ARMY RESEARCH LABORATORY

Innovative Statistical Inference for Anomaly Detection in Hyperspectral Imagery

by Dalton Rosario

ARL-TR-3339 **September 2004**

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

Army Research Laboratory
Adelphi, MD 20783-1197

ARL-TR-3339 **September 2004**

Innovative Statistical Inference for Anomaly Detection in Hyperspectral Imagery

Dalton Rosario
Sensors and Electron Devices Directorate, ARL

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
September 2004	Final	October 2003 to September 2004

4. TITLE AND SUBTITLE
Innovative Statistical Inference for Anomaly Detection in Hyperspectral Imagery

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Dalton Rosario

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U.S. Army Research Laboratory
ATTN: AMSRD-ARL-SE-SE
2800 Powder Mill Road
Adelphi, MD 20783-1197

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TR-3339

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Research Laboratory
2800 Powder Mill Road
Adelphi, MD 20783-1197

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
A statistical motivated idea is proposed and its application to hyperspectral imagery is presented, as a viable alternative to testing a two-sample hypothesis using conventional methods. This idea led to the design of two novel algorithms for object detection. The first algorithm, referred to as *semiparametric* (*SemiP*), is based on some of the advances made on semiparametric inference. A logistic model, based on case-control data, and its maximum likelihood method are presented, along with the analysis of its asymptotic behavior. The second algorithm, referred to as an *approximation* to semiparametric (*AsemiP*), is based on fundamental theorems from large sample theory and is designed to approximate the performance properties of the *SemiP* algorithm. Both algorithms have a remarkable ability to accentuate local anomalies in a scene. The *AsemiP* algorithm is particularly more appealing, as it replaces complicated *SemiP's* equations with simpler ones describing the same phenomenon. Experimental results using real hyperspectral data are presented to illustrate the effectiveness of both algorithms.

15. SUBJECT TERMS
Hyperspectral anomaly detection, large sample theory

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dalton Rosaro
Unclassified	Unclassified	Unclassified	UL	40	19b. TELEPHONE NUMBER (*Include area code*) (301) 394-4235

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Content

List of Figures

Acknowledgment

I would like to take this opportunity to thank Prof. Rama Chellappa (UMD, U. of Md. at College Park), Prof. Benjamin Kedem (UMD), Prof. Eric Slud (UMD), and Dr. Nasser Nasrabadi (ARL) for occasional live discussions during the course of this work, and Dr. Patti Gillespie (ARL) for her encouragement and unconditional support.*

* This work was supported in pat by the ARL Director's Research Initiative award FY04-SED-29.

INTENTIONALLY LEFT BLANK

1. Introduction

1.1 Objective

This report focuses on the design of a new family of local anomaly detectors for hyperspectral sensor imagery (HSI). These detectors employ an indirect sample comparison via the mathematics of *semiparametric* statistics and large sample theory to test the likelihood that local HSI random samples belong to the same population, or class. The employed notion is not based on a physical motivation but on a statistical one, and it is proposed as a viable alternative to testing a two-sample hypothesis using conventional methods. The aim is to achieve a significant reduction of meaningless detections via unsupervised learning methods, while maintaining a high probability of meaningful detections. The notion I seek to implement is relatively simple, but some of the mathematics presented in this report are nontrivial, as *Normality* is not assumed in the models.

1.2 Survey of Prior Art

Hyperspectral sensors are passive sensors that simultaneously record images for many contiguous and narrowly spaced regions of the electromagnetic spectrum. A data cube is created from these images in which each image corresponds to the same ground scene and contains both spatial and spectral information about objects and backgrounds in the scene. These sensors employ several bands and have been used in various fields including urban planning, mapping, and military surveillance. Good references to some of these topics in the context of remote sensing and descriptions of the most popular sensors since the mid 1960's are included in the reference section.[1,2,3]

With the introduction of sensors capable of high spatial and spectral resolution, there has been an increasing interest in using spectral imagery to detect objects or features of interest (targets). However, detection algorithms that presume a target signature are subject to signal mismatch losses because of the complications of converting sensor data into material spectra.[4] Target matching approaches are further complicated by the large number of possible targets and uncertainty as to the reflectance/emission spectra of these objects. For example, the surface of a target 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 targets and complications of

[1] Schowengerdt, 1997.
[2] Coompbell, 1996.
[3] Lillesand, 1994.
[4] Schott, 1977.

atmospheric compensation 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 or target subspaces. Quite often objects consisting of unusual materials are detected as local anomalies in a scene; hence, anomaly detection will be used interchangeably in this paper with object detection.

Anomalies are defined with reference to a model of the background. 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 and from the overall scene, respectively. Both approaches have their merits. The local spectral anomaly detector is susceptible to false alarms that are isolated spectral anomalies. An algorithm commonly referred to as RX algorithm,[5] for instance, has become a benchmark for multispectral data, based on this principle. The RX algorithm is a maximum likelihood anomaly detection procedure that simplifies the clutter to being spatially white. Researchers have also adapted some classic approaches[6] (e.g., Fisher's Linear Discriminant [FLD], Principal Component Analysis [PCA]) in the same spirit of the RX algorithm to anomaly detection in hyperspectral sensor imagery. Global spectral anomaly detection algorithms, on the other hand, are not susceptible to this type of clutter-generated false alarms. However, a global anomaly detector will not find an isolated target in the open if the signature is similar to that of previously classified background material. Examples of global detectors are the global normal mixture mode[7] and the global linear mixture model.[8] A comprehensive comparative analysis among different approaches for HSI anomaly detection can be found in.[9,10,11,12]

Earlier laboratory experiments using SAR (synthetic aperture radar) imagery[13] produced compelling evidences to suggest that using conventional approaches to compare sample pairs, i.e., two-sample hypothesis tests treating the two samples individually, promotes an intolerable high number of meaningless detections (false alarms [FA]) in areas characterized by region discontinuities in the imagery. At that time, I proposed a post-processing method from mathematical morphology to account for those observations and adapted it to a SAR target-detection system.[13,14] The approach produced some surprisingly good results using real SAR data.

[5] Yu, 1997
[6] Kwon, 2003.
[7] Stein, 2002
[8] Stein, 2002.
[9] Manolakis, 2002.
[10] Crist, 1999.
[11] Haskett, 1999.
[12] Grossmann, 1998.
[13] Rosario, 1999.
[14] Rosario, 2000.

The novelty in this paper is the mathematical means that is proposed to address HSI anomaly detection. The main contribution of this report is threefold: (i) a recently proposed local anomaly detector[15] shall be discussed for the first time using extended details; I shall describe a suitable mathematical model[16,17,18,19,20] that elegantly materializes a combining idea and shall study the model's maximum likelihood method and its asymptotic behavior; I shall design an effective local anomaly detector based on the model's asymptotic behavior, which for convenience shall be named the *semiparametric (SemiP)* algorithm; (ii) a second anomaly detector shall be proposed to the community, an *approximation* to the semiparametric *(AsemiP)* algorithm, which may be used to replace the complicated equations of the first model's solution with simpler equations—yet describing the same phenomenon; I shall state a proposition of the second model and prove its statement. Derivation of the *AsemiP* algorithm is motivated by the *SemiP*'s output properties, not by the semiparametric model itself—although, its derivation is also based on approximation theorems of mathematical statistics; and (iii) in order to promote the use of models whose mathematics are based on the statistical assumption of *independent, identically distributed* random samples, an *inside/outside window* mechanism shall be introduced aimed at transforming local HSI information into independent sample pairs. Comparative results are also presented between *SemiP*, *AsemiP* and other approaches commonly used with hyperspectral data.

