a much faster pace than that of rSVD, and note that when $m = 300,000$, the processing time of rSVD is well within 10 seconds, meeting the onboard processing time limit. To judge the accuracy of estimated singular vectors, we compute the correlation or the inner product of singular vectors in U by "*svds*" and $\hat{U}$ by rSVD as shown in Figure 3(b), where we clearly see that both sets are almost identical up to the fifteenth singular vector. To judge the accuracy of estimated singular values, we compute the relative absolute errors, $|\hat{\sigma}_k - \sigma_k|/|\sigma_k|$, and plot them in Figure 3(c). Again we observe the high accuracy of the estimates up to the first 15 singular values. In most HSI datasets, the singular values decay rate is generally faster than $1/k$, and hence we should expect even higher accuracy of the estimates. Also, it is sufficient to estimate the first 15 or even 10 singular vectors and singular values, which would often cover more than 90% of the original variance of the HSI data [1].

Then we numerically test the three special cases discussed in Section 2.1.

Case 1. We have considered square Toeplitz matrices with increasingly large sizes, $n = 15, 30, \ldots, 1500$. Figure 4 shows the rapid decay of a $1,000 \times 1,000$ matrix, and hence they are suitable for testing the algorithm. Figure 5(a) shows that the relative Frobenius norm errors rise and fall in the order of 10^{-12} and remain in the same order even when the size of a matrix increases. Figure 5(b) demonstrates that the computational time of rSVD is very short for the Toeplitz matrices whose singular values decay rapidly.

Case 2. Here we simulate $10,000 \times 100$ matrices with slowly decaying singular values, that is, $\sigma_i/\sigma_1 = i^{-s}$, with $s = 0.2, 0.4, \ldots, 1.0$. For each matrix, we run the rSVD algorithm 100 times for each power q in the set, $\{q = 1, 2, \ldots, 20\}$. The norm errors of reconstructed matrices are averaged across 100 runs and normalized by the norm error when $q = 1$, that is, e_q/e_1. Figure 6 shows that increasing the power q improves the reconstruction quality or decreases the norm error, and greater effects are observed for larger s because the singular values of $(AA^T)^q$ are $\sigma_i^{2q} = \sigma_1^{2q} i^{-2qs}$.

Case 3. To simulate covariance matrices, we generate a sequence of positive definite symmetric matrices with increasing size as $n = 100, 110, \ldots, 2,000$. The eigenspectrum follows a power decay with the power set as -1. We apply the modified B as in Case 3 to compute its SVD, rather than using $Q^T A$. We set $k = 25$. Figure 7(a) shows the computation time compared with using regular SVD in MATLAB, while Figure 7(b) shows the relative Frobenius norm errors between the original matrix and its low-rank approximation. Apparently the computation time used by rSVD is far less than the regular SVD, while the accuracies are quite high.

3.2. rSVD on a Large HSI Dataset and a Lossless Compression. The rSVD algorithm was also applied to a relatively large HSI dataset consisting of a $920 \times 4,933 \times 58$ image cube collected over Gulfport MS by a commercial hyperspectral sensor having a spectral range of 0.45 to 0.72 microns. This cube was then unfolded into a large matrix of size $4,538,360 \times 58$. Running an exact SVD algorithm is almost impossible on a regular desktop computer with limited memory and speed,

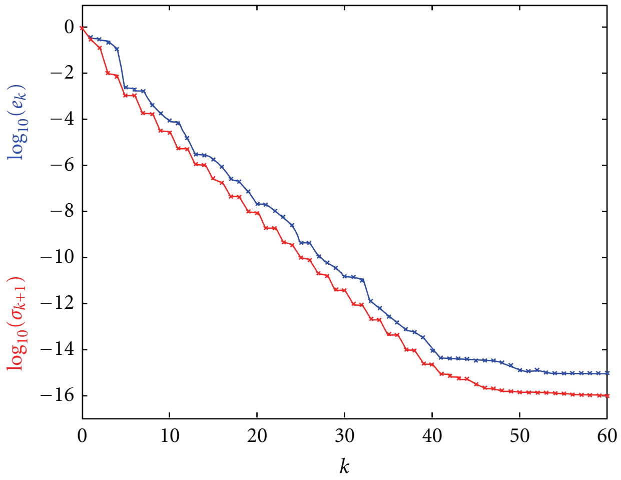


Figure 1: Comparison of the computed error e_k (blue) and the theoretical error σ_{k+1} (red).

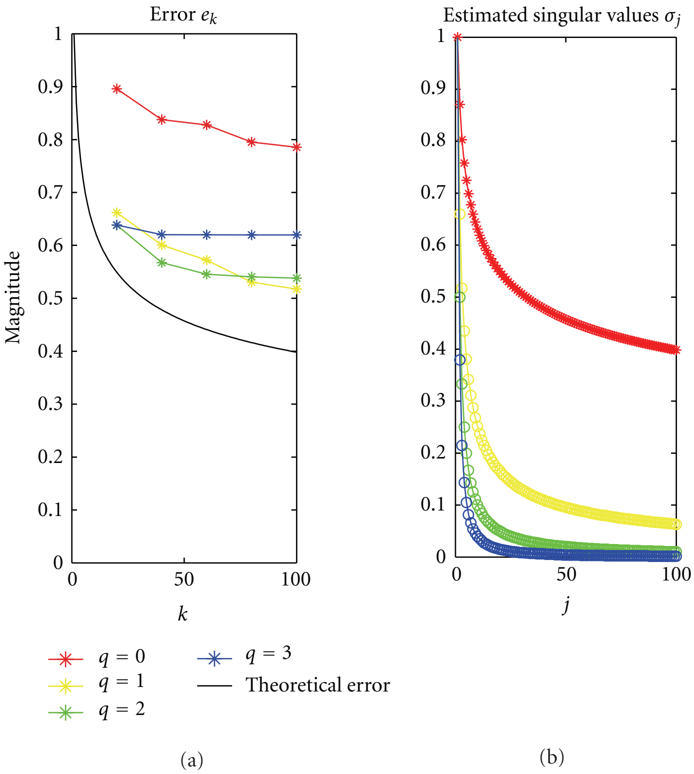
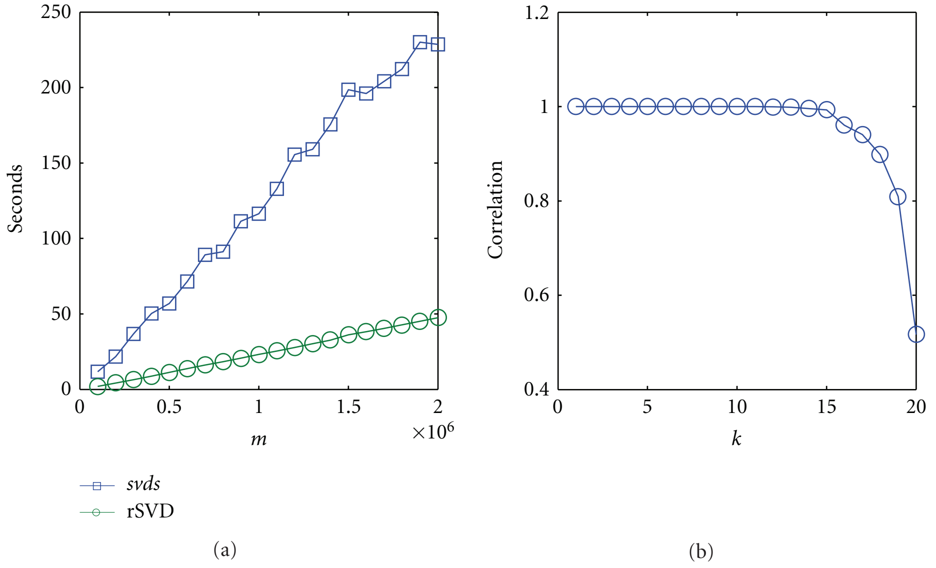


Figure 2: Comparison of the error e_k and the theoretical error σ_{k+1}. The red curve shows that the error e_k is greater than the theoretical error σ_{k+1}. Note that the singular values decay more rapidly as q increases.

while the rSVD algorithm gives us singular values and vectors very close to the true ones, with a significantly reduced amount of flops and memory required. For the first 25 singular vectors and singular values, it only takes 68 seconds on a desktop computer with Xeon 3.2 GHz Quadcores and 12 GB memory. From the computed singular values and vectors, we observed that the singular vectors after the ninth singular vector all appear to be noise, indicating that the data matrix does have a low-rank representation.

Figure 8 shows the results for a small scene from the large dataset described above, consisting of part of the University of Southern Mississippi Campus, extracted from the large Gulfport MS dataset. Notice the targets placed in the scene, for detection and identification tests. The first eight singular vectors, $\hat{u}_i$, are folded back from the transformed data. The first singular vector shown in Figure 8 is the mean image across 58 spectral bands, while the second singular vector shows high intensity at the grass and foliage pixels, the third



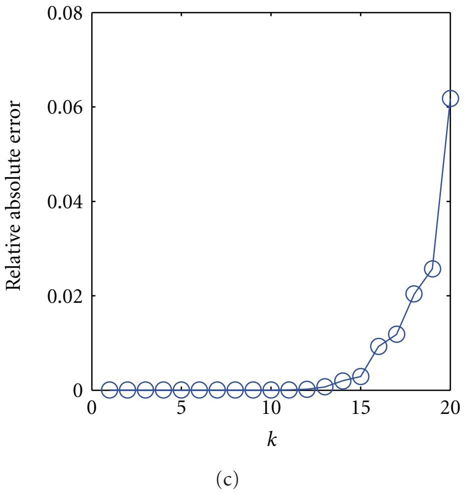


FIGURE 3: (a) Computation time of "*svds*" and rSVD. (b) Correlations between the singular vectors in U by "*svds*" and rSVD. (c) Relative absolute differences between the singular values estimated by "*svds*" and rSVD.

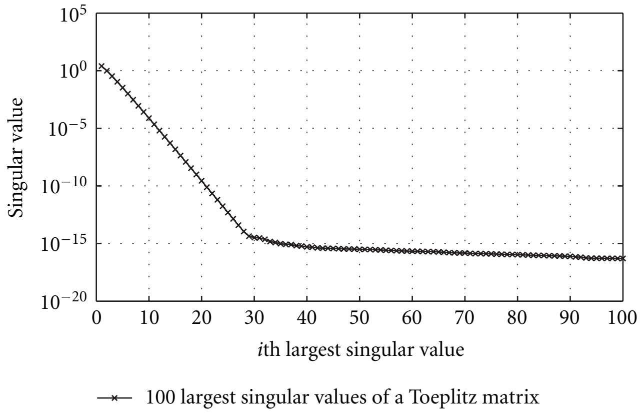
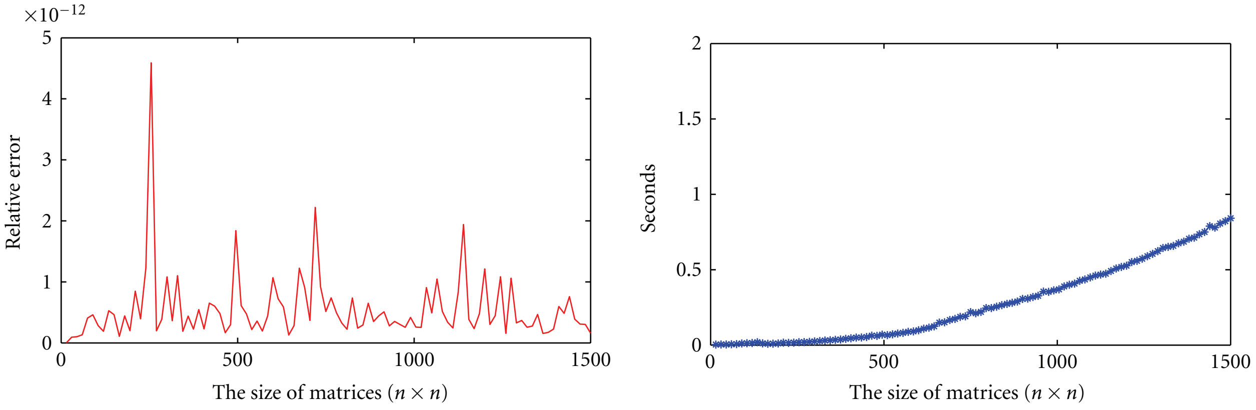


FIGURE 4: Illustration rapidly decaying singular values of matrix of size 1000 × 1000.



(a)

(b)

Figure 5: (a) The relative errors between the rSVD low-rank approximation and the original matrix A are shown by the red curve. (b) The computational time for rSVD.

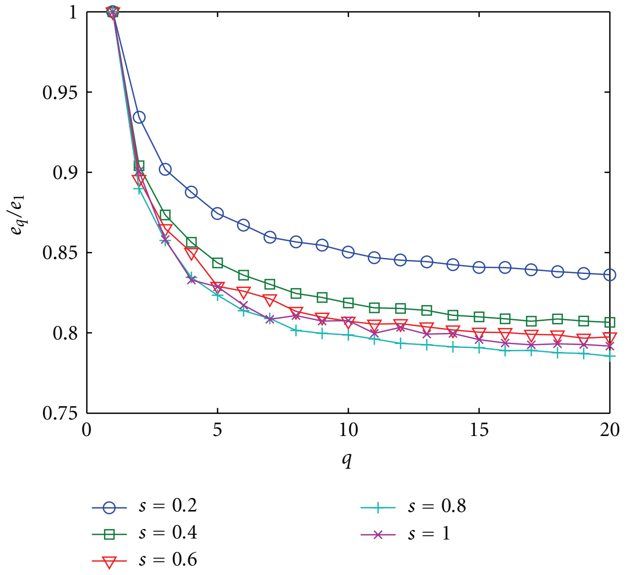
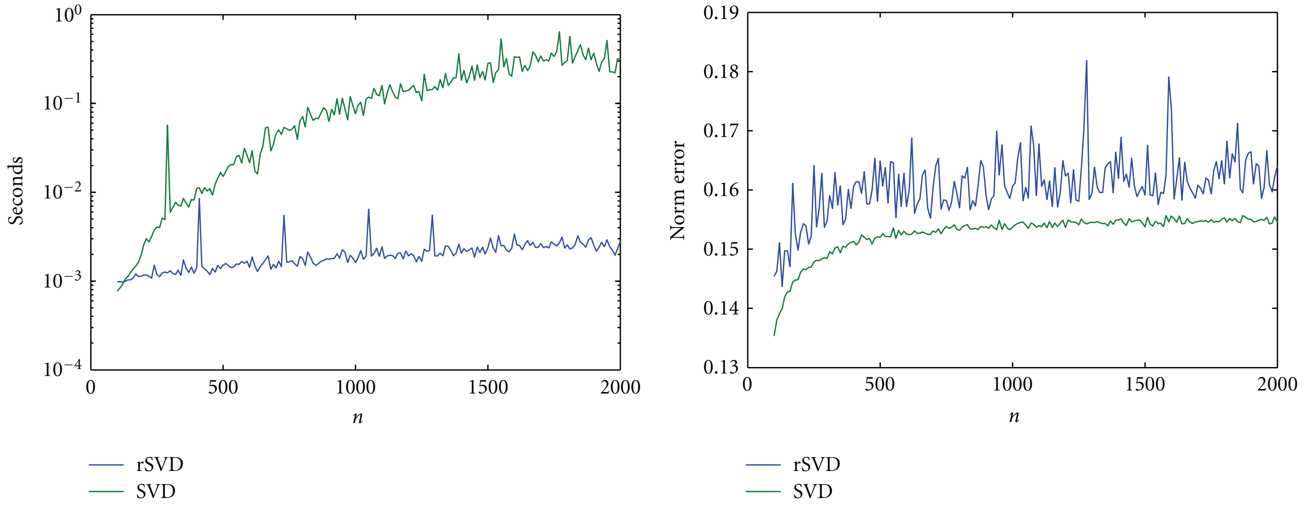


Figure 6: The reduction of norm error by increasing the power q for singular value spectrums decaying with various powers s.

shows the targets quite clearly, as well as high-reflectance sandy areas or rooftops, the fifth shows the low-reflectance pavement or roof tops, and shadows, and the seventh shows vehicles at various places marked by the circles. Starting from the eighth, the rest of the singular vectors appear to be mostly noise.

In Figure 9(b), the histogram of entries in $\widehat{R}_k$ shows that residuals are roughly distributed as a Laplacian distribution, and all residuals are within the range of $[-.1, .1]$, which is significantly smaller than the original range of A in Figure 9(a). Moreover, most of the residuals (93%) are within the range $[-.01, .01]$ (notice the log scale on the y axis), which means that the entropy of residuals is significantly smaller than the entropy of the original. This justifies a further coding step on the residuals so as to complete a lossless compression scheme. Here we apply the Hoffman coding due to its fast computation and show the compression ratios at various error rates, corresponding to the numbers of bits required to code the residuals. For example, a 16-bit coding would result in an error in the range of 10^{-5}. Figure 10 provides us options on balancing compression with accuracy. For practical purposes, an error rate in the order of 0.001 might be sufficient, and this would result in a compression ratio of 2.5 to 4. For comparison purpose, the 3D-SPECK [7] on a small dataset of size 320 × 360 × 58 results in a compression ratio of 1.12 at the 16-bit coding. If more sophisticated coding algorithms than Hoffman coding are applied here, we could see more improvements on the compression ratios. For computing the compression ratios, we have assumed 16-bit coding (2-byte) for all the matrices, including B_j, Q_j, the residual matrix and the coded (compressed) residual matrix.



(a)

(b)

Figure 7: (a) Computation time of rSVD in blue and SVD in green. (b) Relative norm errors by rSVD and SVD.

To test the suitability of the onboard real-time processing by Algorithm 3, we apply the rSVD on a $300,000 \times 100$ matrix and see it is finished in 7 seconds on a low-end dual-core laptop computer, and if, with a parallel coding algorithm for the residuals, we should finish Algorithm 3 within the required 10-seconds time frame.

3.3. Small Target Detection Using rSVD. For a small target detection experiment using rSVD, we choose a version of the Forest Radiance HSI dataset, which has been analyzed by using numerous target detection methods, see, for example, [18, 24–26] Our rationale behind using an SVD algorithm in target detection lies in the fact that even though targets might be of small size, if all the spectrally similar targets have sufficient presence, some singular vectors of the HSI data matrix will reflect these features, and hence the presence of targets can be simply shown by these singular vectors. After removing the water-absorption and other noisy bands, we unfold the $200 \times 150 \times 169$ data cube into a $30,000 \times 169$ matrix and apply Algorithm 2 for the singular vectors $\hat{u}_i$. Figure 11 shows the sum of first twelve $\hat{u}_i$ folded back into a 200×150 matrix, and we can clearly detect 25 of the 27 targets, while the other two are slightly visible.