2. Approach

2.1 Problem Formulation

Figure 1 shows a hyperspectral scene that consists of 14 ground vehicles parked across a grassy area along a *treeline*. For surveillance applications in similar scenes, human analysts do well quickly ignoring most of the imagery and concentrating their attentions to those vehicles and their shadows. Humans, of course, use both global and local information to focus their attention to meaningful objects in the scene, a capability that can only be reproduced by applying, for instance, layers of unsupervised learning methods complementing each other to perform this single task. For example, a suite of algorithms that includes an edge detector, an edge elongation, a clustering method, and a morphological size test would probably reproduce the humans' performance in such a scene, but with a huge cost: computational time.

[15] Rosario, 2004
[16] Qin, 1997.
[17] Fokianos, 2001
[18] Anderson, 1972.
[19] Prentice, 1979.
[20] Cox, 1996.

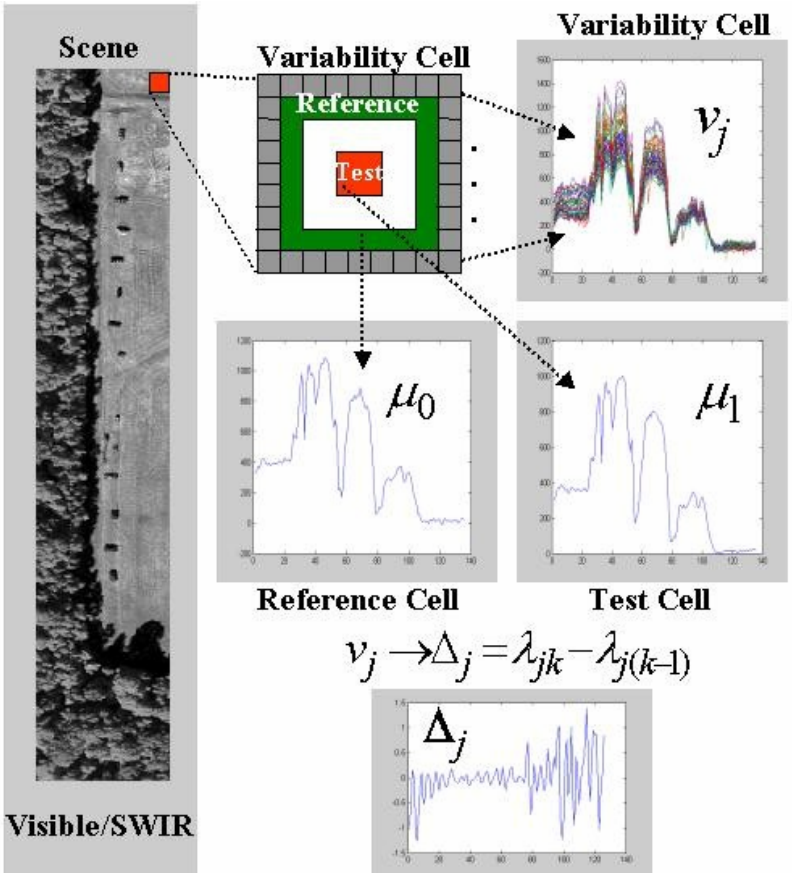

Figure 1. Nonhomogeneous, multicomponent scene from the hyperspectral digital imagery collection experiment (HYDICE). (A pixel in an HYDICE imagery is represented by a vector.) Typically, local anomaly detectors produce an intolerable high number of false alarms (non-anomalies) in similar scenes; local region discontinuities degrade detectors' performances. The sampling mechanism is discussed in the text.

My interest is to approach humans' performance using a single unsupervised learning algorithm that functions as a local anomaly detector. Figure 1 illustrates the sampling mechanism I propose to transform local imagery information into sample pairs for our statistical models. It introduces three window cells from which samples will be drawn from the data. These windows are referred to as: *test cell*, *reference cell*, and *variability cell*. Information between the *variability* and *reference cells* will be used to form a control or reference feature vector, and information between the *variability* and *test cells* will be used to form an unknown test feature vector.

The *test cell* provides a spectral sample average (μ_1) from a (w x w) window; the *reference cell* provides a spectral sample average (μ_0) from M vectors surrounding a guard region, i.e., a *blind* area between test and reference cells to account for larger than (w x w) targets; and the *variability cell* provides J individual spectral vectors (v_j) each consisting of $k = 1,\ldots,K$ spectral

responses (λ_{jk}) for K distinct wavelengths in the visible to SWIR (shortwave infrared) region of the electromagnetic spectrum, i.e., region from *0.4* μm to *2.4* μm.

Hyperspectral data have highly correlated—hence, dependent—spatial and spectral clutter, so to promote statistical independence, given that we will make this assumption in our models, I propose to apply a high-pass (HP) filter in the spectral domain, thus transforming v_j into Δ_j (see figure 1), and then use Δ_j to compute a feature that promotes spatial independence. The feature is known as spectral angle mapper (SAM),[21] which in essence computes the angle between two vectors, or

$$x_{ij} = \frac{180}{\pi} \arccos\left(\frac{\Delta_j^t \Delta_{\mu_i}}{\|\Delta_j\| \|\Delta_{\mu_i}\|} \right), \tag{1}$$

where $\Delta_j = \lambda_{jk} - \lambda_{j(k-1)}$ (Δ_j is a [*K-1*] x *1* vector); $\Delta_{\mu i} = \lambda_{\mu i,k} - \lambda_{\mu i,(k-1)}$ (using spectral components from vectors μ_0 and μ_1), $x_i = [x_{ij}]$; $i=0,1$; $j=1,\ldots,J$ (J is the total number of vectors in the variability cell); x_{ij} range from *0* to *90* degrees (*0* representing minimum class difference between reference and test samples and *90* representing the maximum class difference between these samples); the operator $\| z \|$ denotes the squared root of $z^t z$; and $[*]^t$ denotes the vector transpose operator.

Let x_0 denote the reference feature vector, x_1 the test feature vector, and let both vectors be distributed (~) by unknown joint distributions f_0 and f_1, respectively, or

$$\begin{aligned} x_1 &= [x_{11}, \cdots, x_{1n_1}] \sim f_1(x) \\ x_0 &= [x_{01}, \cdots, x_{0n_0}] \sim f_0(x) \end{aligned} \tag{2}$$

where, $n_0 = n_1 = J$ in this particular implementation.

The window cells are expected to systematically move throughout the imagery and at each location this question will be posed: Do x_0 and x_1 belong to the same population, or class, in the feature space? If the answer is *no*, the test sample will be labeled as an anomaly with respect to its surroundings at that location.