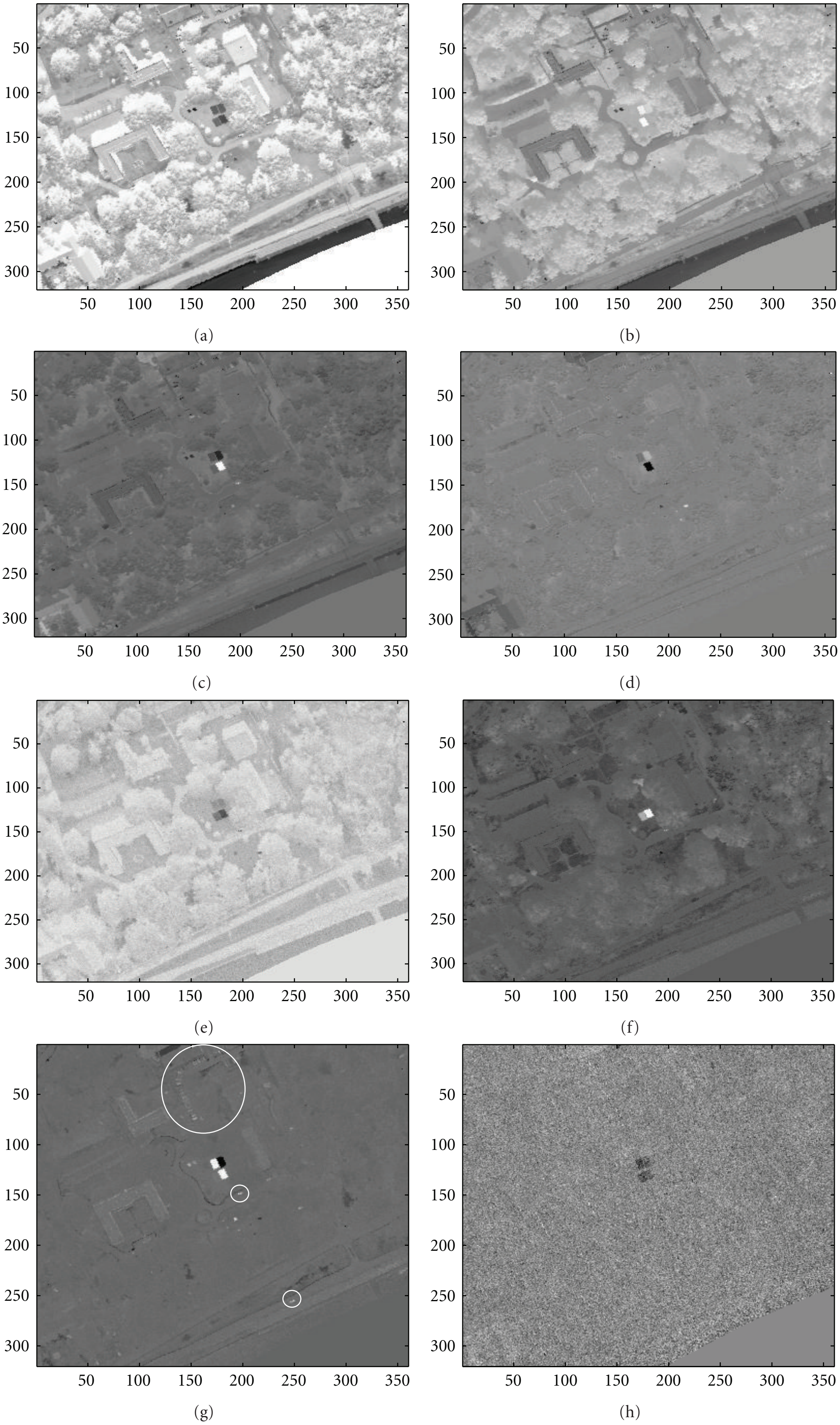
3.4. Comparison between rSVD and CPPCA. In this section, we will compare rSVD with CPPCA from the aspects of accuracy and computation time, first on simulated data and then on a real HSI dataset. We first simulate a set of matrices with increasing number of rows (pixels), $m = 10,000, 20,000, \ldots, 100,000$, while fixing $n = 100$ as 100 spectral bands. The singular value spectrum is simulated as following a power decay rate with the power set as -1. Both CPPCA and rSVD algorithms are applied to each simulated matrix, and results are compared in terms of their reconstruction quality and the computation time. Figure 12(a) shows that the running time of rSVD increases linearly with m, while that of CPPCA remains constant, which is not surprising since CPPCA mainly works on eigenvectors of fixed dimension n. However in terms of reconstruction quality, Figure 12(b) shows the advantage of rSVD. Here we set the number of reconstructed eigenvectors by CPPCA to 3 since it provides the best norm errors, while for rSVD we set it to 25.

Table 1: Computation times (seconds) of rSVD and CPPCA.

k/n	0.1	0.15	0.2	0.3	0.4	0.5
rSVD	0.212	0.292	0.390	0.707	0.897	1.264
CPPCA	0.247	0.305	0.331	0.368	0.399	0.509

For the real dataset, we use a small section of the Gulfport dataset and fold it into a matrix A with size 115200×58. From the reconstructed matrices $\hat{A}$ by both methods, with varying rank k we compare their reconstruction qualities in terms of signal-to-noise ratio (SNR) in Figure 13, and the computation time in Table 1. Again we observe the better reconstruction quality though slightly slower computation of rSVD when compared to CPPCA.

Next, we compare the accuracy of the reconstruction of the eigenvectors, $\hat{v}_i$, of the covariance matrix of A by these two methods. Given that PCA is an extremely useful tool in HSI data analysis, for example, for classification and target detection, it is essential to obtain a quality reconstruction of the eigenvectors $\hat{v}_i$ by rSVD in terms of accuracy and efficiency. Here we simulate a $10,000 \times 100$ matrix A with orthogonalized random matrices, U_o and V_o, and a power-decay singular value spectrum in Σ_o, with the power set as -1. Then we run both algorithms for 1,000 times, and, within each time, we compute the angles between eigenvectors by CPPCA and the true ones, and between eigenvectors by rSVD and the true ones. The first row of Figure 14 shows the histograms of angles between the first eight reconstructed eigenvectors by CPPCA and the true ones, while the second row shows the histograms



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

Figure 8: The first eight singular vectors, $\hat{u}_i$, are turned into images. The circles on the image of the seventh singular vectors indicate identified vehicles.

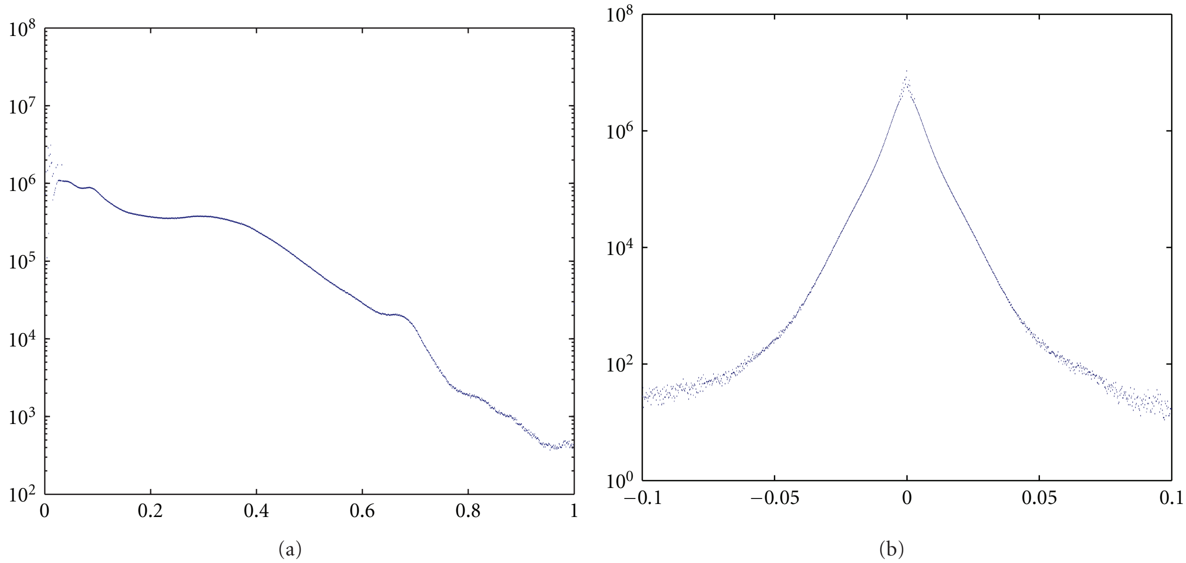


Figure 9: (a) The distribution of intensities of the original Gulfport HSI cube. (b) The distributions of residuals after subtracting the TSVD from the original.

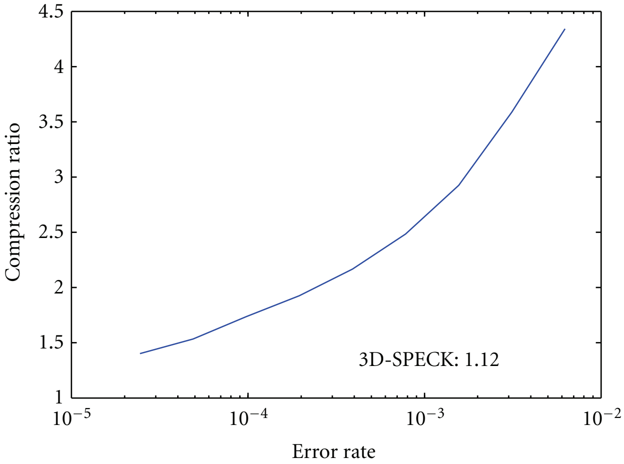


Figure 10: The compression ratio as a function of error rate.

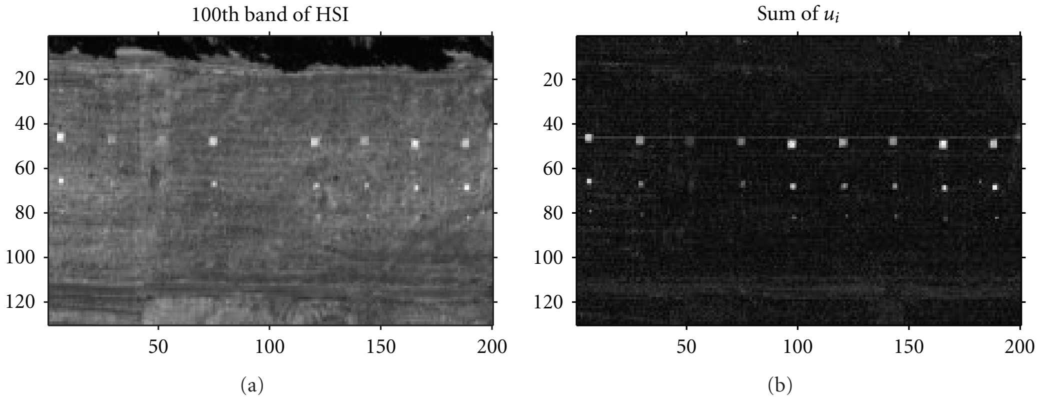
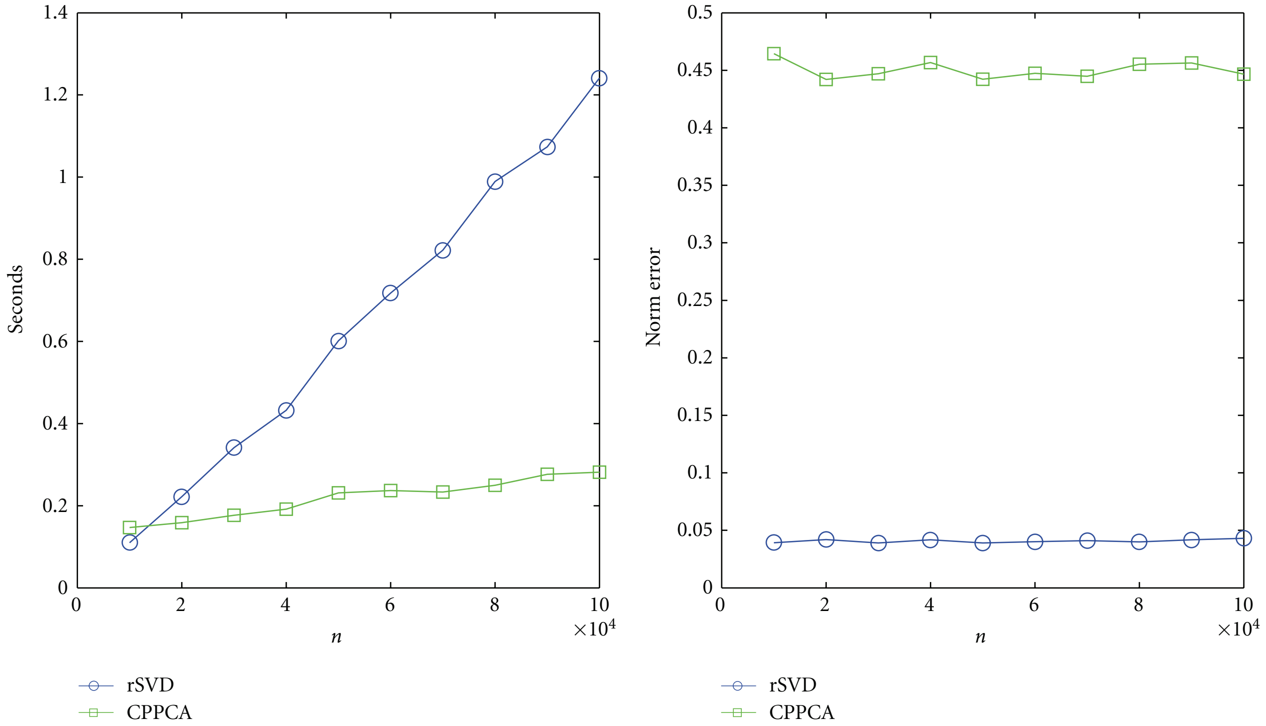


Figure 11: One band of the Forest Radiance HSI dataset is shown on the left. Binary target detections are shown on the right, obtained after a summation of the first 12 $\hat{u}_i$.



(a)

(b)

Figure 12: (a) Running times of rSVD and CPPCA. (b) Reconstruction qualities of rSVD and CPPCA.

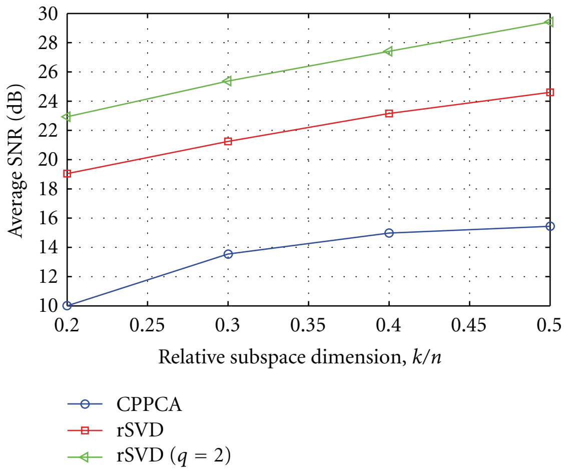


Figure 13: Comparison of reconstruction qualities of rSVD and CPPCA in terms of SNR.

of the angles between the first eight eigenvectors by rSVD and the true ones. We can see that the first three or four eigenvectors by CPPCA appear to be close to the true ones, while the rest are not. Hence if using more than four eigenvectors reconstructed by CPPCA, we observe a decrease in reconstruction quality or an increase in the norm error. However in the second row, we see good accuracy of the eigenvectors computed by rSVD.

3.5. Classification of HSI Data by rSVD.

Since the projection of a HSI data matrix by its truncated singular matrix, that is,

$$A_P = AV_k, \tag{13}$$

contains most of the information in the original matrix A, we can use any classification algorithm, such as the popular k means, to classify HSI data, but also use its representation

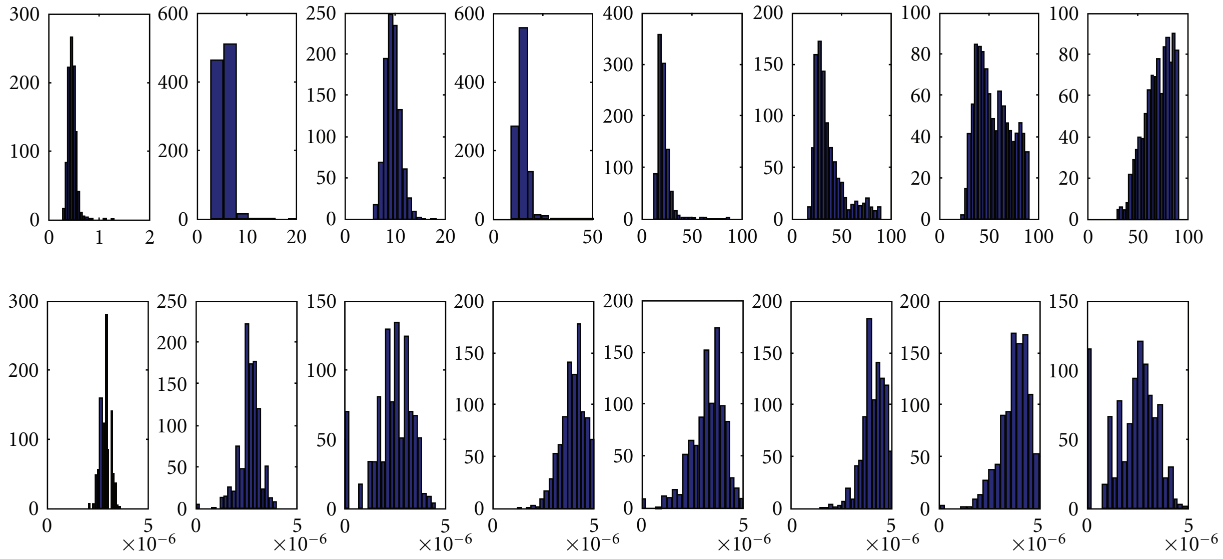


FIGURE 14: Comparison of reconstructed eigenvectors by rSVD and CPPCA with the true ones.

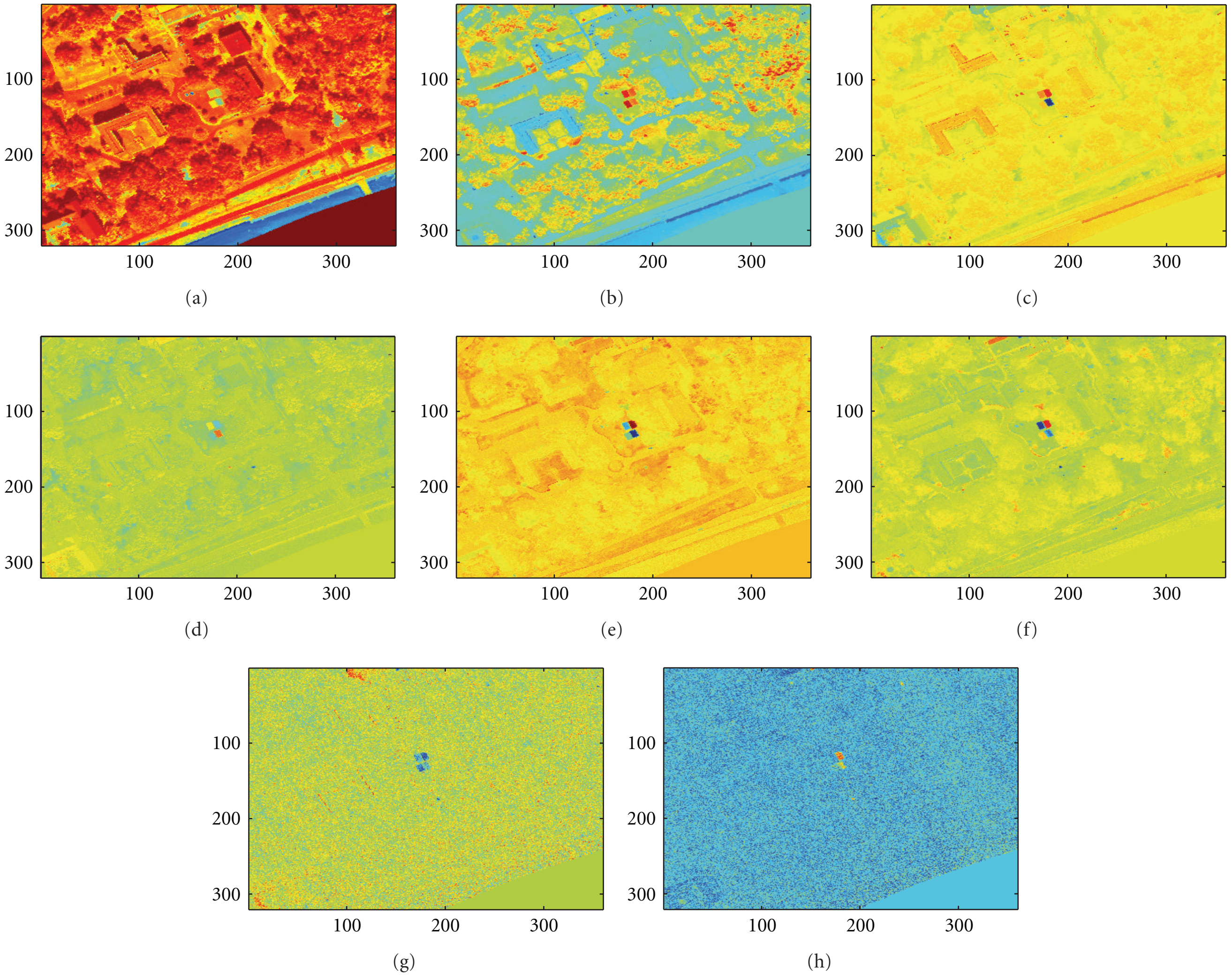


FIGURE 15: Plots of the first 8 columns of A_P.

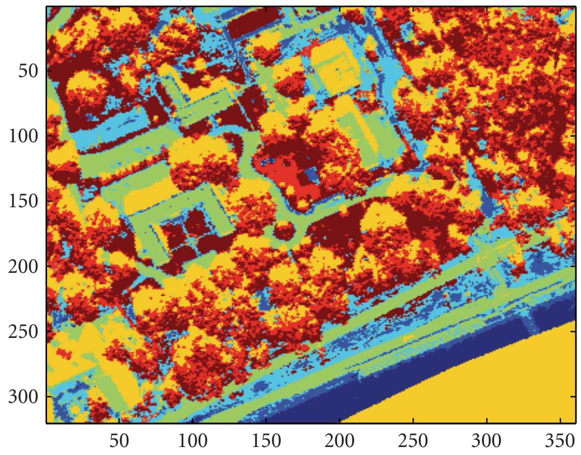


Figure 16: The classification result of k-means using A_k.

in a lower-dimensional space. Consider a small section of the Gulfport dataset. Figure 15 shows the first 8 columns of A_P. From the first subfigure, we see that most information of the hyperspectral image is contained in the first column, while the second column almost contains the rest of the information which the first column does not contain. The rest of the columns contain information at more detailed and spatially clustered levels. Figure 16 shows the result of classification by k-means, where we can see the low-reflectance water and shadows in yellow, the foliage in red, the grass in dark red, the pavement in green, high-reflectance beach sand in dark blue, and dirt/sandy grass in blue and light blue.

A comparison with results from running k-means on the full dataset shows that only 13 pixels of all the $320 \times 360 = 115,200$ pixels are classified differently between the full dataset and its low-dimensional representation. Hence it is highly suitable to use this low-dimensional representation for classification.

4. Conclusions

As HSI data sets are growing increasingly massive, compression and dimensionality reduction for analytical purposes has become more and more critical. The randomized SVD algorithms proposed in this paper enable us to compress, reconstruct, and classify massive HSI datasets in an efficient way while maintaining high accuracy in comparison to exact SVD methods. The rSVD algorithm is also found to be effective in detecting small targets down to single pixels. We have also demonstrated the fast computation in compression and reconstruction of the proposed algorithms on a large HSI dataset in an urban setting. Overall, the rSVD provides a lower approximation error than some other recent methods and is particularly well suited for compression, reconstruction, classification, and target detection.

Acknowledgments

The authors wish to thank the referees and the project sponsors for providing very helpful comments and suggestions for improving the paper. Research by Jiani Zhang and Jennifer Erway was supported by NSF grant DMS-08-11106. Research by Robert Plemmons and Qiang Zhang was supported by the U.S. Air Force Of- fice of Scientific Research (AFOSR), award number FA9550-11-1-0194, and by the U.S. National Geospatial-Intelligence Agency under Contract HM1582-10-C-0011, public release number PA Case 12-433.

References

[1] M. T. Eismann, *Hyperspectral Remote Sensing*, SPIE Press, 2012.

[2] H. F. Grahn and E. Paul Geladi, *Techniques and Applications of Hyperspectral Image Analysis*, Wiley, 2007.

[3] J. M. Bioucas-Dias, A. Plaza, N. Dobigeon et al., "Hyperspectral unmixing overview: geometrical, statistical, and sparse regression-based approaches," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 2, pp. 354–379, 2012.

[4] J. C. Harsanyi and C. I. Chang, "Hyperspectral image classification and dimensionality reduction: an orthogonal subspace projection approach," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 32, no. 4, pp. 779–785, 1994.

[5] A. Castrodad, Z. Xing, J. Greer, E. Bosch, L. Carin, and G. Sapiro, "Learning discriminative sparse models for source separation and mapping of hyperspectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, pp. 4263–4281, 2011.

[6] C. Li, T. Sun, K. F. Kelly, and Y. Zhang, "A compressive sensing and unmixing scheme for hyperspectral data processing," *IEEE Transactions on Image Processing*, vol. 21, no. 3, pp. 1200–1210, 2012.

[7] X. Tang and W. Pearlman, "Three-dimensional wavelet-based compression of hyperspectral images," *Hyperspectral Data Compression*, pp. 273–308, 2006.

[8] H. Wang, S. D. Babacan, and K. Sayood, "Lossless hyperspectral-image compression using context-based conditional average," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 12, pp. 4187–4193, 2007.

[9] G. H. Golub and C. F. V. Loan, *Matrix Computations*, The Johns Hopkins University Press, 3rd edition, 1996.

[10] I. Jolliffe, *Principal Component Analysis*, Springer, 2nd edition, 2002.

[11] J. E. Fowler, "Compressive-projection principal component analysis," *IEEE Transactions on Image Processing*, vol. 18, no. 10, pp. 2230–2242, 2009.

[12] P. Drineas and M. W. Mahoney, "A randomized algorithm for a tensor-based generalization of the singular value decomposition," *Linear Algebra and Its Applications*, vol. 420, no. 2-3, pp. 553–571, 2007.

[13] L. Trefethen and D. Bau, *Numerical Linear Algebra*, Society For Industrial Mathematics, 1997.

[14] N. Halko, P. G. Martinsson, and J. A. Tropp, "Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions," *SIAM Review*, vol. 53, no. 2, pp. 217–288, 2011.

[15] Q. Zhang, R. Plemmons, D. Kittle, D. Brady, and S. Prasad, "Joint segmentation and reconstruction of hyperspectral data with compressed measurements," *Applied Optics*, vol. 50, no. 22, pp. 4417–4435, 2011.

[16] M. E. Gehm, R. John, D. J. Brady, R. M. Willett, and T. J. Schulz, "Single-shot compressive spectral imaging with a dual-disperser architecture," *Optics Express*, vol. 15, no. 21, pp. 14013–14027, 2007.

[17] A. Wagadarikar, R. John, R. Willett, and D. Brady, "Single disperser design for coded aperture snapshot spectral imaging," *Applied Optics*, vol. 47, no. 10, pp. B44–B51, 2008.

[18] Y. Chen, N. M. Nasrabadi, and T. D. Tran, "Effects of linear projections on the performance of target detection and classification in hyperspectral imagery," *Journal of Applied Remote Sensing*, vol. 5, no. 1, Article ID 053563, 2011.

[19] G. Martinsson, "Randomized methods for computing the singular value decomposition of very large matrices," in *Workshop on Algorithms for Modern Massive Data Sets*, 2012.

[20] A. D. Meigs, L. J. Otten, and T. Y. Cherezova, "Ultraspectral imaging: a new contribution to global virtual presence," in *Proceedings of the IEEE Aerospace Conference*, vol. 2, pp. 5–12, March 1998.

[21] A. Berk, G. P. Anderson, P. K. Acharya et al., "MODTRAN 5, a reformulated atmospheric band model with auxiliary species and practical multiple scattering options: update," in *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XI*, vol. 5806 of *Proceedings of SPIE*, pp. 662–667, 2005.

[22] B. Parlett, *The Symmetric Eigenvalue Problem*, vol. 20, Society for Industrial Mathematics, 1998.

[23] M. A. O'Neil and M. Burtscher, "Floating-point data compression at 75 Gb/s on a GPU," in *Proceedings of the 4th Workshop on General Purpose Processing on Graphics Processing Units (GPGPU '11)*, p. 7, March 2011.

[24] R. C. Olsen, S. Bergman, and R. G. Resmini, "Target detection in a forest environment using spectral imagery," in *Imaging Spectrometry III*, vol. 3118 of *Proceedings of SPIE*, pp. 46–56, July 1997.

[25] C. I. Chang, "Target signature-constrained mixed pixel classification for hyperspectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 40, no. 5, pp. 1065–1081, 2002.

[26] B. Thai and G. Healey, "Invariant subpixel material detection in hyperspectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 40, no. 3, pp. 599–608, 2002.

Hindawi Publishing Corporation
Journal of Electrical and Computer Engineering
Volume 2012, Article ID 103286, 15 pages
doi:10.1155/2012/103286

Research Article

Evaluating Subpixel Target Detection Algorithms in Hyperspectral Imagery

Yuval Cohen,[1] Yitzhak August,[1] Dan G. Blumberg,[2] and Stanley R. Rotman[3]

[1] *The Unit of Electro-Optics Engineering and the Earth and Planetary Image Facility, Ben-Gurion University of the Negev, P.O. Box 653, 84105 Beer-Sheva, Israel*
[2] *The Department of Geography and Environmental Development and the Earth and Planetary Image Facility, Ben-Gurion University of the Negev, P.O. Box 653, 84105 Beer-Sheva, Israel*
[3] *Department of Electrical and Computer Engineering and the Earth and Planetary Image Facility, Ben-Gurion University of the Negev, P.O. Box 653, 84105 Beer-Sheva, Israel*

Correspondence should be addressed to Stanley R. Rotman, srotman@ee.bgu.ac.il

Received 28 February 2012; Accepted 5 June 2012

Academic Editor: James Theiler

Our goal in this work is to demonstrate that detectors behave differently for different images and targets and to propose a novel approach to proper detector selection. To choose the algorithm, we analyze image statistics, the target signature, and the target's physical size, but we do not need any type of ground truth. We demonstrate our ability to evaluate detectors and find the best settings for their free parameters by comparing our results using the following stochastic algorithms for target detection: the constrained energy minimization (CEM), generalized likelihood ratio test (GLRT), and adaptive coherence estimator (ACE) algorithms. We test our concepts by using the dataset and scoring methodology of the Rochester Institute of Technology (RIT) Target Detection Blind Test project. The results show that our concept correctly ranks algorithms for the particular images and targets including in the RIT dataset.

1. Introduction

Ideally, one would like to choose a hyperspectral detection algorithm for use in a particular scenario with the assurance that it would be "optimal," that is, that the type of algorithm to be used and its free parameters would be optimized for the particular task for which it is being considered. Of course, in such cases, the complexity of real-world scenarios and the difficulties of predicting the exact target signature *in situ*, make it hard to believe that we can predict the optimal target detection algorithm ahead of time. Because the responses of these algorithms can vary depending on target placement, we adapted the Rotman-Bar Tal Algorithm (RBTA) [1] for comparing point target detection algorithms, used for infrared broadband images, to the analysis of hyperspectral imagery [2–4]. The RBTA implants targets and evaluates the response of the detecting algorithm to their presence in every pixel in the dataset. Indeed, our development of new algorithms based on this tool has been validated by results obtained by other researchers in actual field tests [5, 6].

An inherent weakness of the RBTA method is its assumption that subpixel targets will each be contained within a single pixel. In light of our recent work [7], which showed that even very small targets can affect several pixels, here we fine-tuned the RBTA method to account for this possibility.

Sections 2–6 describes the RBTA in detail. We show how the simulation of target detection performance is dependent on the spatial correlation of the pixels present in the target.

Sections 7–12 analytically considers the expected performances of several detection algorithms under conditions of "pixel phasing," that is, a small target located simultaneously in several adjacent pixels. Our improved RBTA (IRBTA) takes into account target blurring and pixel phasing. The results presented in Sections 13–16 show that the superiority of the ACE algorithm and the importance of accounting for target blurring are validated in a real data analysis based on the

RIT target detection blind test experiment. Conclusions are presented in Section 17.

2. Determining the "Best Algorithm" for Target Detection

Manolakis et al. [8] claimed that to identify the best algorithm for target detection, we need datasets with reliable ground truth spanning a diverse range of targets and backgrounds under various scenarios, target fill factors, and atmospheric conditions. Statistically significant numbers of target and background pixels are necessary to construct reliable ROC curves. Because in many cases this degree of data confirmation is unavailable, we suggest an alternative approach for estimating the best algorithm from among several detectors for specific backgrounds and targets. We start by presenting the RBTA [1]. The algorithm was originally developed for broadband infrared images with subpixel targets, but we altered it to account for pixel blur (atmospheric and system effects which would cause the emitted power of the target to be spread over several pixels) and multipixel targets in hyperspectral imagery.

To estimate detector performance, Rotman and Bar-Tal proposed a multistep process that begins with an analysis of the unmodified reflectance image that is available in the website without any embedded targets. (We assume that ideally no targets are present in the datacube being analyzed; if one were present, it would slightly distort the histogram of the image. We trust that such a distortion will not disturb the overall analysis of the image statistics). The algorithm being tested is evaluated for each pixel, and the results are summarized in what we call a false-alarm histogram. Next we embed targets into every pixel and evaluate each of the algorithms. This is done independently for each pixel (rather than simultaneously) so that surrounding pixels are not changed prior to algorithm evaluation. The results are arranged in a target detection histogram. Each histogram (false-alarm and target detection) is then normalized; a variable threshold is set and the area of the false-alarm and target histograms to the right of the threshold are measured. For any particular threshold, a pair of P_{FA} and P_D (probability of false alarm and probability of detection) values are generated. The threshold is swept through all possible detector outputs, generating a set of these pairs. When graphed, these points produce the ROC (receiver operating characteristic) curves.

We note that the target implantation mechanism as given here has ignored several possibly significant effects which would affect the values found of PD. In particular, the target spectrum is a nearly noiseless lab spectrum that does not have the same artifacts, noise, and degradation as the real imagery. Additionally, this approach assumes the data has been perfectly atmospherically compensated by RIT's algorithm, which is not necessarily true. In our opinion, this seems to limit the use of our method rather than to invalidate it. Since the atmospheric conditions at the time of the measurement were not known, we cannot implant atmospherically corrected signatures or validate the reflectance dataset that is available in RIT website. Instead, we are testing the response of the algorithms to an implanted nonatmospherically corrected target which has been substituted in the reflectance dataset as described above; in each examined pixel, the fraction of the laboratory signature replaced the fraction of the background signal. While inaccurate atmospheric correction may result in an unknown decrease in the target detection, we note that the final comparisons are for variations in algorithm selection for a given target signature. The method should not be used to calculate absolute values for the probability of detection of a particular target which indeed has been altered by atmospheric and other effects. Rather, we are attempting to determine which algorithms will have a superior probability of detecting a target of this type in the scenario. Future work should include a quantitative determination to what degree atmospheric effects change the ranking of different algorithms.

This methodology can be used for the following reasons: as a rule, the ROC curve, which are generated tend to have probabilities of detection which range from 0 to 1; the value of probabilities of false alarm, on the other hand, vary from 0 to some chosen threshold $P_{\text{FA-MAX}}$. This threshold is normally set quote low; a standard value would be 0.01. This is appropriate since the acceptable use of most detection algorithms could only be in the range where a small percentage of the pixels in the image would be false alarms.

Now, the exact distribution of the background pixels is crucial for the analysis of our detection algorithms; it will indeed be the exceptional pixels in the tail of the distribution which will determine the ROC curve. However, since the probability of detection is being determined by the entire P_D scale from 0 to 1, all pixels contribute. In other words, the target detection scheme in this paper is extremely sensitive to a few false alarms; it is much less sensitive to a few pixels with missed "synthetic" target signatures. As such, subtle effects affecting the exact form of the target signature *in situ* are not being measured; rather the average response of the algorithm to the target signatures placement in all the pixels is the key factor. For our above goal, that is, the comparison of different target algorithms, we believe our method to be reliable.

To summarize, ROC curve evaluation entails the following steps as demonstrated in Figure 1.