A conventional two-sample hypothesis test would work very well if samples x_0 and x_1 do belong to distinct classes C_0 and C_1, or to one of these classes. Problems occur, however, when one of the samples (e.g., x_0) belongs to a composite class consisting of both classes C_0 and C_1, denote $x_0(C_0C_1)$, and then it is compared to $x_1(C_1)$. In those cases, standard statistical tests may reject the hypothesis that $x_0(C_0C_1)$ and $x_1(C_1)$ belong to the same class. This rejection—correct as it may seem—is arguably the most dominant driving force affecting the number of FA produced by most—if not all—local anomaly detectors using sensor imagery. The reason is that region discontinuities (e.g., boundaries between tree clusters and their shadows) are abundant in sensor imagery, and they are not taken into account in conventional statistical models.

[21] Kelly, April 1989.

2.2 Statistical-Motivated Idea

I propose an indirect comparison approach to circumvent the problem discussed in section 2.1. The approach is not based on a physical motivation but on a statistical one. The key is not to compare samples x_0 and x_1 directly, but to make that comparison indirectly by constructing a new sample z, consisting of both x_0 and x_1, and then by comparing z in some form to either x_0 or x_1. To clarify this notion, consider the following case studies when comparing x_0 and x_1: (1) samples belong to distinct classes, $x_0(C_0)$ and $x_1(C_1)$; (2) samples belong to a single class, $x_0(C_0)$ and $x_1(C_0)$; and (3) one of the samples holds information from two classes while the other sample belong to one of these classes, $x_0(C_0C_1)$ and $x_1(C_1)$. Now consider the following, where U denotes the union of samples:

	Case 1	Plausible Result
Conventional	$x_0(C_0) =? x_1(C_1)$	*No*
Proposed	$z(C_0C_1) = \{x_0(C_0) \, U \, x_1(C_1)\} =? x_1(C_1)$	*No*

	Case 2	Plausible Result
Conventional	$x_0(C_0) =? x_1(C_0)$	*Yes*
Proposed	$z(C_0C_0) = \{x_0(C_0) \, U \, x_1(C_0)\} =? x_1(C_0)$	*Yes*

	Case 3	Plausible Result
Conventional	$x_0(C_0C_1) =? x_1(C_1)$	*No*
Proposed	$z(C_0C_1C_1) = \{x_0(C_0C_1) \, U \, x_1(C_1)\} =? x_1(C_1)$	*Maybe*

It is plausible that the proposed comparison approach would yield the same results produced by conventional methods for case studies *1* and *2*, where *No* and *Yes* denote *anomalies* and *non-anomalies*, respectively. In particular, z in *Case 1* would have been labeled as a strong anomaly in respect to x_1 for the same reason a conventional test would have rejected the hypothesis that a composite sample (e.g., tree and shadow) belongs to the same class of a relatively pure sample (e.g., shadow). In *Case 3*, however, the proposed and conventional approaches would probably disagree in the intensity of their results because $z(C_0C_1C_1)$ shows more evidence of C_1 than does

$x_0(C_0C_1)$. Sample $x_1(C_1)$ would seem statistically closer to $z(C_0C_1C_1)$ than to $x_0(C_0C_1)$, and that difference would help to interpret $x_1(C_1)$ as a *soft-anomaly* (labeled in *Case 3* as *Maybe*).

My goal is to propose statistical models that are capable of accentuating, significantly, local *anomalies* (*No*) from *soft-anomalies* (*Maybe*) and *non-anomalies* (*Yes*). Such a capability would allow a detector to retain a high probability of correct detections while significantly reducing the number of nuisance detections. The first statistical model that I propose combines samples by relating in some form the probability distribution functions of x_0 and x_1. The model is discussed next.

2.3 Logistic Model

Let vectors x_k have their components independently, identically distributed (iid). Let x_0 be independent of x_1. And consider the following:

$$\begin{aligned} x_1 &= [x_{11}, \cdots, x_{1n_1}] \quad iid \sim g_1(x) \\ x_0 &= [x_{01}, \cdots, x_{0n_0}] \quad iid \sim g_0(x), \end{aligned} \tag{3}$$

$$\frac{g_1(x)}{g_0(x)} = \exp(\alpha + \beta h(x)), \tag{4}$$

where g_1 is regarded as an exponential distortion of g_0 and $h(x)$ is an arbitrary but known function of x. Note in (3)-(4) that no parametric assumption is given to g_0 and that g_1 depends only on the unknown parameters α and β, hence, justifying the model's name: semiparametric.

The rationale for proposing to use (4), as our baseline, is that many common distribution families can be expressed as a canonical exponential function. These families fall under a category of probability density functions called *exponential families*, which are known to have many nice mathematical and statistical properties. (Some of these properties are discussed, for instance, in.[22]) One of these mathematical properties, for example, is that an exponential-family distribution can always be expressed as a shift of another exponential-family distribution, as shown in (4). Reference[23] provides a good discussion on this topic.

Model (3)-(4) is based on case-control data and its mathematical development depends on some of the advances made on the theory of semiparametric inference. Semiparametric approaches are commonly used in analyzing binary data that arise in studying relationships between disease and

[22] Casella, 1990.
[23] Kay, 1987.

environment of genetic characteristics.[18,19,20] Equation 4 has its roots in the standard logistic function having the general form

$$P(z) = \frac{\exp(\gamma + \beta z)}{1 + \exp(\gamma + \beta z)} \equiv \eta(z), \tag{5}$$

where γ is a scale parameter and β interpreted as a constant rate both defining a proportion *P,* which is dependent on a variable *z*. The logistic function was invented in the 19th century for the description of the growth of populations and the course of autocatalytic chemical reactions. Pierre-Francois Verhulst (1804-1849), a Belgian statistician, named[5] as the *logistic* function and published his suggestions between 1838 and 1847. For an elaborate historical account, see.[24]

Using the independence assumptions in model (3)-(4), with $h(x) = x$, the MLE (maximum likelihood estimate) of α and β can be attained via the likelihood function,

$$\begin{aligned} \zeta(\alpha, \beta, g_0) &= \prod_{i=1}^{n_0} g_0(x_{0i}) \prod_{j=1}^{n_1} \exp(\alpha + \beta x_{1j})\, g_0(x_{1j}) \\ &= \prod_{i=1}^{n=n_1+n_0} g_0(t_i) \prod_{j=1}^{n_1} \exp(\alpha + \beta x_{1j}). \end{aligned} \tag{6}$$

where, $n = n_0+n_1$ and the combined feature vector *t* is

$$t = [x_{11}, \cdots, x_{1n_1}, x_{01}, \cdots, x_{0n_0}] = [t_1, \cdots, t_n]. \tag{7}$$

Notice in (6) that the part involving $g_0(t_i)$ (reference distribution) reflects the combined-sample property of a model we sought, and the part involving the exponential distortion reflects only the property of the sample that is not the reference. Both properties fit well into the proposed framework of merging samples and then, in some form, comparing this combined structure with one of its original samples.