3. Subpixel Target Detection: Global Methods

3.1. CEM. In many cases, it is convenient to scale the matched filter such that it has a value of 1 when the target signature fills the pixel being examined. This scaling can be achieved by normalizing the matched filter to its value when operating on the designated target spectrum:

$$\mathrm{CEM}(\bar{x}) = \frac{\bar{s}^T \bar{\bar{R}}^{-1} \bar{x}}{\bar{s}^T \bar{\bar{R}}^{-1} \bar{s}}, \tag{1}$$

where s is the reference signature of the target, R is the background correlation matrix, that is, an [L × L] matrix, L is the number of bands, and x is the observed pixel. Geometrically speaking, the CEM algorithm measures the projection of x onto s normalized by the length of s in the

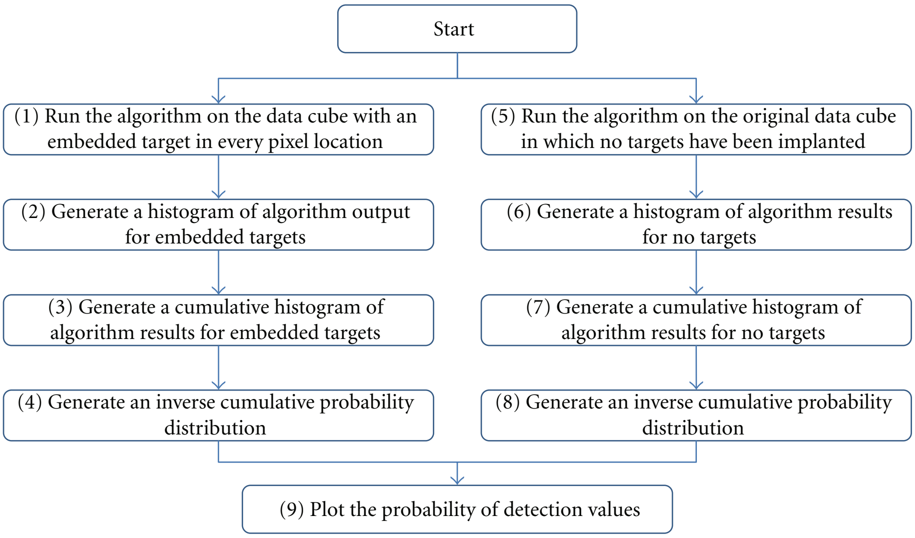

Start
(1) Run the algorithm on the data cube with an embedded target in every pixel location
(2) Generate a histogram of algorithm output for embedded targets
(3) Generate a cumulative histogram of algorithm results for embedded targets
(4) Generate an inverse cumulative probability distribution
(5) Run the algorithm on the original data cube in which no targets have been implanted
(6) Generate a histogram of algorithm results for no targets
(7) Generate a cumulative histogram of algorithm results for no targets
(8) Generate an inverse cumulative probability distribution
(9) Plot the probability of detection values

Figure 1: RBTA flow chart.

whitened space and thus leads to planar decision surfaces in that space. An important characteristic of the CEM algorithm is that its output is correlated to the target's fractional abundance in signature *x*, assuming the target signature is well isolated from the other endmembers, mixing is linear, and the relative abundances of the endmembers follow a Dirichlet distribution [9].

3.2. GLRT and ACE. Manolakis and his group [10–14] have described a number of stochastic target detection algorithms, including that attributed to Kelly [15] for solving to the Neyman-Pearson decision/detection theory for maximizing the probability of detection of a target with a fixed probability of false alarms. The solution uses a GLRT expressed as

$$\begin{aligned}\mathrm{GLRT}(\bar{x}) = &\left(\left[\left(\bar{s}-\overline{m_g}\right)^T \overline{\overline{G}}^{-1}\left(\bar{x}-\overline{m_g}\right)\right]^2\right) \\ &/\left(\left[\left(\bar{s}-\overline{m_g}\right)^T \overline{\overline{G}}^{-1}\left(\bar{s}-\overline{m_g}\right)\right]\right. \\ &\left.\cdot\left[1+(1/M)\cdot\left(\bar{x}-\overline{m_g}\right)^T \overline{\overline{G}}^{-1}\left(\bar{x}-\overline{m_g}\right)\right]\right),\end{aligned} \tag{2}$$

where *s* and *x* are the same as for (1), m_g is the global mean, *G* is the background covariance matrix, and *M* is the total number of samples.

The ACE algorithm, a variation of the GLRT algorithm, is expressed as

$$\begin{aligned}\mathrm{ACE}(\bar{x}) = &\left(\left[\left(\bar{s}-\overline{m_g}\right)^T \overline{\overline{G}}^{-1}\left(\bar{x}-\overline{m_g}\right)\right]^2\right) \\ &/\left(\left[\left(\bar{s}-\overline{m_g}\right)^T \overline{\overline{G}}^{-1}\left(\bar{s}-\overline{m_g}\right)\right]\right. \\ &\left.\cdot\left[\left(\bar{x}-\overline{m_g}\right)^T \overline{\overline{G}}^{-1}\left(\bar{x}-\overline{m_g}\right)\right]\right),\end{aligned} \tag{3}$$

with a maximum value of 1 for the case of $x = s$ and a minimum value of 0 when $x = m_g$.

In the context of target detection, the sign of $(s - m_g)^T G^{-1}(x - m_g)$ is important, as only positive abundances are of interest. (In contrast, this would not be the case for thermal gas detection, for example, where the target could be either absorptive or emissive in nature). Thus, in practice, a signed version of the GLRT algorithm is used as follows:

$$\mathrm{GLRT}_{\mathrm{sign}}(\bar{x}) = \mathrm{sign}\left[\left(\bar{s}-\overline{m_g}\right)^T \overline{\overline{G}}^{-1}\left(\bar{x}-\overline{m_g}\right)\right]\cdot \mathrm{GLRT}(\bar{x}). \tag{4}$$

The corresponding ACE algorithm for target detection, also a variation of the GLRT algorithm, is expressed as

$$\mathrm{ACE}_{\mathrm{sign}}(\bar{x}) = \mathrm{sign}\left[\left(\bar{s}-\overline{m_g}\right)^T \overline{\overline{G}}^{-1}\left(\bar{x}-\overline{m_g}\right)\right]\cdot \mathrm{ACE}(\bar{x}). \tag{5}$$

Because real data does not necessarily match the assumptions from which the above algorithms are derived, that is, a background probability distribution function assumed to be multivariate Gaussian with zero mean bias and an additive target model, we generally cannot expect that any of the algorithms will be optimal or even that one will consistently outperform another [8]. Nevertheless, it was shown by Manolakis [13] that for a limited dataset, although each of the algorithms exhibited some degree of success in target detection, the ACE algorithm performed best on the limited dataset tested.

In Figure 1, step 1, note that the target is not in all the positions simultaneously; rather, the result is obtained sequentially. Steps 4 and 8 are generated by one minus the cumulative histogram using the results from step 3 and 7, respectively, (these are the probability of detection-PD). In Step 9, we plot PD values (step 4) versus the PFA (step 8).

4. Subpixel Target Detection Using Local Spatial Information

Improving target detection involved replacing the global mean with the local mean. Using the local mean is definitely double edged: on one hand, we would expect that the closer the points used to evaluate the background are to the suspected target, the more likely it is that the estimate will be accurate. On the other hand, the noise in the estimate will decrease given more points entering into the estimation, assuming that the background is stationary and the noise is linearly added to the background and independent thereof. Our empirical experience confirmed by several studies (4) and (5) is that the closer we choose the pixels the better, with the condition that we do not have target contamination of the background pixels. It is this proviso that we wish to test here.

We note that we are not dealing here with a "local" covariance matrix which would change when evaluating each pixel in the image. Rather, we use the same covariance matrix throughout the image; it will simply be based on the difference of the sample pixels and their "local" background.

Since we are dealing with a subpixel target, which in the physical domain can affect only pixels in a limited spatial area surrounding the center of the target, we used the eight nearest neighbors approach to estimate the value of the test pixels. The CEM algorithm does not use the mean and will therefore be unaffected by the above changes. The GLRT can be improved as follows:

$$\begin{aligned}\mathrm{GLRT}_{\mathrm{local}}(\overline{x}) = {} & \left(\left[\left(\overline{s}-\overline{m_g}\right)^T\overline{\overline{G}}^{-1}(\overline{x}-\overline{m_8})\right]^2\right) \\ & /\left(\left[(\overline{s}-\overline{m_8})^T\overline{\overline{G}}^{-1}(\overline{s}-\overline{m_8})\right]\right. \\ & \left.\cdot\left[1+(1/M)\cdot(\overline{x}-\overline{m_8})^T\overline{\overline{G}}^{-1}(\overline{x}-\overline{m_8})\right]\right),\end{aligned} \tag{6}$$

and for target detection

$$\begin{aligned}\mathrm{GLRT}_{\mathrm{sign-local}}(\overline{x}) = {} & \operatorname{sign}\left[(\overline{s}-\overline{m_8})^T\overline{\overline{G}}^{-1}(\overline{x}-\overline{m_8})\right] \\ & \cdot\mathrm{GLRT}_{\mathrm{local}}(\overline{x}),\end{aligned} \tag{7}$$

with m_8, the mean of the eight nearest neighbors, replacing the global mean m_g. For the ACE detector, the same procedure (replacing m_g by m_8) may be followed.

Segmentation [16–18] or even more local covariance matrices [2, 4, 6, 19] can be used to improve the covariance matrix. Common to all these methods is an increased need for high performance computational resources, while the corresponding influence each method has on detection ability is uncertain and highly dependent on the pictures being analyzed. Used in parallel, the algorithms create new difficulties through the combination of results from different segments. We used a global covariance matrix, but adapted it to local variations by using the local rather than the global mean, that is,

$$\begin{aligned}\overline{\overline{G}}_{\mathrm{global}} &= \frac{\left[\overline{\overline{X}}-\overline{\overline{M_g}}\right]^T\left[\overline{\overline{X}}-\overline{\overline{M_g}}\right]}{M}, \\ \overline{\overline{G}}_{\mathrm{local}} &= \frac{\left[\overline{\overline{X}}-\overline{\overline{M_8}}\right]^T\left[\overline{\overline{X}}-\overline{\overline{M_8}}\right]}{M},\end{aligned} \tag{8}$$

where X is a two-dimensional matrix ($M \times L$), in which M is the number of pixels and L is the number of bands, m_g $[1 \times L]$ is the mean vector of X, and M_g *is* m_g replicate M times. When we use M_8 for the covariance matrix, we do not need to replicate the mean, because M_8 is also of size $[M \times L]$, and this is the appropriate covariance matrix for whitening $X - M_8$.

5. Data

We tested our algorithms on the online reflectance data sets and the hyperspectral data collected over Cooke City. The Cooke City imagery was acquired on 4 July 2006 using the HyMap VNIR/SWIR sensor with 126 spectral bands. Two hyperspectral scenes are provided with the dataset, one intended to be used for development and testing (the "Self Test" scene, where the positions of some targets are known) and the other intended to be used for detection performance evaluation (the "Blind Test" scene, where the position of targets is unknown). The data was corrected for atmospheric effects and available in the website but the exact atmospheric condition and the atmospheric correction algorithm are not available in the website and we assume that the reflectance dataset is good but not perfect. In Figure 2, we present the image in false color.

The target signatures, used both in the algorithm for detection and in the implantation of the synthetic targets in the RBTA method were laboratory measured and in reflectance units. The GSD is approximately 3 m. In Figure 3, we present the spectral signature of the targets in the blind test image.

The list of all targets is presented in Table 1 below.

6. Spatial Effect

6.1. Analytical and Simulated Performances of GLRT and ACE

6.1.1. Simple Case. The general form for local target detection as described in Section 3 is

$$\begin{aligned}D_{\mathrm{Local}}(\overline{x}) = {} & \left(\left[(\overline{s}-\overline{m_8})^T\overline{\overline{G}}^{-1}(\overline{x}-\overline{m_8})\right]^2\right) \\ & /\left(\left[(\overline{s}-\overline{m_8})^T\overline{\overline{G}}^{-1}(\overline{s}-\overline{m_8})\right]\right. \\ & \left.\cdot\left[\Psi_1+\Psi_2\cdot(\overline{x}-\overline{m_8})^T\overline{\overline{G}}^{-1}(\overline{x}-\overline{m_8})\right]\right),\end{aligned} \tag{9}$$

Table 1: Targets description.

Target ID	Target description	size (m^2) No. 1	size (m^2) No. 2	Self test ground truth	Blind test ground truth
F1	Red cotton fabric panel	3 × 3	N/A	Yes	No
F2	Yellow nylon fabric panel	3 × 3	N/A	Yes	No
F3	Blue cotton fabric panel	2 × 2	1 × 1	Yes	No
F4	Red nylon fabric panel	2 × 2	1 × 1	Yes	No
F5	Maroon nylon fabric panel	2 × 2	1 × 1	No	Web score
F6	Gray nylon fabric panel	2 × 2	1 × 1	No	Web score
F7	Green cotton fabric panel	2 × 2	1 × 1	No	Web score
V1	Chevy Blazer, green	4 × 2	N/A	Yes	Web score
V2	Toyota T100, white with black plastic liner	3 × 1.7	N/A	Yes	Web score
V3	Subaru GL Wagon, Red	4.5 × 1.6	N/A	Yes	Web score

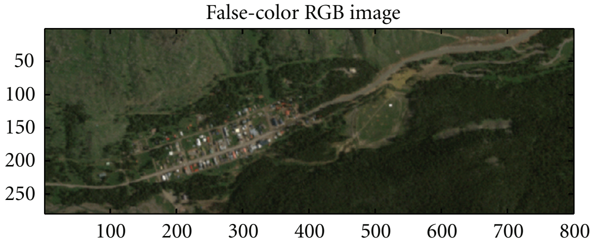


Figure 2: False-color RGB of the Cooke City imagery.

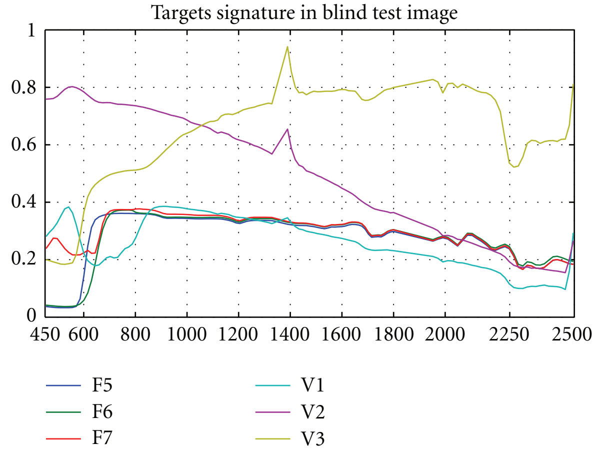


Figure 3: Spectral signatures of the targets that are present in the Blind test image x-axis is the wavelength [nm] and y-axis is the reflectance unit less.

with m_8 as the mean of eight neighbors. G, the global-local covariance matrix, is computed as

$$G_{\text{global\_local}} = \frac{\left(\overline{\overline{X}} - \overline{\overline{M_8}}\right)^T \cdot \left(\overline{\overline{X}} - \overline{\overline{M_8}}\right)}{L}, \tag{10}$$

where we can get GLRT and ACE as functions of $\Psi_1 + \Psi_2$:

$$\begin{aligned} \text{GLRT}: \Psi_1 &= M, \qquad \Psi_2 = 1, \\ \text{ACE}: \Psi_1 &= 0, \qquad \Psi_2 = 1. \end{aligned} \tag{11}$$

For the case in which the PUT (pixel under testing) x is exactly s, we obtain the following results:

$$\begin{aligned} D_{\text{Local}}(\bar{x}) = {} & \left(\left[(\bar{s} - \overline{m_8})^T \overline{\overline{G}}^{-1} (\bar{s} - \overline{m_8})\right]^2\right) \\ & / \left(\left[(\bar{s} - \overline{m_8})^T \overline{\overline{G}}^{-1} (\bar{s} - \overline{m_8})\right]\right. \\ & \left. \cdot \left[(\Psi_1 + \Psi_2) \cdot (\bar{s} - \overline{m_8})^T \overline{\overline{G}}^{-1} (\bar{s} - \overline{m_8})\right]\right). \end{aligned} \tag{12}$$

Let us define the scalar C as–

$$C = (\bar{s} - \overline{m_8})^T \overline{\overline{G}}^{-1} (\bar{s} - \overline{m_8}). \tag{13}$$

Therefore, when $\bar{x}$ is exactly $\bar{s}$, GLRT and ACE can be written as

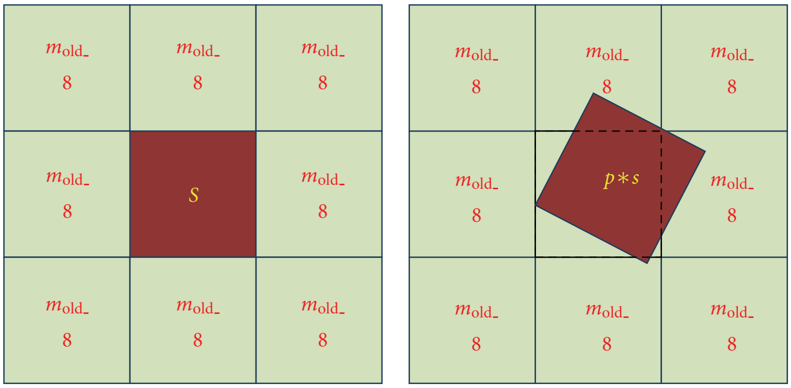
$$\begin{gathered} \text{GLRT}(\bar{s}) = \frac{C}{[M + C]} = \frac{1}{M/C + 1}, \\ \text{ACE}(s) = 1. \end{gathered} \tag{14}$$

Assuming that the data is normally distributed, C is chi-square distributed with $E(C) = \text{L}$, where L is the number of bands. For the case in which $M \gg E(C) = \text{L}$, we can assume that

$$\text{GLRT}(s) \cong \frac{C}{M}. \tag{15}$$

7. Pixel Phasing Case

When imaging, the target can often fall across several pixels even if its total size is only a single pixel; we will call this effect pixel phasing even though it is a natural consequence of imaging system quantization. The pixel phasing effect can be demonstrated by a target one pixel in size, the imaging of which leads to pixel phasing registration defined by p,



(a) Simple case

(b) Pixel phasing case

Figure 4: Pixel phasing schema.

such that $0 < p < 1$ (0 corresponds to perfect sampling, with the target completely replacing the background) (Figure 4). From the point of view of the central pixel, it is not important the spatial location of the fraction within the pixel nor the location of the remainder of the target signature. Assuming uniform backgrounds of $m_{\text{old\_8}}$ for both center pixels, they can now be given as in Figure 4.

We obtain the following:

$$\overline{x_{\text{New}}} = p \cdot \bar{s} + (1-p) \cdot \overline{m_{\text{old\_8}}}, \tag{16}$$

where x_{new} is the new PUT for the pixel phasing case and

$$m_{\text{new\_8}} = \frac{7+p}{8} \cdot \overline{m_{\text{old\_8}}} + \frac{1-p}{8} * \bar{s}, \tag{17}$$

where $m_{\text{new\_8}}$ is the new mean for the background.

We now evaluate the terms $(\bar{s} - \overline{m_{\text{new8}}})$ and $(\bar{x} - \overline{m_{\text{new8}}})$ as follows:

$$\begin{aligned}(\bar{s} - \overline{m_{\text{New8}}}) &= \bar{s} - \left[\frac{7+p}{8} \cdot \overline{m_{\text{old\_8}}} + \frac{1-p}{8} * \bar{s}\right] \\ &= \frac{(7+p)}{8} \cdot (\bar{s} - \overline{m_{\text{old\_8}}}),\end{aligned} \tag{18}$$

$$\begin{aligned}(\overline{x_{\text{New}}} - \overline{m_{\text{New8}}}) &= (p \cdot \bar{s} + (1-p) \cdot \overline{m_{\text{old\_8}}}) \\ &\quad - \left(\frac{7+p}{8} \cdot \overline{m_{\text{old\_8}}} + \frac{1-p}{8} * \bar{s}\right) \\ &= \frac{9 \cdot p - 1}{8} \cdot (\bar{s} - \overline{m_{\text{old\_8}}}).\end{aligned} \tag{19}$$

The GLRT result now becomes

$$\begin{aligned}D_{\text{Local}}(\bar{x}) = &\left([((9 \cdot p - 1)/8) \cdot (((7+p)/8) \cdot C)]^2\right) \\ &/([((9 \cdot p - 1)/8) \cdot ((9 \cdot p - 1)/8) \cdot C] \\ &\cdot [\Psi_1 + \Psi_2((7+p)/8) \cdot ((7+p)/8) \cdot C]),\end{aligned} \tag{20}$$

where $D_{\text{Local}}(\bar{x})$ is the general local detector for the pixel phasing case.

For the case in which $N \gg C$, we calculate that

$$\begin{aligned}\text{GLRT}_{\text{miss\_sampling}}(\bar{x}) &\cong \frac{(p^2 + 14p + 49)}{64} \cdot \frac{C}{M} \\ &= \frac{(p^2 + 14p + 49)}{64} \cdot \text{GLRT}_{\text{local}},\end{aligned} \tag{21}$$

where $\text{GLRT}_{\text{miss\_sampling}}(x)$ is the expected GLRT value for the pixel phasing case and $M \gg L$. The GLRT expected value degrades as a function of p. But for ACE ($\Psi_1 = 0$, $\Psi_2 = 1$) we still get expected values of 1:

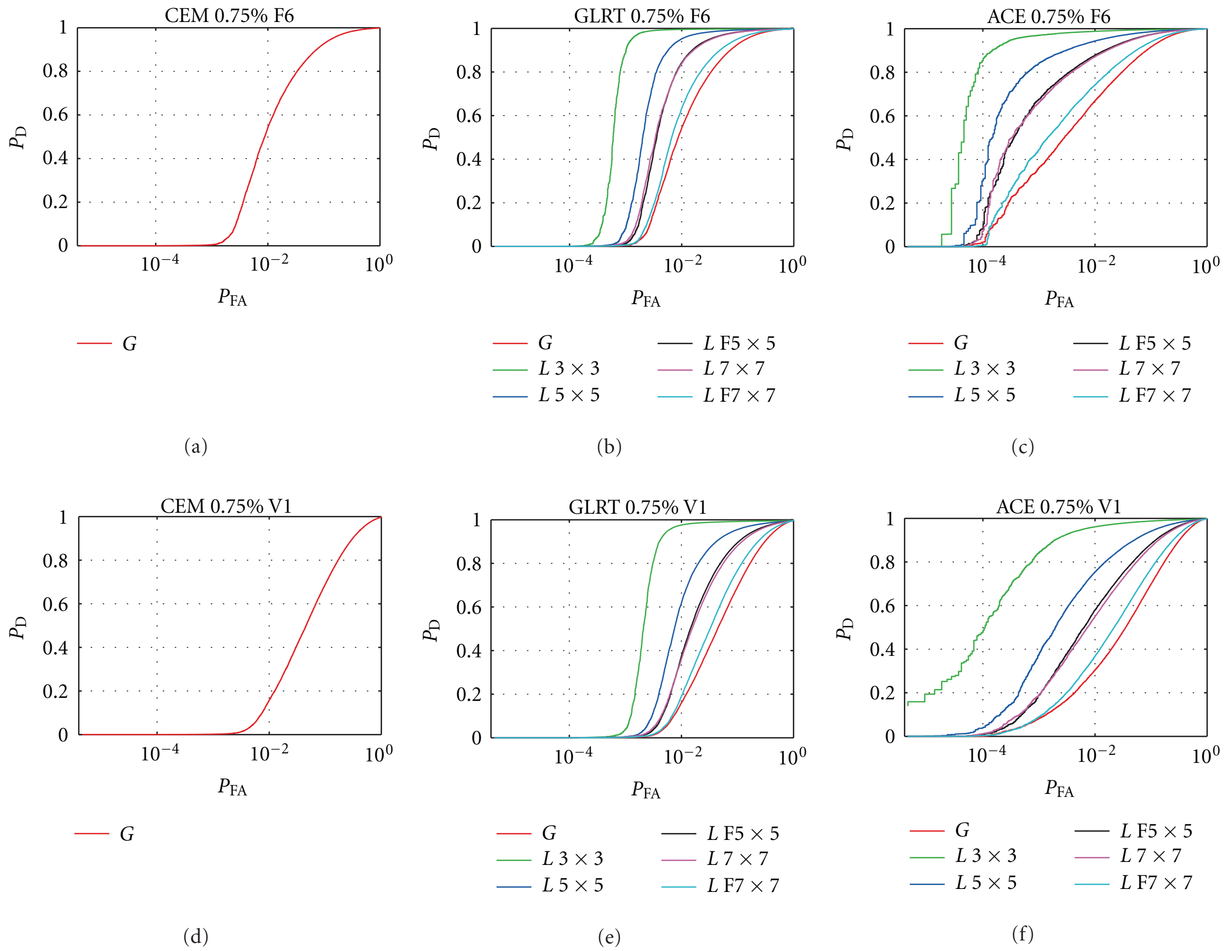
$$\text{ACE}_{\text{miss\_sampling}}(x) = 1 = \text{ACE}_{\text{local}}(x). \tag{22}$$

In this model, the complete lack of ACE degradation as a function of pixel phasing may explain why ACE is a more robust detector than GLRT in many test cases, as noted in the literature [8, 20].

8. Ranking the Algorithms by RBTA

The difficult task of synthesizing a synthetic image to help predict which algorithm to select is simplified and detector selection is facilitated if we synthesize only the target signature of our real image. Suppose we want to determine the proper detector for a specific target. We have already selected our method (e.g., CEM, GLRT, or ACE), and now we want to select the size of the local window. One approach is to assume that the best size for the local window is that under which the PUT value can be predicted with minimum error vis-à-vis the real PUT (in which we normalize each band by the mean values of the pixels in that band).

The approach outlined above depends only on the background image, not on the target signature, and it entails two assumptions: first, estimating signature values will improve our detector results independent of the different target signatures and second, the target has no effect on its neighbors. Address these assumptions in the following sections.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 5: RBTA results for different size of local windows.

9. How to Use RBTA

The implementation of RBTA, which depends on our ability to implant realistic signals into backgrounds and measure detector response, should be done carefully. We cannot expect the real signature to be identical to a library signature, but we can hope for a high level of similarity. The low percentage of the target signature that actually enters any particular pixel is demonstrated in Figure 6; the response of the CEM filter, which responds proportionally to the percentage of the target fill in the tested filter, was maximum at 0.06.

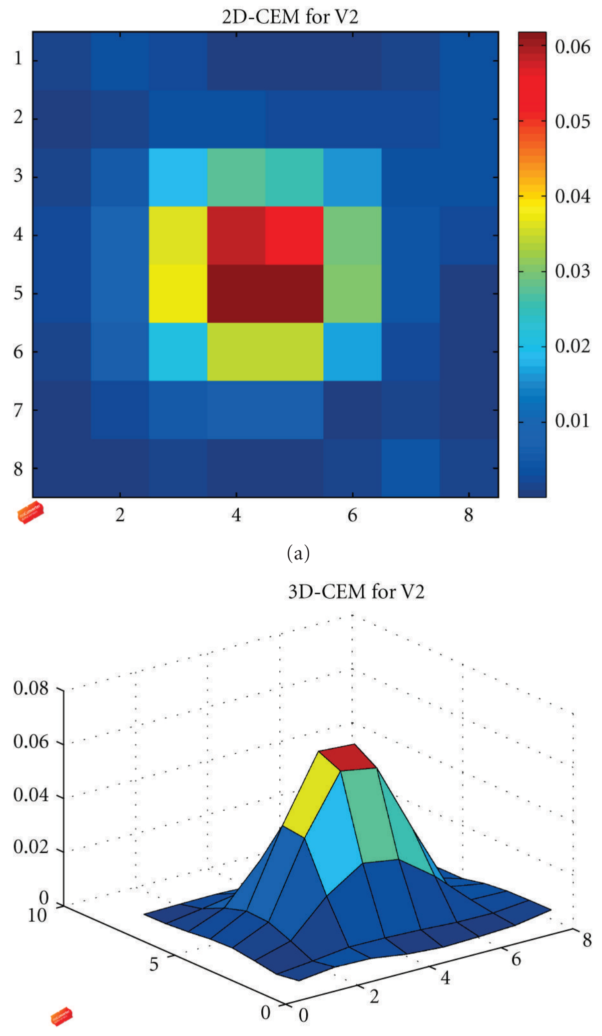
As a rule, to test and challenge our algorithms by examining the area under the ROC curve, we need to test targets which neither "saturate" the ROC curve (with a probability of detection close to one with no false alarms detected) nor result in a "diagonal" ROC curve (in which the probability of detection equals the probability of false alarms. As the allowable false alarm rate decreases, the strength of our synthetic implanted target would need to increase; if we know what the acceptable false alarm rate is, we can select the target percent that will demonstrate the dynamic range around this rate and get results for our detectors (Figure 5).

For the values found experimentally for p, the target was easily detectable and saturated our ROC curve. Thus, we only embedded 0.0075 of the target signature in the background pixels to generate the target detection histogram. In our results, we found that for GLRT and ACE, the best local window size is 3 × 3 pixels (CEM has no local form). We also see that using bigger windows to estimate the pixel signature value gets us closer to the performance of global detectors that use a global mean. In this case, it is clea that local detectors are superior to global detectors.

In terms of real data, we must expect each target to affect more than one pixel even if its total physical size is at the subpixel level. A discussion of this point follows below and leads to improvement of the RBT algorithm.

10. Improvements to RBTA

10.1. Target Size. As will be discussed in Sections 11-12 the apparent target size in the final digital image is related both to its physical size and to various atmospheric and sensor effects, for example, its point spread function (PSF), Gibbs effect, crosstalk between pixels, spatial sampling, band-to-



(a)

(b)

Figure 6: CEM results (2D and 3D) for target with pixel size.

band misregistration [5], and motion compensation. Thus, a target of a single pixel could actually occupy several pixels. In the RIT blind test, there are two 3×3-m targets, that is, exactly the size of the ground sample resolution (GSD) for the self and blind test images. Figure 6 shows a sample target of this size.

11. Spatial Sampling Effect

If we take into account only the spatial sampling effect, we can estimate the percent of pixel area partially occupied by the target. Notice that even targets of subpixel size often spread over neighboring pixels (Figure 7).

Put formally, the percent of pixel area covered as a function of target size, target location, and target orientation is

$$S_{\text{target\_in\_pixel}} = \iint_{-0.5}^{0.5} \frac{4}{\pi} \cdot (\alpha \cdot \beta) \cdot \text{UnitBox}[\tilde{x}, \tilde{y}] \cdot dx\, dy,$$

$$\tilde{x} = [\alpha \cdot [(x - \Delta x) \cdot \cos(\theta) - (y - \Delta y) \cdot \sin(\theta)]],$$
$$\tilde{y} = [\beta \cdot [(x - \Delta x) \cdot \sin(\theta) + (y - \Delta y) \cdot \cos(\theta)]], \tag{23}$$

where θ represents the clockwise rotation of the target relative to the pixel grid, and α, β represent the proportions of target length and width, respectively, relative to pixel physical dimension, Δx, Δy are the transition of the target origin relative to the pixel origin, as demonstrated in Figure 8

$$\text{UnitBox}[x, y] = \begin{cases} 1 & -\frac{1}{2} < x < \frac{1}{2} \text{ and}, -\frac{1}{2} < y < \frac{1}{2}, \\ 0 & \text{else}. \end{cases} \tag{24}$$

The expected value $E[S\alpha, \beta]$ is

$$E\left[S_{\alpha,\beta}\right] = \iiint_D S_{\text{target\_in\_pixel}}\, d\Delta x \cdot d\Delta y \cdot d\theta,$$
$$D = \begin{Bmatrix} 0 \le \Delta x \le 0.5 \\ 0 \le \Delta y \le 0.5 \\ 0 \le \theta \le \pi \end{Bmatrix}. \tag{25}$$

If we set θ to a constant value of zero and the target length and width are half the size of a pixel, we can simulate all the locations of the target where it's covering the same percent of pixel area (Figure 9).

Calculating (25) for a different size target using a numerical example produced the results shown in Figure 10.

In the graphs depicting pixel coverage as a function of physical target size, the x-axis is the ratio either between the target area and the pixel area (Figure 10(a)) or between the target length and the pixel length (Figure 10(b)). The blue line in both figures represents the percent of target within the pixel, while the green line is the percent of the pixel expected to be covered. It is intuitive that a very small target will be located in only one pixel, covering a small percent of that pixel. It is less intuitive, however, that the expected pixel to be covered will be entirely covered only by a target with an area four times that of the pixel.

12. Point Spread Function Effect

The PSF effect, present in any optical system, is not always known. Let us assume that the PSF is a typical, rotationally symmetric Gaussian filter of size 3 × 3 with standard deviation sigma 1/2.

Figure 11 demonstrates the synthetic spread effect that emerges from the convolution of the optical PSF (Figure 11(a)), and the physical pixels phasing due to target size (Figure 11(b)). For the spatial sampling we took the mean case representing the average pixel phasing that we could expect. Figure 11(c) represents the total effect, for example, convolution between (a) and (b). We devised an improved RBTA (IRBTA) and embedded the pixel signature and its neighbors with ratios as shown in Figure 11(c).

A comparison of Figures 5 and 12 shows that global detector and local detectors that both use only 7 × 7 frames

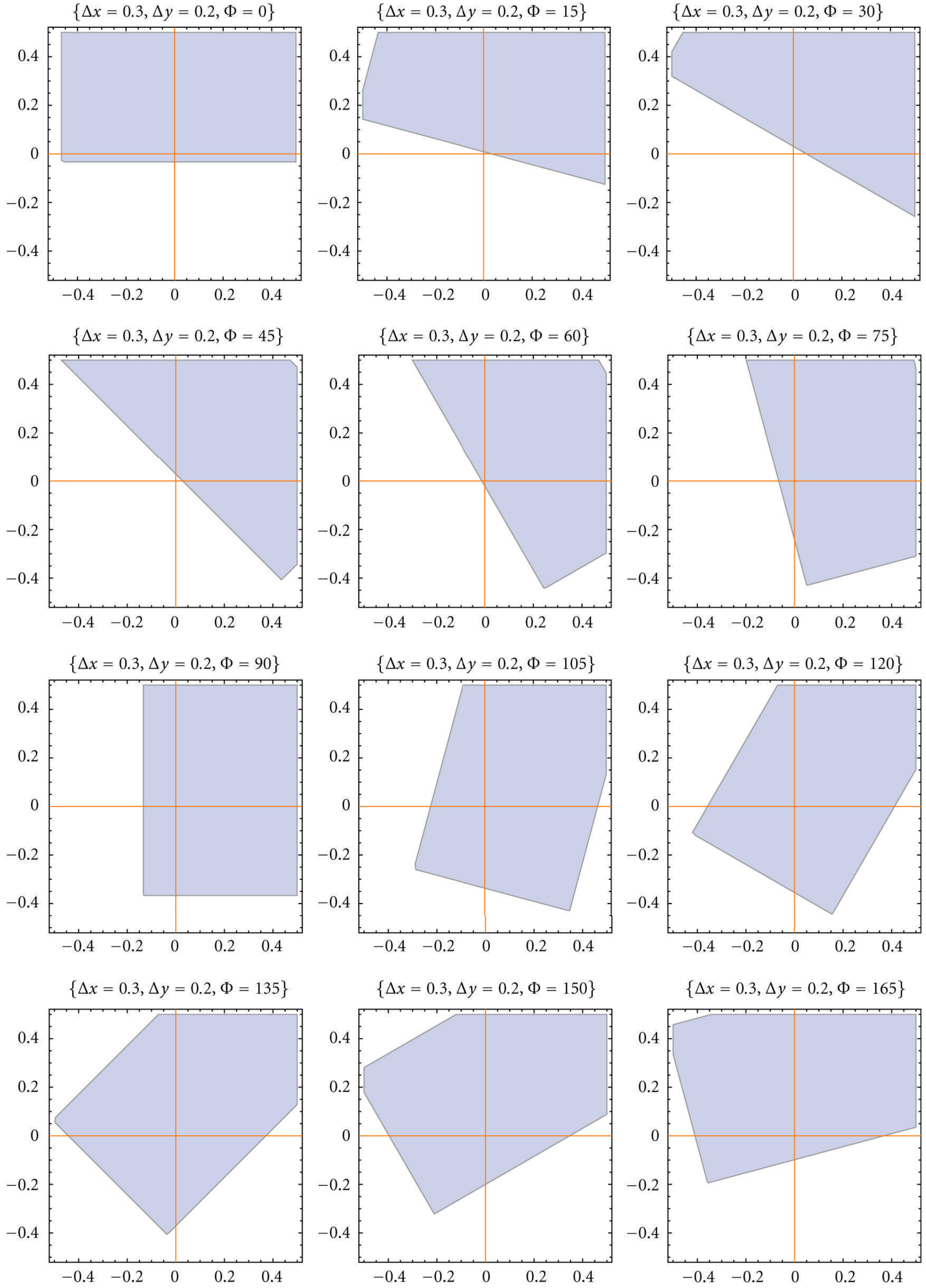


Figure 7: Demonstration for target size of 4/3 over 2/3 with origin at (0.3,0.2) and different rotation angle.

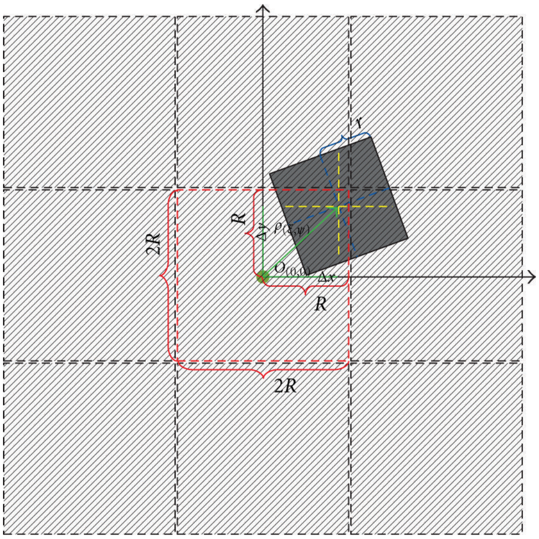


Figure 8: Variable demonstration.

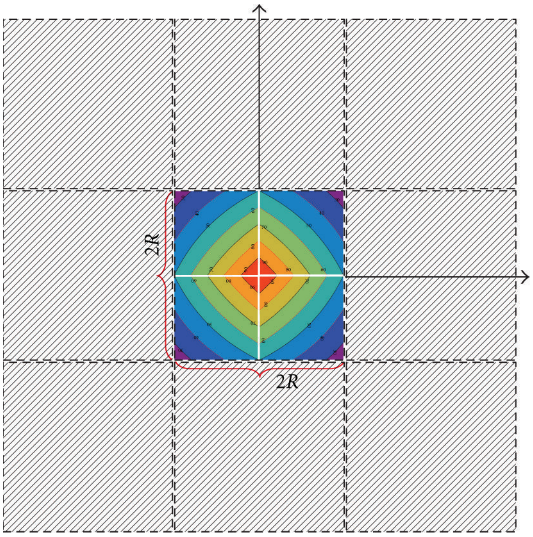


Figure 9: Pixel covered as function of target size for $\theta = 0$, $\alpha = 2$, and $\beta = 2$.

have exactly the same performance. Evidently, the target spread effect is only localized in a 5×5 window. Furthermore, the most significant effect is for a local 3 × 3 window. Indeed, the local 3 × 3 seems robust, and it outperformed all other detectors. We expected these detectors to be the best for this target in this image.