Notice also that g_0 in (6) is unknown, thus, deriving MLE of α and β via standard procedures cannot be attained. However, using profiling one can express g_0 in terms of α and β and then replace g_0 with its new representation back in (6). Using profiling, the method of Lagrange multiplier was proposed[18] to attain the maximization of ζ by fixing (α,β) and then maximizing ζ with respect to $g_0(t_i)$ for $i=1,\ldots,n$, subject to constraints

$$\begin{aligned} &\sum g_0(t_i) = 1, \quad g_0(t_i) \geq 0, \\ &\sum [\exp(\alpha + \beta t_i) - 1] g_0(t_i) = 0, \end{aligned} \tag{8}$$

[5] Yu, 1997.
[18] Anderson, 1972.
[19] Prentice, 1979.
[20] Cox, 1996.
[24] Cramer, 2002.

where the last constraint reflects the fact that $exp(\alpha+\beta x)g_0(x)$ is a distribution function. Following this approach, it can be shown[16] that the maximum value of ζ is attained at

$$g_0(t_i) = \frac{1}{n_0}\frac{1}{1+\rho\exp(\alpha+\beta t_i)}, \tag{9}$$

where $\rho=n_1/n_0$, and that ignoring a constant, the log-likelihood function is

$$l(\alpha,\beta) = \sum_{j=1}^{n_1}(\alpha+\beta x_{1i}) - \sum_{i=1}^{n}\log[1+\rho\exp(\alpha+\beta t_i)]. \tag{10}$$

A system of score equations that maximizes (10) over (α,β) is shown below,

$$\frac{\partial l(\alpha,\beta)}{\partial\alpha} = -\sum_{i=1}^{n}\frac{\rho\exp[\alpha+\beta t_i]}{1+\rho\exp[\alpha+\beta t_i]} + n_1 = 0 \tag{11}$$

$$\frac{\partial l(\alpha,\beta)}{\partial\beta} = -\sum_{i=1}^{n}\frac{t_i\rho\exp(\alpha+\beta t_i)}{1+\rho\exp(\alpha+\beta t_i)} + \sum_{j=1}^{n_1}x_j = 0. \tag{12}$$

Let (α^*,β^*) satisfy (11)-(12), then using (9) it can be shown[16] that the MLE of $g_0(x)$ is

$$\hat{g}_0(t_i) = \frac{1}{n_0}\frac{1}{1+\rho\exp(\alpha^*+\beta^* t_i)}. \tag{13}$$

2.4 SemiP Algorithm

In this subsection, I adapt the theory in section 2.3 to the framework of object (anomaly) detection. The mathematics for this adaptation is nontrivial, as we employ the asymptotic behavior of (α^*,β^*) under a model that does not assume *Normality*. We strive to reach a reasonable balance between showing the most important parts of the mathematical development and length of this report.

First, notice that for $g_1(x) = exp(\alpha+\beta x)g_0(x)$ to be a density, a hypothesis of $\beta = 0$ in (4) must imply $\alpha = 0$, as the term $exp(\alpha)$ functions as a normalizing factor so that g_1 integrates over x to a total mass unity. Second, notice that the hypothesis H_0: $\beta = 0$ (given that α must also be equal to zero) in (4) implies that a test population and a reference (control) population are equally distributed, namely, $g_1 = g_0$. With this logic, one can design an anomaly detector from the following composite hypothesis test:

$$\begin{aligned} H_0: &\quad \beta = 0 \quad (g_1 = g_0) \qquad \text{anomaly absent} \\ H_1: &\quad \beta \neq 0 \quad (g_1 \neq g_0) \qquad \text{anomaly present.} \end{aligned} \tag{14}$$

[16] Qin, 1997.

Under this test, local regions in the entire imagery would be individually tested to reject the null hypothesis (H_0) yielding in the process a binary surface of values of *1*, depicting a rejection of H_0, and values of *0*, depicting a non-rejection of H_0. An isolated object in a scene would be expected to produce a cluster of *1* values (anomalies) in the resulting binary surface. But to design the hypothesis test in (14), one must know the asymptotic behavior of the extremum estimator β^*, which one can verify (see Appendix A) that it converges to Normality, or

$$\sqrt{n}\beta^* \to N\left(0, \frac{\rho^{-1}(1+\rho)^2}{Var(t)}\right). \tag{15}$$

Squaring the left side of (15) and normalizing the result by the asymptotic variance, one can conclude (see convergence results, for instance, in[22]) that the resulting random variable

$$\chi = n\rho(1+\rho)^{-2}\beta^{*2}V^*(t) \to \chi_1^2 \tag{16}$$

converges to a chi square distribution with *1* degree of freedom, where $V^*(t)$ estimates $Var(t)$.

2.5 Approximating SemiP Performance: AsemiP Algorithm

In reference to (16), there are two major factors working in harmony and in complementary fashion to promote maximum separation between signal (anomalies) and noise (non-anomalies), they are: the squared value of β^* and the estimated combined variance, $V^*(t)$, which is also quadratic.

These factors work in the following way: When two samples from the same class are compared (i.e., is H_0: $\beta=0$ [$g_1=g_0$] true?), the term $(\beta^*)^2$ tends to approach zero very fast, specially for β^* values less than unity. On the other hand, if two samples from distinct classes are compared, the term $V^*(t)$ tends to a relatively high number, also very fast, asserting the fact that a combined sample vector t consists of components belonging to distinct populations.

Motivated by these properties, I shall state and prove an approximation algorithm based on large sample theory that replaces complicated *SemiP* equations with simpler ones describing the same phenomenon.

Proposition 1 (AsemiP Algorithm). Let

$$\begin{aligned} x_1 &= [x_{11}, \cdots, x_{1n_1}] \text{ be iid},\ E(x_{1i}) = \mu_1, Var(x_{1i}) = \sigma_1^2 < \infty; \\ x_0 &= [x_{01}, \cdots, x_{0n_0}] \text{ be iid},\ E(x_{0i}) = \mu_0,\ Var(x_{0i}) = \sigma_0^2 < \infty; \end{aligned} \tag{17}$$

assume that x_0 and x_1 are independent and that, for some x_0 and x_1, the combined vector t

$$t = [x_{11}, \cdots, x_{1n_1}, x_{01}, \cdots, x_{0n_0}] = [t_1, ..., t_{n=n_1+n_0}] \text{ is iid},\ Var(t) = \sigma^2 < \infty; \tag{18}$$

[22] Casella, 1990.

and define

$$\tilde{\beta} \equiv \mu_1 - \mu_0; \quad \zeta_1 \equiv \frac{\sigma^2}{\sigma_0^2}, \zeta_2 \equiv \frac{\sigma^2}{\sigma_1^2}. \tag{19}$$

Under some regularity conditions, if hypothesis H_0: $\left(\tilde{\beta} = 0; \zeta_1 = \zeta_2 = 1\right)$ is true and

$$\hat{\tilde{\beta}} = \bar{x}_1 - \bar{x}_0, \qquad \bar{x}_i = n_i^{-1} \sum_{k=1}^{n_i} x_{ik}, \quad i = 0,1; \tag{20}$$

$$\tilde{V}(t) = \sum_{i=1}^{n} (t_i - \bar{t})^2 \cdot g(x_1, x_0), \qquad n = n_1 + n_0, \ \ where \ \ \bar{t} = n^{-1} \sum_{i=1}^{n} t_i, \tag{21}$$

$$g(x_1, x_0) = \frac{(n-2)^2}{\left[\sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2 + \sum_{i=1}^{n_0} (x_{0i} - \bar{x}_0)^2 \right]^2}; \ and \quad \tilde{\rho} = \left(\frac{1}{n_1} + \frac{1}{n_0} \right)^{-1}; \tag{22}$$

where, $\hat{\tilde{\beta}}$ is an estimate of $\tilde{\beta}$, then the random variable

$$\tilde{\chi} = \tilde{\rho} (n-1)^{-1} \hat{\tilde{\beta}}^2 \tilde{V}(t) \rightarrow \chi_1^2 \tag{23}$$

converges in distribution to a chi-squared distribution with 1 degree of freedom.