For our next stage of proof of concept, we need to compare real detection performances to these simulated results. We obtain this by using the self-test dataset and by submitting our algorithms to the RIT target detection blind test and comparing the ranking of our algorithms for each of the target signatures available in the set; we can then see if the IRBTA predictions of the preferred algorithms are true.

13. Detection Performance: Experimental Results

13.1. Scoring Methodology and Result Presentation. Detection algorithm performances are given according to the RIT target detection methodology applied to the aforementioned Cooke City hyperspectral dataset. The score is based on a comparison of the values given to the background pixels in the image to the value given to the target. The target value is defined as the maximum value given the pixels in the target area; the metric then counts how many pixels there are in the overall image greater than or equal to the target value. Since the threshold needed to detect the target would have to be less than or equal to the target value, all points above this value are false alarms. The score given for any algorithm/target combination would thus be the number of pixels above the target value. Perfect detection would equal the value 1, since the only pixel equal or above the target value would be the target itself; no false alarms are present.

In Tables 2, 3, and 4 the scores for the self-test targets were calculated, and those for the blind test were obtained by uploading our results to the RIT website. We have colored coded the results for the ease of the reader. A value of 1 (italic background) indicates perfect performance (no false positives). We marked all scores greater than 448 (e.g., PFA = $2 * 10^{-3}$) with bold italic backgrounds. We will associate a false alarm score greater than this as a fundamentally undetected target; ranking above these values is irrelevant All results (i.e., scores) were divided among the three tables: Table 2 presents the global method (CEM, GLRT and ACE) while Tables 3 and 4 contain the results for the local GLRT and ACE, respectively, with different sized windows. The best score for each group is marked with a bold. Ignoring (as stated above) the PFA values $> 2 * 10 - 3$ (bold italic background), then the global ACE is clearly the best method from among the global methods. The local GLRT and ACE with the 3 × 3 windows are the best of the local GLRT and ACE methods, respectively. Indeed, the overall performance of the local ACE algorithm with 3 × 3 windows was superior to all other algorithms. All these results were obtained using the IRBTA without any need of ground truth. Some targets, for example, V3 in the blind test and V2 and V3 in the self test, degraded with the 3 × 3 window of the local ACE algorithm, but those targets were effectively undetectable with all three algorithms.

14. Global Methods: Results

Within the results for the global methods (Table 2), we marked detector scores indicating a significant advantage relative to other detectors in bold. Our results clearly show that the ACE detectors outperformed the other global detectors.

15. Local GLRT and Local ACE: Results

Similar to the global methods analysis, here (Tables 3 and 4) we also applied **yellow** highlights to cells with exceptional

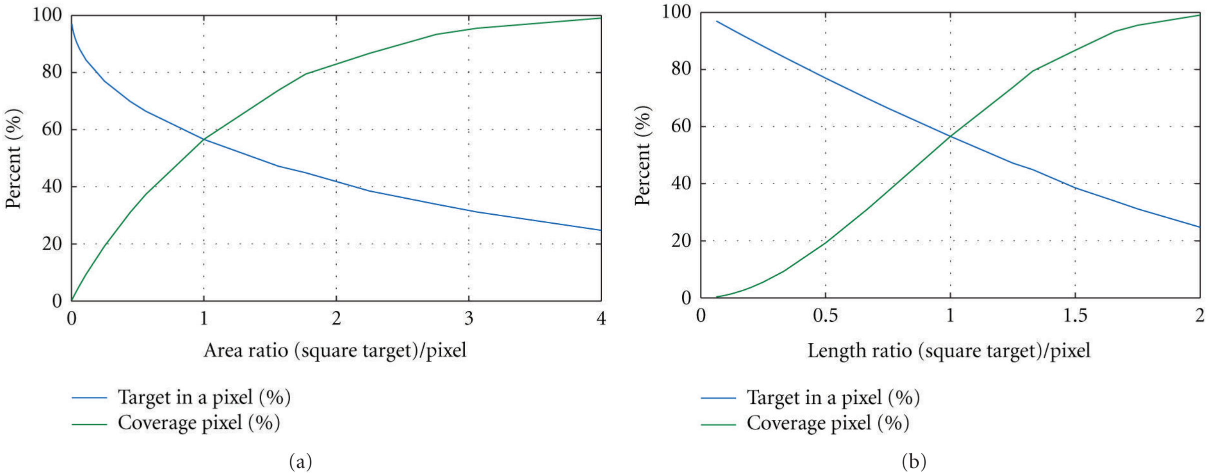


Figure 10: Pixel coverage as a function of target size.

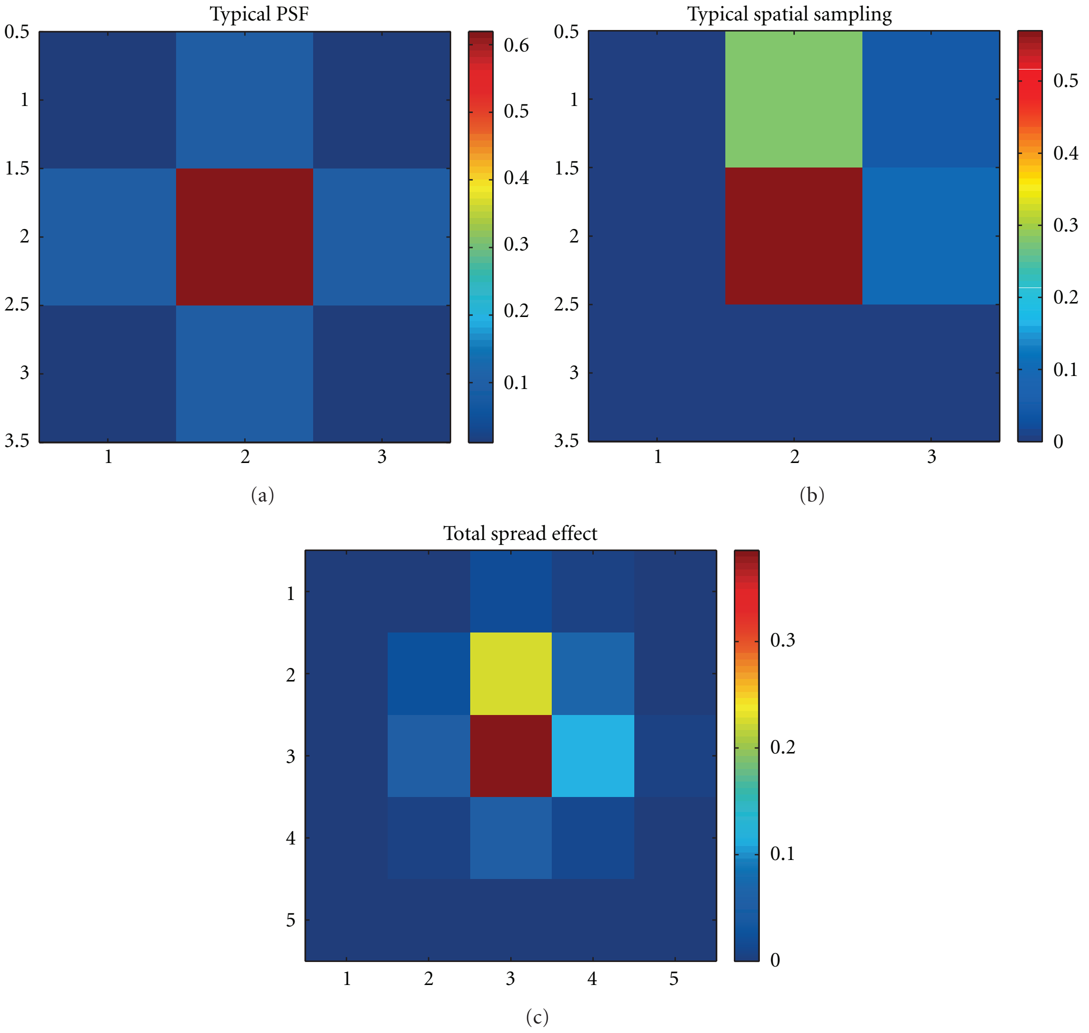
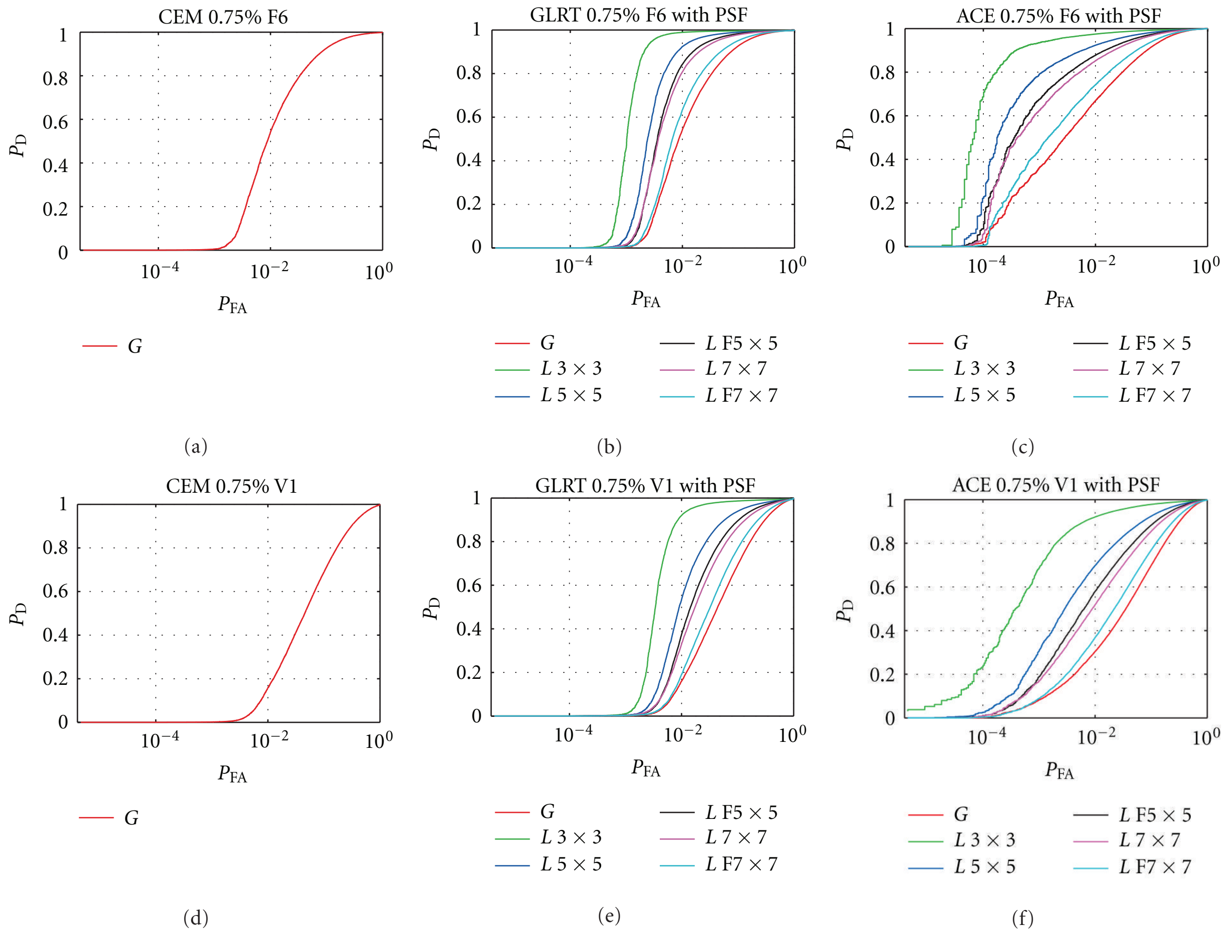


Figure 11: Synthetic spread emulation.



(a) (b) (c)

(d) (e) (f)

Figure 12: IRBTA results for different type of detectors.

detector scores that signified a significant advantage relative to other detectors. From our analysis, it is clear that for both the GLRT (Table 3) and ACE (Table 4) detectors, the optimal size of the local window for this size target is 3 × 3, a result that is in accordance with our estimation by IRBTA.

16. Benchmark Results

The RIT website provides a comparison between the results of different algorithms which have been submitted throughout the world. Table 5, which presents the web rank for the blind test, clearly shows that the performance of the local ACE 3×3 algorithm was superior relative to dozens of results that have been uploaded to the site. We achieved perfect detection for V1 (the nearest score was 15 for algorithms that work well only for V1) and reasonable detection for V2 (the nearest score was 196). There was no algorithm that achieved detection for V3 (the best score was 112). (The above conclusions are correct if we do not take into account one algorithm that was submitted to the RIT site, labeled “WTA”, i.e., “Winner Take All.” This algorithm exhibited perfect detection for all targets, including the vehicles, However, since the WTA algorithm has not been published, we cannot test it. In addition, there are objective reasons to believe that the algorithm does not actually exist; perfect detection of all targets with no false alarms ever would be an almost impossible result). While we included WTA in our scoring in Table 5, our comments following the Table do not consider this algorithm).

It is possible to compare the actual results obtained by the target detection algorithms on the RIT system to those obtained from the IRBTA simulation. In Tables 6 and 7, we show the results of several different algorithms when detecting the target V1; when we examine the results from the RIT test, we see that the ACE algorithm outperformed the GLRT algorithm, and that there was a definite preference for the 3×3 frame compared to the other frames (Table 6). These results are mirrored in our IRBTA results (Table 7). This (and results on the other targets) confirms our hypothesis that the RBTA can be used to simulate and predict target detection probabilities a priori in scenarios where actual targets are not yet present.

17. Conclusion

In this paper, we showed that there is no “best hyperspectral detection algorithm” for all images and targets. We noted the significant effect spatial distribution has on detector

Table 2: Results of global methods.

			Global		
	Target ID		CEM	GLRT	ACE
Self-test		F1	15	13	*1*
		F2	*1*	*1*	*1*
	F3	1 m^2	***1350***	***1366***	***1157***
		2 m^2	30	28	**1**
	F4	1 m^2	208	207	**118**
		2 m^2	23	24	**1**
Blind test	F5	1 m^2	163	156	**18**
		2 m^2	19	19	**1**
	F6	1 m^2	75	76	**7**
		2 m^2	13	13	**1**
	F7	1 m^2	***1204***	***1290***	***1792***
		2 m^2	315	318	**2**
Self-test		V1	324	321	**34**
		V2	***2028***	***1852***	***1027***
		V3	***853***	***775***	**367**
Blind test		V1	422	428	**179**
		V2	***931***	***921***	***1365***
		V3	***3230***	***3154***	***2999***

Table 3: Results of local GLRT for different-sized windows.

			GLRT local				
	Target ID		3 × 3	5 × 5	F5 × 5	7 × 7	F7 × 7
Self-test		F1	**4**	5	6	6	8
		F2	*1*	*1*	*1*	*1*	*1*
	F3	1 m^2	**188**	208	230	293	426
		2 m^2	**9**	13	13	18	22
	F4	1 m^2	**65**	101	121	120	152
		2 m^2	**4**	8	10	12	18
Blind test	F5	1 m2	**15**	48	66	78	125
		2 m^2	*1*	*1*	*1*	*1*	3
	F6	1 m^2	**14**	24	34	33	54
		2 m^2	**3**	5	5	5	6
	F7	1 m^2	**92**	152	202	244	388
		2 m^2	**81**	120	152	160	230
Self-test		V1	101	**74**	87	77	107
		V2	***6734***	***3456***	***3286***	***2812***	***2891***
		V3	***2506***	***875***	***741***	***706***	***712***
Blind test		V1	**37**	46	58	77	156
		V2	**283**	392	435	***518***	***694***
		V3	***11274***	***6455***	***5623***	***8259***	***10060***

Table 4: Results of local ACE for different-sized windows.

			ACE local				
	Target ID		3 × 3	5 × 5	F5 × 5	7 × 7	F7 × 7
Self-test		F1	*1*	*1*	*1*	*1*	*1*
		F2	*1*	*1*	*1*	*1*	*1*
	F3	1 m^2	**16**	17	19	33	62
		2 m^2	*1*	*1*	*1*	*1*	*1*
	F4	1 m^2	**50**	68	75	72	75
		2 m^2	*1*	*1*	*1*	*1*	*1*
Blind test	F5	1 m^2	**11**	**11**	12	13	13
		2 m^2	*1*	*1*	*1*	*1*	*1*
	F6	1 m^2	6	5	5	5	5
		2 m^2	*1*	*1*	*1*	*1*	*1*
	F7	1 m^2	**5**	**5**	**5**	**5**	9
		2 m^2	2	2	2	2	*1*
Self-test		V1	**3**	**3**	**3**	**3**	7
		V2	***3148***	***1524***	***1457***	***1171***	***1416***
		V3	***2387***	***1143***	***961***	***812***	***792***
Blind test		V1	**1**	5	8	15	21
		V2	**79**	106	119	198	359
		V3	***12513***	***6859***	***6327***	***10543***	***16328***

Table 5: Benchmark results.

	Target ID		Local ACE 3 × 3	Web rank
Blind test	F5	1 m^2	**11**	12/148
		2 m^2	*1*	1/148
	F6	1 m^2	6	11/90
		2 m^2	*1*	1/90
	F7	1 m^2	**5**	3/82
		2 m^2	**2**	5/82
Blind test	V1		**1**	1/50
	V2		**79**	3/82
	V3		***12513***	52/86

performances, and we showed that the RBTA can be used to select the proper detectors from among several detectors but without any need for ground truth. However, point targets can influence their neighboring pixels, due either to the PSF or to the target spreading across more than one pixel. To account for this potential source of inaccuracy, therefore, we introduced the improved RBTA (IRBTA), whose exact method of use depended on the target size. In addition, we showed that when detectors calculated the mean for estimating the pixel signature value, we did not need ground truth to find the best estimate. We tested our concept through the selection of the best detectors from among stochastic algorithms for target detection, that is, the constrained energy minimization (CEM), generalized likelihood ratio test (GLRT), and adaptive coherence estimator (ACE) algorithms, using the dataset and scoring methodology of the Rochester Institute of Technology (RIT) Target Detection Blind Test project. The results showed that our concepts predicted the best algorithms for the particular images and targets provided by the website.

Table 6: RIT results for the actual detection of V1 in RIT test image are shown in the first two lines. The percentage of implanted target was 0.75%. The GLRT and ACE algorithms were calculated as presented in the text. The size of the background was calculated for 3×3, 5×5 and 7×7 frames, excluding the center pixel. The "Only" 7×7 and 5×5 algorithms only used the outer ring of the window. The third and fourth lines represent the same results normalized by dividing by the values obtained by the 3×3 filter.

Website results						
Window	Only 7×7	7×7	Only 5×5	5×5	3×3	Global
V1_GLRT	156	77	58	46	37	428
V1_ACE	21	15	8	5	1	179
Normalize relative to 3×3						
V1_GLRT	**4.22**	**2.08**	**1.57**	**1.24**	**1**	**11.57**
V1_ACE	**21**	**15**	**8**	**5**	**1**	**179**

Table 7: The A_{th} $((A - \text{th}^2)/(\text{th} - \text{th}^2))$ where A is the area under the Pd-log (P_{fa}) curve and th is the threshold (i.e., the maximum false alarm rate) results of the IRBTA algorithm for V1 in the RIT test image are shown in the first two lines. The maximum false alarm rate is 10^{-3}; other parameters are as given in Table 6. The third and fourth lines represent the same results normalized by dividing the value of the 3×3 filter by the values of the other filters.

Ath by IRBTA-Threshold = 10^{-3}						
Window	Only 7×7	7×7	Only 5×5	5×5	3×3	Global
V1_GLRT	0.0007	0.0011	0.0010	0.0014	0.0032	0.0003
V1_ACE	0.0436	0.0908	0.0946	0.1657	0.4981	0.0420
Normalize relative to 3×3						
V1_GLRT	**4.81**	**2.97**	**3.35**	**2.36**	**1**	**11.83**
V1_ACE	**11.44**	**5.49**	**5.26**	**3.01**	**1**	**11.86**

Acknowledgment

The authors would like to acknowledge the contribution of the Rochester Institute of Technology (RIT) Digital Imaging and Remote Sensing (DIRS) Laboratory to this work by providing the dataset and an open forum to discuss hyperspectral target detection. The authors also acknowledge the Paul Ivanier Center for Robotics and Industrial Production, Beer-Sheva, Israel, for partial support of this work. (The work in this paper was performed and submitted before [21] came to our attention. There are several technical differences between the papers; together, they fortunately confirm the basic theses proposed in this paper).

References

[1] M. Bar-Tal and S. R. Rotman, "Performance measurement in point source target detection," *Infrared Physics and Technology*, vol. 37, no. 2, pp. 231–238, 1996.

[2] C. Caefer, J. Silvermana, O. Orthalb et al., "Improved covariance matrices for point target detection in hyperspectral data," *Optical Engineering*, vol. 47, no. 7, Article ID 076402, 2008.

[3] C. E. Caefer, M. S. Stefanou, E. D. Nielsen, A. P. Rizzuto, O. Raviv, and S. R. Rotman, "Analysis of false alarm distributions in the development and evaluation of hyperspectral point target detection algorithms," *Optical Engineering*, vol. 46, no. 7, Article ID 076402, 2007.

[4] C. E. Caefer and S. R. Rotman, "Local covariance matrices for improved target detection performance," in *Proceedings of the 1st Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS '09)*, August 2009.

[5] J. T. Casey and J. P. Kerekes, "Misregistration impacts on hyperspectral target detection," *Journal of Applied Remote Sensing*, vol. 3, no. 1, Article ID 033513, 2009.

[6] S. Matteoli, N. Acito, M. Diani, and G. Corsini, "An automatic approach to adaptive local background estimation and suppression in hyperspectral target detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 2, pp. 790–800, 2011.

[7] Y. Cohen, D. G. Blumberg, and S. R. Rotman, "Sub-pixel target detection using local spatial information in hyperspectral images," in *Proceedings of the Electro-Optical Remote Sensing, Photonic Technologies, and Applications V.*, Proceedings of SPIE, Prague, Czech Republic, 2011.

[8] D. Manolakis, R. Lockwood, T. Cooley et al., "Is there a best hyperspectral detection algorithm?" in *Proceedings of the Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XV*, Proceedings of SPIE, Orlando, Fla, USA, 2009.

[9] J. Settle, "On constrained energy minimization and the partial unmixing of multispectral images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 40, no. 3, pp. 718–721, 2002.

[10] D. Manolakis, C. Siracusa, and G. Shaw, "Hyperspectral subpixel target detection using the linear mixing model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 7, pp. 1392–1409, 2001.

[11] D. G. Manolakis and G. Shaw, "Detection algorithms for hyperspectral imaging applications," *IEEE Signal Processing Magazine*, vol. 19, no. 1, pp. 29–43, 2002.

[12] D. Manolakis, "Detection algorithms for hyperspectral imaging applications: a signal processing perspective," in *IEEE Workshop Advances in Techniques for Analysis of Remotely Sensed Data*, pp. 378–384, October 2003.

[13] D. Manolakis, "Hyperspectral signal models and implications to material detection algorithms," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04)*, pp. III117–III120, May 2004.

[14] D. Manolakis, "Taxonomy of detection algorithms for hyperspectral imaging applications," *Optical Engineering*, vol. 44, no. 6, Article ID 066403, 2005.

[15] E. J. Kelly, "An adaptive detection algorithm," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 22, no. 2, pp. 115–127, 1986.

[16] S. R. Rotman, J. Silverman, and C. E. Caefer, "Segmentation and analysis of hyperspectral data," in *Proceedings of the 22nd Convention of Electrical and Electronics Engineers in Israel*, 2002.

[17] J. Silverman, C. E. Caefer, J. M. Mooney, M. M. Weeks, and P. Yip, "Automated clustering/segmentation of hyperspectral images based on histogram thresholding," in *Proceedings of the Imaging Spectrometry VII*, vol. 4480 of *Proceedings of SPIE*, pp. 65–75, San Diego, Calif, USA, August 2001.

[18] J. Zhang, J. Chen, Y. Zhang, and B. Zou, "Hyperspectral image segmentation method based on spatial-spectral constrained

region active contour," in *Proceedings of the 30th IEEE International Geoscience and Remote Sensing Symposium (IGARSS '10)*, pp. 2214–2217, Honolulu, Hawaii, USA, July 2010.

[19] X. Yu, I. S. Reed, and A. D. Stocker, "Comparative performance analysis of adaptive multispectral detectors," *IEEE Transactions on Signal Processing*, vol. 41, no. 8, pp. 2639–2655, 1993.

[20] J. Schott, *Remote Sensing: The Image Chain Approach*, Oxford University Press, New York, NY, USA, 2007.

[21] W. F. Basener, E. Nance, and J. Kerekes, "The target implant method for predicting target difficulty and detector performance in hyperspectral imagery," in *Proceedings of the Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XVII*, vol. 8048, 80481H of *Proceedings of SPIE*, Orlando, Fla, USA, 2011.

Hindawi Publishing Corporation
Journal of Electrical and Computer Engineering
Volume 2012, Article ID 628479, 7 pages
doi:10.1155/2012/628479

Research Article

Target Detection Using Nonsingular Approximations for a Singular Covariance Matrix

Nir Gorelik,[1] Dan Blumberg,[2] Stanley R. Rotman,[1] and Dirk Borghys[3]

[1] *Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel*
[2] *Department of Geography and Environmental Development, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel*
[3] *Signal and Image Centre, Royal Military Academy, 1000 Brussels, Belgium*

Correspondence should be addressed to Nir Gorelik, nir.gorelik@gmail.com

Received 1 April 2012; Accepted 7 June 2012

Academic Editor: Xiaofei Hu

Accurate covariance matrix estimation for high-dimensional data can be a difficult problem. A good approximation of the covariance matrix needs in most cases a prohibitively large number of pixels, that is, pixels from a stationary section of the image whose number is greater than several times the number of bands. Estimating the covariance matrix with a number of pixels that is on the order of the number of bands or less will cause not only a bad estimation of the covariance matrix but also a singular covariance matrix which cannot be inverted. In this paper we will investigate two methods to give a sufficient approximation for the covariance matrix while only using a small number of neighboring pixels. The first is the quasilocal covariance matrix (QLRX) that uses the variance of the global covariance instead of the variances that are too small and cause a singular covariance. The second method is sparse matrix transform (SMT) that performs a set of K-givens rotations to estimate the covariance matrix. We will compare results from target acquisition that are based on both of these methods. An improvement for the SMT algorithm is suggested.

1. Introduction

The most widely used algorithms for target detection are traditionally based on the covariance matrix [1]. This matrix estimates the direction and magnitude of the noise in an image. In the equation for a matched filter presented in [1] we have

$$R = t^T \Phi_G^{-1}(x - m), \tag{1}$$

x is the examined pixel, m is the estimate of that pixel based on the surroundings, Φ_G is the global covariance matrix, and t is the target signature. In words, we can say that our matched filter for target detection will detect the target in a particular pixel x if x is different than its surroundings $(x - m)$, unlike the noise (controlled by Φ_G^{-1}) and in the direction of the target. If the target signature is unknown, then the RX algorithm uses the target residual $(x - m)$ as its own match, that is,

$$R = (x - m)^T \Phi_G^{-1}(x - m), \tag{2}$$

Φ_G is traditionally calculated as follows:

$$\Phi_G = \frac{1}{N}\sum_{i=1}^{N}(x_i - m)(x_i - m)^T. \tag{3}$$

Although the equation is theoretically justified if the background is stationary, it is often used in cases where this is not true.

In target detection, the image is not normally statistically stationary; it will however have quasistationary "patches" which connect to each other at the edges. When one estimates the mean and covariance matrix of the background of a particular pixel, the local neighboring pixels will have provided a better estimate than the pixels of the entire image. In [2], we show that much better results can be obtained if one uses a "quasilocal covariance matrix" (QLRX). In general terms, it uses the eigenvectors of the overall global matrix, but the eigenvalues are taken locally. This tends to lower the matched filter scores at edges in the data (when the image is going from one stationary distribution to another), but allows for accurate detection in less noisy areas.

The overall question of using a covariance matrix from local areas in which not enough data is sparse is actually a well-studied issue in the literature. In particular, in [3], Theiler et al. consider the sparse matrix rotation method for determining a covariance matrix based on limited data. In this paper, it is our intention to compare the two methods both in terms of their detection ability and their overall efficiency.

2. Local Covariance Matrix

Assume we are given a dataset X which is composed of n pixels with p dimensions. Φ_G is the covariance matrix of this dataset. An SVD (singular value decomposition) can be used to decompose the global covariance matrix [2] into eigenvectors and eigenvalues; we will refer to this as PCA (principal component analysis) space.

To compare the covariance matrix based on the local area surrounding a pixel and a matrix based on all the available data (referred to as global), consider the statistics of the dataset $\widetilde{X} = E_G^T X_G$. Here E_G is the rotating matrix based on the global eigenvectors [2], and $\widetilde{X}$ is the dataset after rotation into the PCA subspace.

If $\widetilde{X}$ is based on all the pixels in the image, then the covariance matrix of $\widetilde{X}$ consists of a diagonal matrix with the global eigenvalues on the diagonal. However, if $\widetilde{X}$ only contains the local surroundings, then the values on the diagonal of the covariance matrix of $\widetilde{X}$ will represent the variances of the local data in the direction of the global eigenvectors.

Mathematically, using the dataset $\widetilde{X}$, for every pixel we calculate the local covariance $\widetilde{\Phi}_L$ from the nearest neighbors. The diagonal $\widetilde{D}_L$ of the local covariance matrix is the variance of the neighbors in PCA subspace as follows:

$$\widetilde{D}_L = \operatorname{diag}\left(E_G^T \Phi_L E_G\right) = \operatorname{diag}\left(\widetilde{\Phi}_L\right). \tag{4}$$

Since the local covariance is composed from a small number of samples, some of the variances may be inappropriately small or even cause a singular covariance. To avoid singularity, the variance matrix Λ_{QL} will be the maximum between the variances of the global (Λ_G) covariance and the variances of local covariance in the PCA subspace ($\widetilde{D}_L$) as follows:

$$\Lambda_{\text{QL}} = \max\left(\Lambda_G, \widetilde{D}_L\right). \tag{5}$$

In this way, if the local area of the pixel has a large variance in some bands, it will be whitened by the local variance; for the bands that have too small local variances that can even cause a singular covariance matrix, it will use the global variance. The quasilocal covariance will be

$$\Phi_{\text{QL}} = E_G \Lambda_{\text{QL}} E_G^T. \tag{6}$$

In the PCA subspace it will simply be:

$$\widetilde{\Phi}_{\text{QL}} = \Lambda_{\text{QL}}. \tag{7}$$

If we will calculate the RX in the PCA subspace, we will need fewer rotations, we will rotate only once all the data to the PCA subspace, and then we will get:

$$\text{QLRX} = \left(\widetilde{x} - \widetilde{m}_L\right)^T \Lambda_{\text{QL}}^{-1} \left(\widetilde{x} - \widetilde{m}_L\right) = \sum_{i=1}^{p} \frac{\left(\widetilde{x}_i - \widetilde{m}_{Li}\right)^2}{\lambda_i}. \tag{8}$$

$\widetilde{m}_L$—the mean of the selected surrounding pixels in PCA subspace.

For subpixel targets, previous work [1] shows that it will be better to use the mean of 8 neighbors $\widetilde{m}_g$. This can be done assuming that the target does not affect the surroundings; if we fear that the target has entered the surrounding areas, then we will ignore those pixels and only use external pixels to them for our estimate.

In this method, we use sparse matrix rotations to find the nearest covariance matrix to the original one which is still nonsingular. We can use SVD to decompose the local covariance as follows:

$$\Phi_L = E_L \Lambda_L E_L^T, \quad \Phi_L \in R^{p \times p}. \tag{9}$$

Based on the fact that every eigenvectors matrix E (or any unitary matrix) can be extracted from a product of K spare orthonormal rotation matrix [3] we can write

$$E_L = \prod_{k=K-1}^{0} E_k = E_{k-1} E_{k-2} \cdots E_0. \tag{10}$$

Every rotation matrix E_k is a Givens rotation operating on coordinates indices (i_k, j_k); the rotation will be on the surface that contains the vectors i_k, j_k as follows:

$$E_k = \left(\begin{array}{ccccccc} & & i_k & & j_k & & \\ 1 & & & & & & \\ & \cdot & & & & & \\ & & \cos(\theta_k) & \cdot\cdot & \sin(\theta_k) & & \\ & & \vdots & \cdot & \vdots & & \\ & & -\sin(\theta_k) & & \cos(\theta_k) & & \\ & & & & & \cdot & \\ & & & & & & 1 \end{array}\right). \tag{11}$$

With $K = \left(\begin{smallmatrix} \text{N} \\ 2 \end{smallmatrix}\right)$ rotation we can get from the identity matrix to any rotation matrix.

The concept of SMT is to start from the identity matrix, rotate every time two axes in the direction of the axis of the eigenvectors of the local covariance matrix, and to stop the rotations when it gives the best fit without becoming singular.

In other words, from a first set of data we can determine the correlations between the variables. If we did all the possible rotations, we would have diagonalized the matrix, but we cannot do this since we do not have enough data to simultaneously find all the local eigenvectors. Instead, we do these rotations on the most correlated ones, testing our new matrix by the degree that it provides good results on a second dataset. When our correction to the second dataset fails, we stop the diagonalizing procedure.

Table 1: Datasets information—OBP1 and OBP2 are parts of OBP.

Name	Site	Sensor name	No. bands	Waveband (μm)	Spat. Res. (m)	Scene description
OBP	Oberpfaffenhofen (Ge)	Hymap	126	0.44–2.45	4	Airfield with agricultural area around
OBP1	Oberpfaffenhofen (Ge)	Hymap	126	0.44–2.45	4	Agricultural area
OBP2	Oberpfaffenhofen (Ge)	Hymap	126	0.44–2.45	4	Agricultural area

Mathematically, the rotation matrix T will be all the selected rotations combined:

$$T = \prod_{k=K-1}^{0} E_k = E_{k-1}E_{k-2}\cdots E_0, \quad \text{when } K < \binom{N}{2}. \tag{12}$$

The variances will be the variances of the local covariance matrix in the direction of the rotation matrix $T, \Lambda_{\text{SMT}} = \text{diag}(T'\Phi_L T)$ and the inversed covariance matrix will be $\Phi_{\text{SMT}}^{-1} = T\Lambda_{\text{SMT}}^{-1} = T'$. to decide what rotation matrix is best we use the maximum likelihood covariance estimation and "leave-third-out" cross-validation. (Note that the use of leave-one-out cross-validation will give better results but will cost much more in computational efforts).

We divide the group of pixels into three groups. We take one third to be the tested pixels $Y \in R^{(n/3)\times p}$, and we use the other two thirds to make the approximation of the covariance Φ_{SMT}. After every rotation we calculate the likelihood of covariance Φ_{SMT} to describe correctly the group Y. We have processed to data to make sure that Y is zero mean as follows:

$$P_{\Phi_{\text{SMT}}}(Y) = \frac{1}{(2\pi)^{p/2}\left|\Phi_{\text{SMT}}\right|^{1/2}} \exp\left\{-\frac{1}{2}\operatorname{tr}\left\{Y^T\Phi_{\text{SMT}}^{-1}Y\right\}\right\}. \tag{13}$$

We do this three times, each time another third is being taken out as the test data Y; after combining the results of the three tests, we find the number of rotations that gives the best result (based on the highest value of $P(Y)$); we then use the full set and this number of rotations to get the final approximation of the covariance matrix (see Figure 1).

To select every time the rotations that will make the biggest improvement, we perform greedy minimization, that is, always choosing the next rotation that will contribute most to reduce the correlation between data along the axis of the matrix as follows:

$$(i_k, j_k) \longleftarrow \arg\max_{i,j} \frac{S_{ij}^2}{S_{ii}S_{jj}}, \tag{14}$$

S is the current covariance matrix, (i, j) are indices of two rows in the matrix, and S_{ij}, S_{ii}, S_{jj} are the members in the matrix with those indices.

After we calculate the covariance matrix for SMT, we can use it for anomaly detection:

$$\text{RX}_{\text{SMT}} = (x - m_L)^T\Phi_{\text{SMT}}^{-1}(x - m_L), \tag{15}$$

x—the tested pixel.

m_L—the mean of the selected shrouding pixels in PCA subspace.

For subpixel targets as we stated previously, it will be better to use the mean of 8 neighbors m_8.

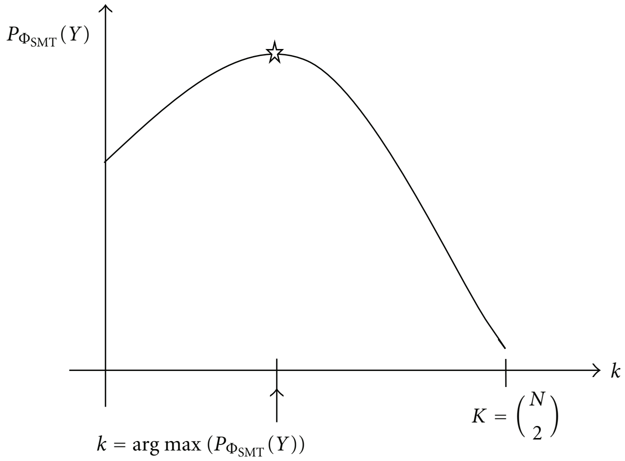


Figure 1: The probability that Φ_{SMT} describes Y correctly after k rotations; the k that will be chosen is the one that gives the maximum probability.