I shall make some insightful comments on (23) and refer the reader to its proof in Appendix B.

By inspection, one should readily recognize the behavior of the chosen function $\tilde{\beta}$, in Proposition 1, as it approximates the behavior of β. If two samples from the same population are compared using (23), the estimate of $\tilde{\beta}$, $\hat{\tilde{\beta}}$ in (20), would also tend to approach zero—as the sample size increases, and tend otherwise for samples belonging to distinct populations.

The real challenge, however, is to derive a relatively simple estimate of *Var(t)*, as defined in (7A), to replace *V*(t)*, as shown in (16). *Var(t)* is a sum of squared errors individually weighted by their probability of occurrence. In Proposition 1, $g(x_1,x_0)$ is proposed to provide that probability feature, but as an average probability of occurrence, instead. In this sense, comparing two samples from distinct populations would produce very high cumulative square errors using the combined vector *t*, but appropriately weighted by an average proportion. The proof of Proposition 1 is presented in Appendix B.

In principle, the overall behavior of (23) seems to track that of (16), and both random variables are asymptotically identically distributed under χ_1^2. Note that the *AsemiP*'s performance will not asymptotically approach that of the *SemiP*'s performance, as the number of samples increases; the former approximates the general behavior of the latter, i.e., it promotes a high separation

between meaningful signals (isolated objects) from noise (homogeneous and non-homogenous local regions).

3. Alternative Techniques

In this section, I make a few general comments on four other anomaly detection techniques, which shall be used in this paper for comparison purposes: their mathematical representations are presented in this section, without proofs, and a reference that fully describes their implementation issues and performances in the HYDICE dataset will be cited.

The four techniques are known as: RX (reed-xiaoli), PCA (principal component analysis), EST (eigen separation transform), and FLD (Fisher's linear discriminant). These techniques—or variations of them—are commonly used in the hyperspectral community. The RX technique is based on the generalized likelihood ratio test and on the assumption that the population distribution family of both test and reference samples are multivariate normal. The FLD technique is also based on the same assumption, but differs in its subtleties in answering the question whether the test and reference samples are drawn from the same normal distribution. The FLD technique promotes separation between classes and variance reduction within each class. The PCA and EST techniques are both based on the same general principle, i.e., data are projected from their original high dimensional space onto a significantly lower dimensional space using a criterion that promotes highest sample variability within each domain in this lower dimensional space. Differences between PCA and EST are better appreciated through their mathematical representations.

These four techniques were implemented with the conventional inside-outside window approach, where *optimum* window sizes were chosen to account for the target-size range of interest (see section 4.1), i.e., *5*x*5* inside window embraced by a *9*x*9* outside window and leaving a guard area between pixels *6* and *7*, inclusive, from the center of the inside window. These techniques are represented by the following set of equations:

$$Score_{RX} = (\bar{x}_{in} - \bar{x}_{out})^t C_{out}^{-1} (\bar{x}_{in} - \bar{x}_{out}) \tag{24}$$

$$Score_{PCA} = E_{in}^t (\bar{x}_{in} - \bar{x}_{out}) \tag{25}$$

$$Score_{EST} = E_{\Delta C}^t (\bar{x}_{in} - \bar{x}_{out}) \tag{26}$$

$$Score_{FLD} = E_{S_b / S_w}^t (\bar{x}_{in} - \bar{x}_{out}) \tag{27}$$

where $\bar{x}_{in}$ is a sample mean vector from the set of inside-window vectors, each having 150 spectral bands (see section 7.1); $\bar{x}_{out}$ is similar but sampled from the outside window; C_{out}^{-1} is

the inverse sample covariance using all vectors sampled from the outside window; E_{in}^{t} is the highest energy eigenvector of the eigenvector decomposition of the inside-window covariance; $E_{\Delta C}^{t}$ is the highest positive energy eigenvector of the eigenvector decomposition of the covariance difference (inside-widow minus outside-widow); and $E_{S_b / Sw}^{t}$ is the eigenvector decomposition of the scatter matrices ratio $S_B S_W^{-1}$, where

$$S_W = \sum_{i=1}^{N} (x_{in}^{(i)} - \overline{X}_{in})(x_{in}^{(i)} - \overline{X}_{in})^t + \sum_{i=1}^{M} (x_{out}^{(i)} - \overline{X}_{out})(x_{out}^{(i)} - \overline{X}_{out})^t \tag{28}$$

$$S_B = \sum_{i=1}^{N} (x_{in}^{(i)} - \overline{X}_{total})(x_{in}^{(i)} - \overline{X}_{total})^t + \sum_{i=1}^{M} (x_{out}^{(i)} - \overline{X}_{total})(x_{out}^{(i)} - \overline{X}_{total})^t \tag{29}$$

and $\overline{X}_{total}$ is the total sample average using all samples from the inside and outside windows.

For additional details on the implementation and performance of these techniques, see.[6]

4. Results

4.1 Data Description

An experiment was carried out on the data set from the hyperspectral digital imagery collection experiment (HYDICE) sensor. The HYDICE sensor records *210* spectral bands in the visible to near infrared (VNIR) and short-wave infrared (SWIR) and, for the purpose of this paper, it is sufficient to know that each pixel in a scene is actually a vector, consisting of *210* components. An extended description of this dataset can be found, for example, in.[25]

The results shown in this section for one data sub-cube are representative for other sub-cubes in the HYDICE (forest radiance) dataset. An illustrative sub-cube (shown as an average of *150* bands; *640* x *100* pixels) is shown in figure 1 (left). (I only used *150* bands by discarding 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*x*4* pixels, and each pixel corresponds to an area of about *1.3* x *1.3* square meters at the given altitude. The main goals of local anomaly detection algorithms on these types of scenes are to detect objects that seem clearly anomalous to its immediate surroundings, in some predetermined feature space, and to yield in the process a tolerable number of nuisance detections. Targets are often found to be anomalous to its immediate surroundings.

[6] Kwon, 2003.
[25] Schweizer, 2000.

4.2 Implementation Notes

I now present some helpful hints on the implementation of the *SemiP* and *AsemiP* algorithms.

1. SemiP Anomaly Detector

- Sampling Mechanism: Use the mechanism described in section 2.1 to sample a pair of random feature vectors x_{ij} (i = 0 [reference], 1 [test]; j=1,…,J) from HSI. I used a 9-pixel (3x3) test window, a 56-pixel reference window, and a 60-pixel variability window, as shown in figure 1. Note that the size of the variability window determines the size of the feature vectors, that is, x_{0j} and x_{1j} have the same size, J = 60.
- Statistical Independence: An attempt should be made to promote statistical independence in HSI. See discussion in section 2.1.
- Function Maximization: Perform an unconstrained maximization of $l(\alpha,\beta)$ in (10), or minimization of $[-l(\alpha,\beta)]$, to obtain the extremum estimates (α^*,β^*). For this report, I used one of the standard unconstrained minimization routines available in MATLAB™ software (i.e., fminsearch), and set the initial values of (α,β) to (0,0).