3. Dataset

Two datasets were used (Figure 2); a description of their origin can be found in Table 1 and in greater detail in [4].

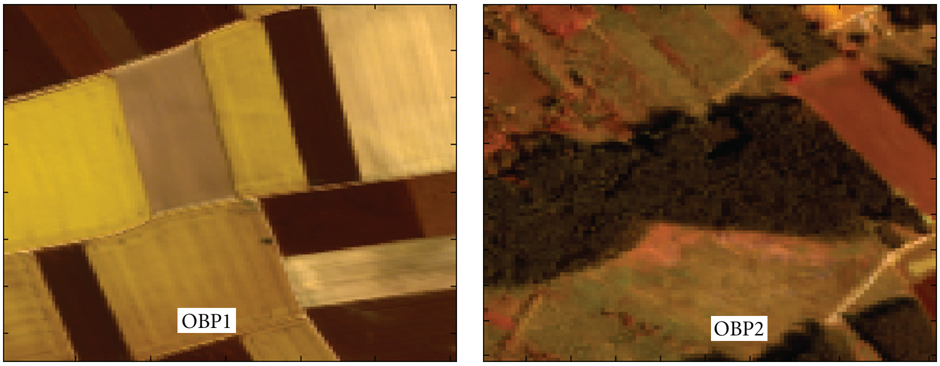
The two data cubes (OBP1 and OBP2) are real data from the Hymap sensor in which anomalies were inserted artificially by linearly mixing the spectra of a green paint pixel with the original background pixel. For display purposes, in Figure 2, images with full-pixel paint spectra are shown. For the evaluation of anomaly detection results, images with a mixing ratio of 0.33 ($P = 33\%$) were used.

4. Results

We now wish to compare the SMT and QLRX algorithms. We will perform RX anomaly detection (2) using the covariance matrices given by each of the algorithms.

Since the dataset being used contains implanted subpixel targets without any danger of overlap into neighboring pixels, the mean in the calculation of (2) was always the mean of the eight nearest neighbors. However, we must consider the correct neighborhood for the calculation of the covariance matrix for SMT that provides the best results.

The first test was done using only the nearest 8 neighbors for the approximation of the covariance; in this test, it is very easy to see that QLRX results are superior to the SMT results. The ROC curves are given in Figure 3. We assume that the area of the target (and of any examined pixel) consists of the square region of dimension OWS by OWS (outer window). The target area itself has area GWS by GWS (guard window);



(a)

(b)

Figure 2: RGB composite of the original data cubes. From left to right: OBP1, OBP2.

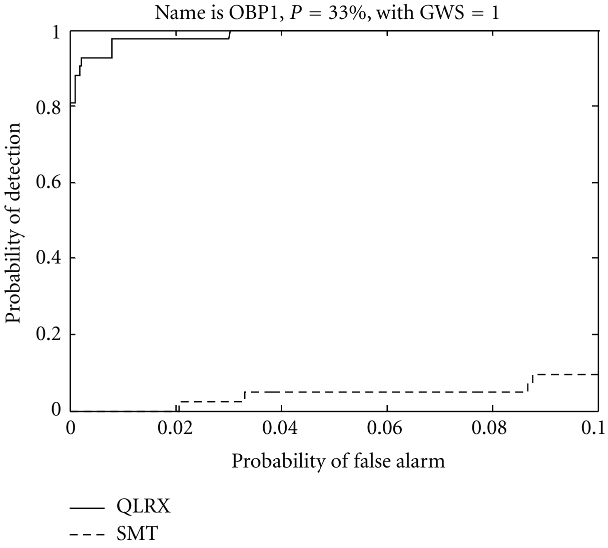


Figure 3: Results of RX algorithm using QLRX and SMT on dataset OBP1 with the stated OWS = 3 and GWS = 1.

for subpixel targets GWS will equal 1. Thus the neighboring pixels are those pixels which are located in the square set of pixels in the area OWS by OWS not contained in the inner GWS by GWS matrix.

In this picture is the result for $P = 33\%$, but the tests for 10, 25, 50, and 100 percents gave similar results.

Results from the OBP2 dataset were comparable.

When the OWS is larger, the results of the SMT improve dramatically.

For the dataset OBP1 we can see a large improvement as OWS increases. For this dataset QLRX gives better results, especially in the low CFAR (constant false alarm rate).

For the dataset OBP2, the differences between QLRX and SMT are reduced but still QLRX performs better in the low CFAR regime (see Figure 5).

Similar results were received for the cases in which 10, 25, 50, or 100 percent of the target is in the pixel.

The SMT has two main difficulties: first, the algorithm calculates a new covariance matrix at every point. This calculation needs a sequential set of rotations based on the

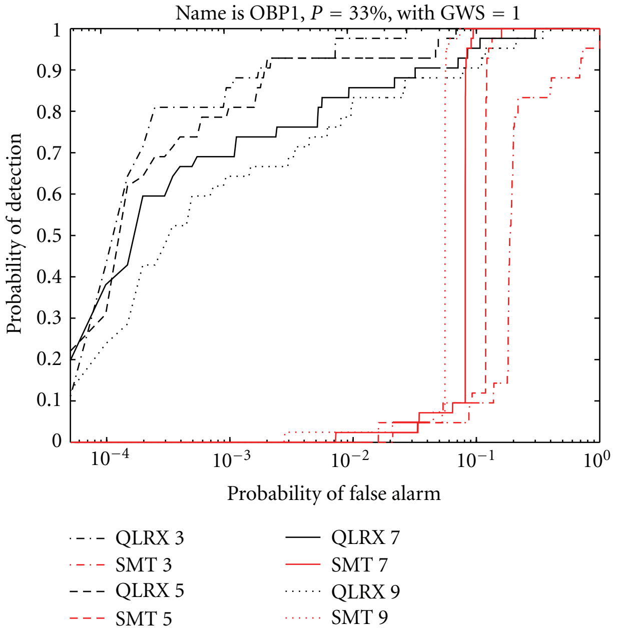


Figure 4: OBP1 results with OWS given by the number in the legend.

training set followed by evaluations of the test set. Both sets are taken from the pixel surrounding only, so none of the information outside the selected group is used in the calculation. In QLRX, the eigenvectors are the same for all points (the eigenvectors of the global covariance). All that we need to do is measure the variance in the local area in the spectral direction of the eigenvalues and calculate the new covariance matrix. Second, the calculation of the SMT itself is highly dependent on the size of the “local” area. While a larger area improves the results, it also increases the time for calculation (see Table 2).

5. Improvements for SMT

A small change in the published method for doing SMT could lead to a large improvement.

Table 2: This table shows the time it took to complete the calculation of a dataset.

“Name”	“OWS”	“GWS”	“QLRX” time in seconds	“SMT” time in seconds	Time ratio
“OP1_T1_S10”	9	1	26	3845	148
“OP1_T1_S33”	9	1	26	3794	146
“OP1_T1_S100”	9	1	25	3820	153
“OP1_T1_S10”	7	1	25	2781	111
“OP1_T1_S33”	7	1	24	2769	115
“OP1_T1_S100”	7	1	24	2770	115
“OP1_T1_S10”	5	1	23	1965	85
“OP1_T1_S33”	5	1	23	1969	86
“OP1_T1_S100”	5	1	23	1965	85
“OP1_T1_S10”	3	1	23	1195	52
“OP1_T1_S33”	3	1	22	1187	54
“OP1_T1_S100”	3	1	23	1201	52
“OP2_T1_S10”	9	1	14	2135	153
“OP2_T1_S33”	9	1	14	2105	150
“OP2_T1_S100”	9	1	14	2095	150
“OP2_T1_S10”	7	1	14	1480	106
“OP2_T1_S33”	7	1	14	1481	106
“OP2_T1_S100”	7	1	14	1464	105
“OP2_T1_S10”	5	1	13	961	74
“OP2_T1_S33”	5	1	14	962	69
“OP2_T1_S100”	5	1	13	960	74
“OP2_T1_S10”	3	1	12	556	46
“OP2_T1_S33”	3	1	13	555	43
“OP2_T1_S100”	3	1	14	552	39

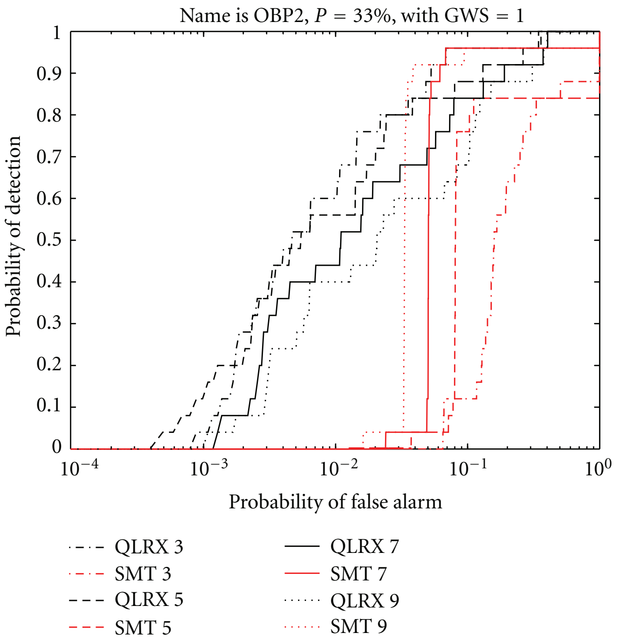


Figure 5: Similar to Figure 4 for the OBP2 dataset.

In the original algorithm, the initial assumed axes of the covariance matrix are in the direction of the original dataset. Then the axes are rotated in pairs into the directions of the “local eigenvectors” to create new covariance matrices.

When we stop, some of the axes will be almost the same direction as the local covariance eigenvectors and some will be closer to the direction of the original axes.

Now since the original directions were random, that is, not related to the correlations between the axes, it is easy to see that there is no reason that this should be optimum. In particular, would it not make more sense to start from the global eigenvectors and rotate into the local ones? Another benefit we will get from this approach is that we will start the rotation from a condition that most probably will be closer to the optimum point (see Figure 6); within fewer rotations, we will get to the maximum likelihood. We will call this new algorithm SMT PCA.

For the OBP1 dataset (Figure 7), the result after starting with the subspace based on the global eigenvectors are better than QLRX when OWS is big (7,9). SMT-PCA gives better results than SMT for any OWS.

For the OBP2 dataset (Figure 8), the result after starting with the subspace based on the global eigenvectors are better than QLRX when OWS is big (7,9). SMT-PCA gives better results from SMT for any OWS (for OWS = 3 SMT and SMT-PCA almost the same).

Examining the number of rotations needed in the SMT and in the SMT-PCA (see Table 3).

Table 3: This table shows the time it took to complete the calculation of a dataset.

"Name"	"OWS"	"GWS"	Original SMT number of rotations	SMT after PCA number of rotations	Rotations number ratio
"OP1_T1_S10"	9	1	3845	1851	2.1
"OP1_T1_S33"	9	1	3794	1831	2.1
"OP1_T1_S100"	9	1	3820	1806	2.1
"OP1_T1_S10"	7	1	2781	1508	1.8
"OP1_T1_S33"	7	1	2769	1498	1.8
"OP1_T1_S100"	7	1	2770	1457	1.9
"OP1_T1_S10"	5	1	1965	1177	1.7
"OP1_T1_S33"	5	1	1969	1203	1.6
"OP1_T1_S100"	5	1	1965	1122	1.8
"OP1_T1_S10"	3	1	1195	442	2.7
"OP1_T1_S33"	3	1	1187	407	2.9
"OP1_T1_S100"	3	1	1201	421	2.9
"OP2_T1_S10"	9	1	2135	839	2.5
"OP2_T1_S33"	9	1	2105	840	2.5
"OP2_T1_S100"	9	1	2095	827	2.5
"OP2_T1_S10"	7	1	1480	647	2.3
"OP2_T1_S33"	7	1	1481	641	2.3
"OP2_T1_S100"	7	1	1464	635	2.3
"OP2_T1_S10"	5	1	961	440	2.2
"OP2_T1_S33"	5	1	962	435	2.2
"OP2_T1_S100"	5	1	960	431	2.2
"OP2_T1_S10"	3	1	556	203	2.7
"OP2_T1_S33"	3	1	555	205	2.7
"OP2_T1_S100"	3	1	552	206	2.7

The smaller the number of rotations, the less time needed for the calculation.

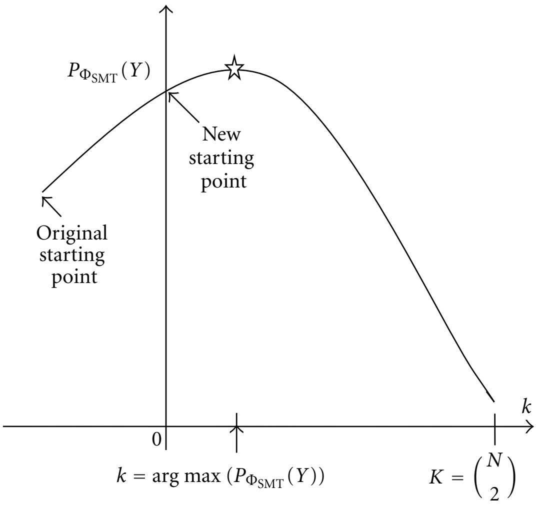

$P_{\Phi_{\mathrm{SMT}}}(Y)$
New starting point
Original starting point
0
k
$k = \arg\max\,(P_{\Phi_{\mathrm{SMT}}}(Y))$
$K = \binom{N}{2}$

Figure 6: When starting from PCA subspace, we will start from a closer point to the maximum so we need fewer rotations; the delta in k between the original location to the current one is the rotations done by transforming to the PCA subspace.

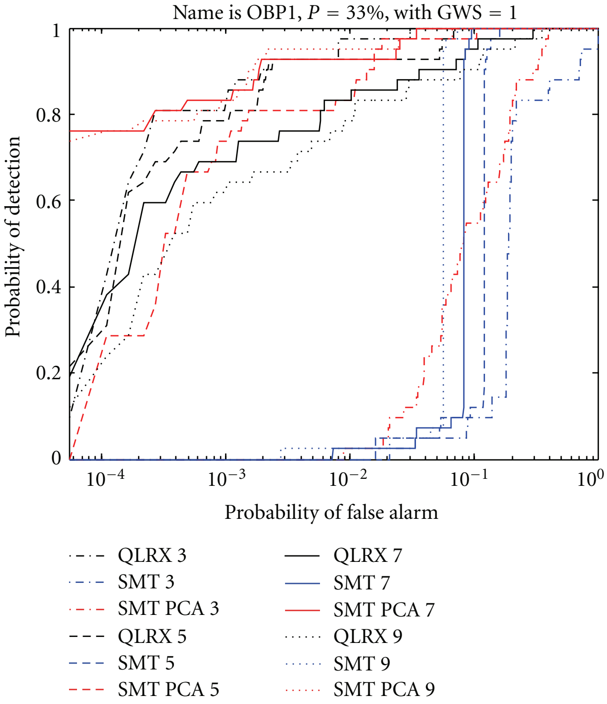


Figure 7: OBP1 results with OWS given by the number and the legend.

6. Conclusions

As a preliminary to our conclusions, please note that when we discuss using a small or large number of pixels, that in all

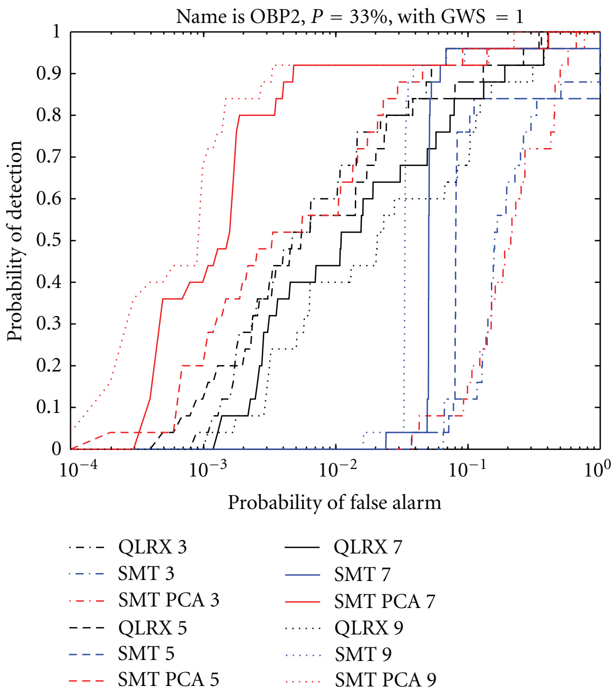


FIGURE 8: Similar to Figure 7 for the HAR dataset.

cases the number of pixels used is less than the number of spectral bands.

Two methods in this paper have been considered for dealing with possibly singular covariance matrices. In the first (QLRX), we use global eigenvectors and local eigenvalues as an approximation of the inverse covariance matrix. In the second (SMT), we use an iterative process to slowly "twist" our axes to come closer to those determined by the data.

In our two datasets, we found that if a small area was used for estimating the background, the QLRX algorithm was superior. For large areas of background, QLRX remains superior, although SMT greatly improves as follows:

$$\frac{\text{Number of pixels}}{\text{Number of bands}} = \frac{p}{n} \in (0,1), \qquad (16)$$

(i) the calculation time of QLRX is much smaller (two orders of magnitude) than both SMT and SMT-PCA,

(ii) the calculation time of SMT PCA is less than the calculation time of the original SMT by about a factor of two,

(iii) SMT-PCA and QLRX performance are better than those of SMT for any number of pixels,

(iv) for a small number of pixels ($p/n \leq 0.1$), the QLRX performance is better than that of SMT-PCA,

(v) for a large number of pixels ($0.2 \leq p/n < 0.1$), the performance of SMT-PCA is better than that of QLRX.

Acknowledgments

The authors gratefully recognize the partial support for this work from the Paul Ivanier Center for Robotics Research and Production Management, Beer-Sheva, Israel. The test images are part of the Oberpfaffenhofen HyMAP scene, collected in 2004, during an airborne campaign sponsored by the Belgian Science Policy Office (BelSPO). Flights were operated by the German Aerospace Center (DLR).

References

[1] C. E. Caefer, M. S. Stefanou, E. D. Nielsen, A. P. Rizzuto, O. Raviv, and S. R. Rotman, "Analysis of false alarm distributions in the development and evaluation of hyperspectral point target detection algorithms," *Optical Engineering*, vol. 46, no. 7, Article ID 076402, 2007.

[2] C. E. Caefer, J. Silverman, O. Orthal, D. Antonelli, Y. Sharoni, and S. R. Rotman, "Improved covariance matrices for point target detection in hyperspectral data," *Optical Engineering*, vol. 47, no. 7, Article ID 076402, 2008.

[3] J. Theiler, G. Cao, L. R. Bachega, and C. A. Bouman, "Sparse matrix transform for hyperspectral image processing," *IEEE Journal on Selected Topics in Signal Processing*, vol. 5, no. 3, pp. 424–437, 2011.

[4] D. Borghys and C. Perneel, "Study of the influence of preprocessing on local statistics-based anomaly detector results," in *Proceedings of the Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS '10)*, pp. 1–4, June 2010.