Variance Under the Null Hypothesis: $V^*(t)$ in (16) should be computed using (13) and a discrete version estimate of (7A):

$$\hat{E}(t^k) = \sum_i t_i^k \hat{g}_0(t_i); V^*(t) = \hat{E}(t^{k=2}) - \hat{E}^2(t^{k=1}). \tag{31}$$

- *Decision Threshold:* Using (16), high values of χ reject hypothesis H_o, hence, detecting anomalies. Set a decision threshold based on the Type I error, i.e., 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*), or alternatively find and use a suitable threshold that yields a value at the *knee* of the *SemiP*'s corresponding ROC curve.

2. AsemiP Anomaly Detector

In contrast to the *SemiP* algorithm, the *AsemiP* algorithm is significantly simpler to implement, as the latter suite does not require specialized subroutines (unconstrained minimization) to perform its function. Using the sampling mechanism described in section 2.1, the variables in Proposition 1 are straightforward to implement. One is also expected to promote statistical independence and to take a sufficiently large number of samples (larger than *30*) to justify the use of approximation theorems of mathematical statistics. We used the sampling mechanism proposed in section 2.1 to obtain a pair of random feature vectors x_{ij} (*i = 0* [*reference*], *1* [*test*]; *j=1,…,J*). I used a *9*-pixel (*3x3*) *test window*, a *56*-pixel *reference window,* and a *60*-pixel *variability window*, as shown in figure 1, where *J = 60*. For a statistical decision, high values obtained by using (23), or equivalently (14B), reject the hypothesis H_o in Proposition 1, thus, detecting anomalies. Set a decision threshold based on a choice of type I error using, as the base

distribution, the chi-square distribution with *1* degree of freedom. Alternatively, find and use a suitable threshold that yields a value at the *knee* of the *AsemiP'*s corresponding ROC curve.

4.3 Comparative Results

ROC curves will be used in this subsection as a means to quantitatively compare the six approaches described in this paper.

As described in section 4.1, the set of *14* ground vehicles near the treeline in figure 1 constitutes our target set. However, 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 non-targets among the different algorithms.

Figure 2 shows ROC curves produced by the output of the six algorithms on the HYDICE data scene shown in figure 1. 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; call it *TargetTruth*. If we denote the region outside the *TargetTruth* as *ClutterTruth,* then the intersection between *TargetTruth* and *ClutterTruth* is *zero* and the entire scene is the union of *TargetTruth* and *ClutterTruth*. In this paper, for a given decision threshold, the proportion of target detection (PD) is measured as the proportion between the number of detected pixels belonging to *TargetTruth* over all pixels belonging to *TargetTruth*. On the other hand, the proportion of false alarms (PFA) is measured as the proportion between the number of detected pixels belonging to *ClutterTruth* over all pixels belonging to *ClutterTruth.*

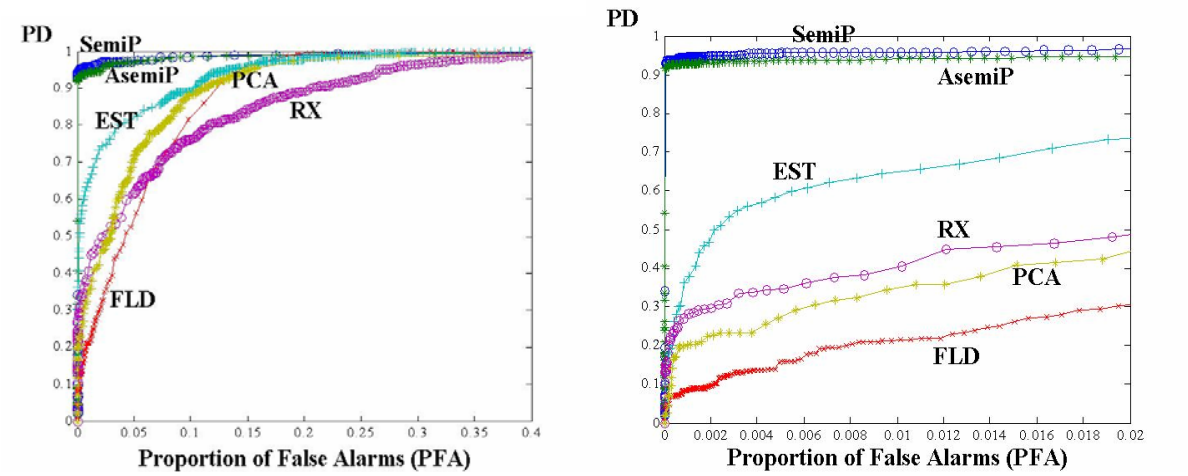

Figure 2. ROC curves for the HYDICE data scene shown in figure 1. The *SemiP* and *AsemiP* detectors are noticeably less sensitive to different decision thresholds; their performances almost achieve an ideal ROC curve for that scene, i.e., a step function starting at point (PFA=0,PD=1).

In general, the quality of a detector can be readily assessed by noticing a key feature in the shape of its ROC curve: The closer the *knee* of a ROC curve is to the PD axis, the less sensitive the approach is to different decision thresholds. In other words, PFA does not change significantly as PD increases. (An ideal ROC curve resembles a step function that starts at point [PFA=0,PD=1].) As it can be readily assessed from figure 2, the *SemiP* and *AsemiP* detectors clearly outperform the other four techniques on the tested scene. That difference in performance can be better appreciated in figure 2 (right), where PFA is further restricted to a maximum value of *0.02* compared to *0.4* in figure 2 (left). Beyond the value of PFA = *0.4*, these ROC curves reach PD near *1.0*.

The significant display of performance shown in figure 2 by the proposed algorithms can be further appreciated by taking a closer look at the output surfaces produced by all six detectors, as they show evidences of local candidate anomalies in the scene. The intensity of local peaks shown in figure 3 reflects the strength of the evidence. It is evident from figure 3 that the surfaces produced by RX, PCA, EST, and FLD detectors are expected to be significantly more sensitive to changing decision thresholds then the ones produced by the *SemiP* and *AsemiP* detectors.

Although the *2-dim* (2-dimensional) version of the output surfaces shown in figure 3 displays useful differences among the responses of the six detectors, they do not make full justice to the quality of the *SemiP* and *AsemiP* detectors, as a *3-dim* perspective of their surfaces would. A *3-dim* perspective of the *SemiP*'s and *AsemiP*'s output surfaces are depicted in figure 4.

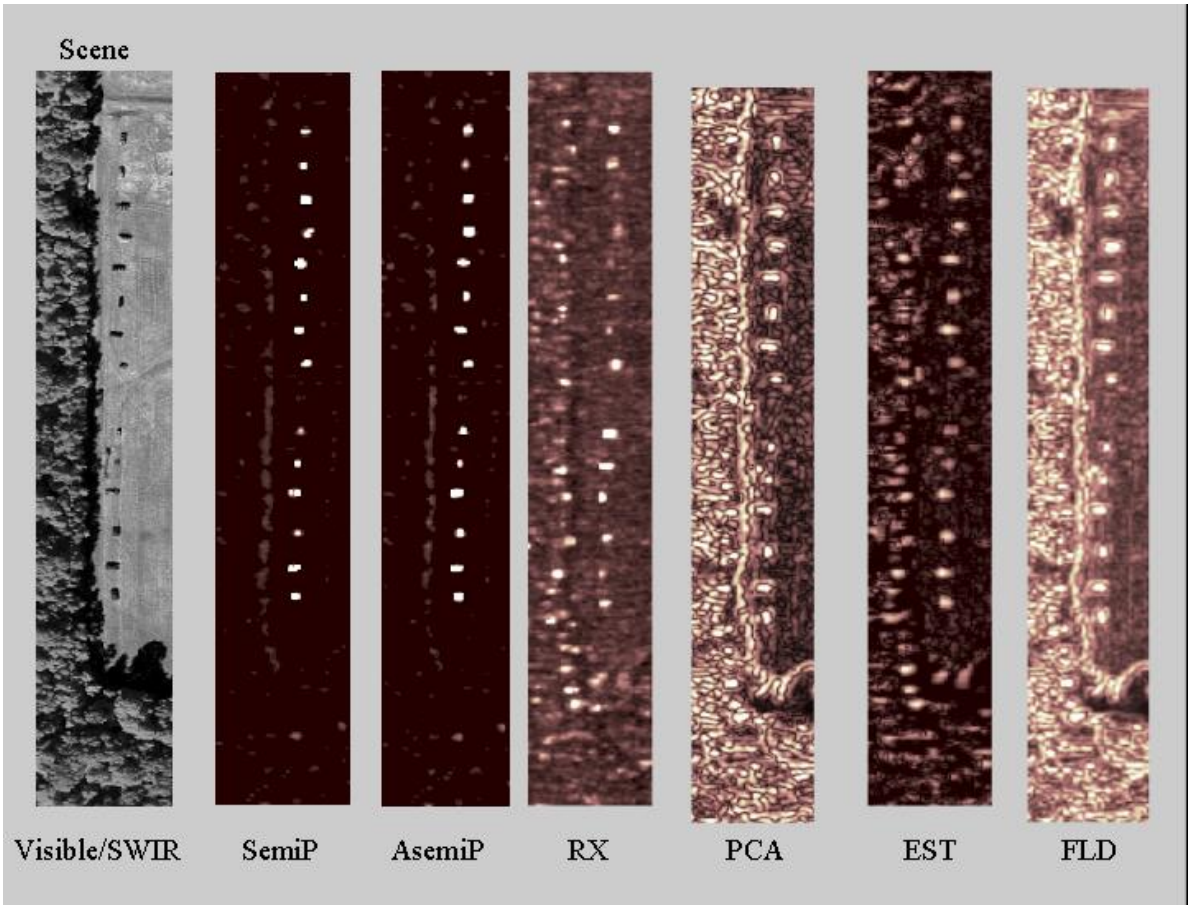

Figure 3. Decision surfaces for a HYDICE data scene (far left). The intensity of local peaks reflects the strength of evidences as *seen* by different anomaly detectors. Boundary issues were ignored in this test; surfaces were magnified to about the size of the original image only for the purpose of visual comparison.

Both surfaces in figure 4 were clipped at the value of *3,000*, but some of their values do continue to significantly higher numbers. The dominant (clipped) peaks are the results produced by the fourteen targets near the *treeline*.

Areas containing the presence of clutter mixtures (e.g., edge of terrain, edge of tree clusters), where other methods usually yield a high number of false alarms (false anomalies), are suppressed by the new approaches. The reason for this suppression is that, as part of the overall comparison strategy, the reference and test feature vectors are not compared as two individual samples. Instead, a new sample is constructed—the union of both cells—and compared to the test sample. This indirect comparison approach, which is inherent in both detectors, ensures that a component of a mixture (e.g., shadow) shall to be detected as a local anomaly when it is tested

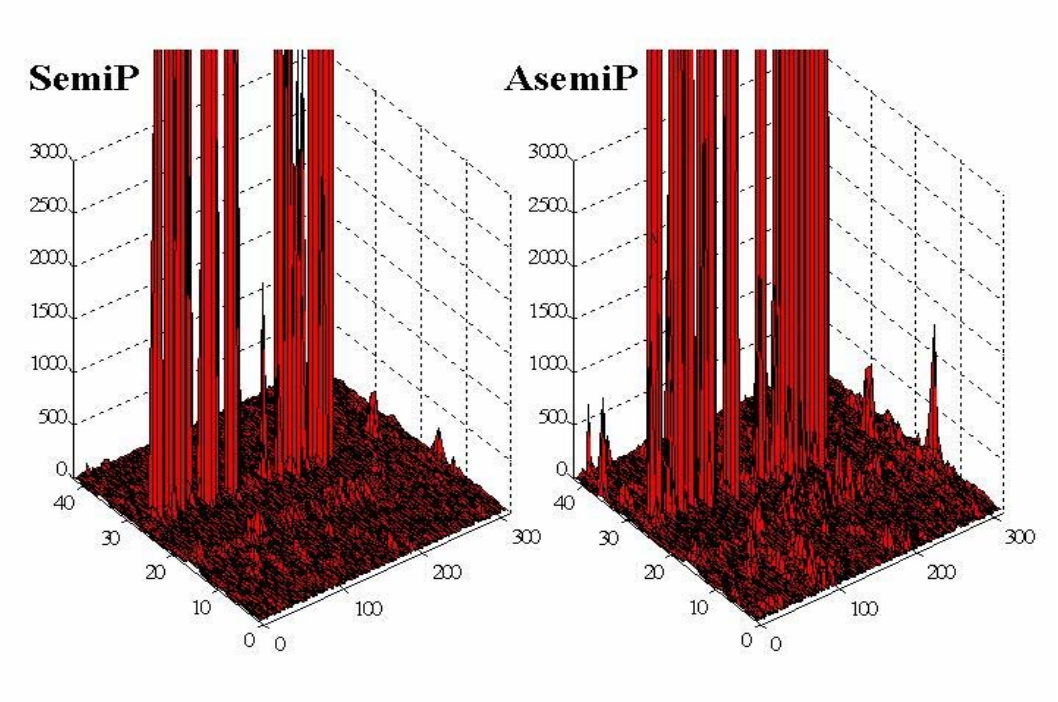

Figure 4. Decision surfaces (*3-dim* version) produced by the *SemiP* and *AsemiP* detectors, their *2-dim* versions are shown in figure 3. The dominant peaks represent the presence of the 14 targets parked near the *treeline*. Less-dominant peaks represent areas in the scene most prone to cause *nuisance* detections (e.g., region discontinuity). The *SemiP* and *AsemiP* detectors show a remarkable ability to accentuate genuine local anomalies from their surroundings.

against the mixture itself (e.g., trees and shadows). Performances of such cases are represented in figure 4 in the form of softer anomalies (significantly less-dominant peaks).

It is evident from figure 3 and figure 4 that both detectors perform remarkably well accentuating the presence of dominant local anomalies (e.g., targets and their shadows) from softer anomalies (e.g., region discontinuity). This ability explains the *SemiP's* and *AsemiP's* superior ROC-curve performances shown in figure 2.

Processing Time: For completeness, we report the processing time in minutes (min) for a cube of size *640* x *100* (pixels) x *150* (bands) using a personal computer (CPU speed: *1.80* GHz; RAM memory: *1.0* Gbytes), MATLAB™ software (release *13*), and three detectors (RX, *AsemiP*, and *SemiP*). The recorded times were: *20.6* min (RX), *22.1* min (*AsemiP*), *42.9* min (*SemiP*). Computing the local variance-covariance matrix and its inverse dominated the RX processing time. Applying locally a HP filter in the spectral domain and applying SAM on the resulting vectors dominated the *AsemiP* processing time. And, finally, applying locally a HP filter and a

spatial SAM, and using an unconstrained minimization routine (see section 4.3) dominated the *SemiP* processing time.

5. Conclusion

I have presented an indirect approach to test two-sample hypotheses in HSI. The approach is not based on a physical motivation but on a statistical one. Using this approach, I designed two fully unsupervised object detectors (*SemiP* and *AsemiP*) for HSI. I showed performance agreement between these detectors through ROC curves, and other means. Performances of both detectors in the visible to short-wave infrared region of the spectrum were compared to performances of four other techniques. The proposed algorithms clearly outperform the others. The asymptotic distribution of the test statistic under H_o in both proposed algorithms is independent of the unknown parameters, which implies that both *SemiP* and *AsemiP* have the constant probability-of-error property. With this property, one can—in theory—select a decision threshold that yields a virtual zero probability of error. *Error* in this context means detection of non-anomalies, which is purely based on class similarities between test samples and their immediate surroundings, not necessarily non-targets (targets defined as particular objects of interest). This distinction should be noted.

6. References

1. Schowengerdt, R. A. *Remote Sensing, Models and Methods for Image Processing*, 2nd ed. Academic: San Diego, CA, 1997.
2. Campbell, J. B. *Introduction to Remote Sensing*, 2nd ed. Guilford: New York, 1996.
3. Lillesand, T. M.; Kiefer, R. W. *Remote Sensing and Image Interpretation*. Wiley: New York, 1994.
4. Schott, J. R. *Remote Sensing: The Image Chain Approach*. Oxford Univ. Press: Oxford, U.K., 1997.
5. Yu, X.; Hoff, L.; Reed, I.; Chen, A.; Stotts, L. Automatic target detection and recognition in multiband imagery: A unified ML detection and estimation approach, *IEEE Tran. Image Processing*, **January 1997**, *6*, pp. 143–156.
6. Kwon, H.; Der, S.Z.; Nasrabadi, N.M. Adaptive anomaly detection using subspace separation for hyperspectral imagery. *Opt. Eng*. **November 2003**, *42 (11)*, pp. 3342–3351.
7. Stein, D. W.; Beaven, S. G. The fusion of quadratic detection statistics applied to hyperspectral imagery. *IRIA-IRIS Proc. 2000 Meeting of the MSS Specialty Group on Camouflage, Concealment, and Deception*, March 2002, pp. 271–280.
8. Stein, D.; Beaven, S.; Hoff, L.; Winter, E.; Scheaum, A.; Stocker, A. Anomaly detection from hyperspectral imagery. *IEEE Signal Processing Mag*, **January 2002**, *19*, pp. 58–69.
9. Manolakis, D.; Shaw, G. Detection algorithms for hyperspectral imaging applications. *IEEE Signal Processing Magazine* **January 2002**, pp 29.
10. Crist, E., Schwartz, C.; Stocker, A. Pairwise adaptive linear matched-filter algorithm. in *Proc DARPA Adaptive Spectral Reconnaissance Algorithm Workshop*, January 1999.
11. Haskett, H. T.; Sood, A. K. Adaptive real-time endmember selection algorithm for sub-pixel target detection using hyperspectral data. in *Proc. 1997 IRIS Specialty Group Camouflage, Concealment, Deception*, October 1999.
12. Grossmann, J.; Bowles, J.; Hass, D.; Antoniades, J.; Grunes, M.; Palmadesso, P.; Gillis, D.; Tsang, K.; Baumback, M.; Daniel, M.; Fisher, J.; Triandaf, I. Hyperspectral analysis and target detection system for the adaptive-spectral reconnaissance program (ASRP). *Proc SPIE*, April 1998, *372*, pp.2–13.

13. Rosario, D. S. Simultaneous Multi-Target Detection and False Alarm Mitigation Algorithm for the Predator UAV TESAR ATR System. *Proc. of the 10th International Conference on Image Analysis and Processing (ICIAP'99),* sponsored by the *IEEE Computer Society*, Venice, Italy, September 1999.

14. Rosario, D. S. End-to-End Performance of the TESAR ATR System. *Proc. of the SPIE 14th Annual International Symposium on Aerospace/Defense Sensing, Simulation, and Control*, Orlando, FL, April 2000.

15. Rosario, D. S. Highly Effective Logistic Regression Model for Signal (Anomaly) Detection. in *Proc. 2004 IEEE ICASSP,* Montreal, Canada, May 2004.

16. Qin , J.; Zhang, B. A goodness of fit test for logistic regression models based on case-control data. *Biometrika,* **1997**, *84*, 609–618.

17. Fokianos, K.; Kedem, B.; Qin, J. A semiparametric approach to the one-way layout. *Technometrics* **2001**, pp. 56–65.

18. Anderson, J. A. "Separate sample logistic discrimination," *Biometrika* **1972**, *59*, 19–35.

19. Prentice, R.; Pyke, R. Logistic disease incidence models and case-control studies. *Biometrika* **1979**, *66*, pp. 403–411.

20. Cox, D. R. Some procedures associated with the logistic qualitative response curve. *Research Papers in Statistics: Festschrift for J. Neyman,* John Wiley: New York, 1966, pp. 55–71.

21. Kelly, E. J. *Adaptive Detection and Parameter Estimation for Multidimensional Signal Models*, MIT Lincoln Laboratory: Lexington, MA, Tech. Rep. 848, April 1989.

22. Casella, G.; Berger, R. L. *Statistical Inference*, Duxbury Press: Belmont, CA 1990, pp. 112–120, 184, 222.

23. Kay, R.; Little, S. Transformations of the explanatory variables in the logistic regression model for binary data. *Biometrika* **1987**, *74*, 495–501.

24. Cramer, J. S. The Origin of Logistic Regression. Tinbergen Institute, U. of Amsterdam: TI 2002-119/4. Available on: http://ideas.repec.org/p/dgr/uvatin/20020119.html.

25. Schweizer, S. M.; Moura, J. M. F. Hyperspectral Imagery: Clutter Adaptation in Anomaly Detection. *IEEE Trans. Information Theory* **August 2000**, *46*, no. 5, pp. 1855–1871.

26. Amemiya, T. *Advanced Econometrics*, Harvard U. Press : Cambridge, MA, 1985, pp. 110–112.

27. Lehmann, E. L. *Theory of Point Estimation*, Wadsworth & Brooks: Pacific Grove, CA, 1991, pp. 333–336, and Chapter 5.

28. Serfling, R. J. , *Approximation Theorems of Mathematical Statistics*, Wiley: New York, 1980, pp. 19.

Appendix A

Proof of the SemiP algorithm: Lemma 1A is relevant to estimators based on function maximization with respect to unknown parameters.

Lemma 1A.[26] *Assumptions:*

(i) Let Θ be an open subset of the Euclidean K-space. (Thus the true value θ_o is an interior point of Θ.)

(ii) $Q_T(y,\theta)$ is a measurable function of vector y for all $\theta \in \Theta$ and $\partial Q_T / \partial \theta$ exists and is continuous in an open neighborhood $N_1(\theta_o)$ of θ_o. (Note that this implies $Q_T(y,\theta)$ is continuous for $\theta \in N_1$, where T is the sample size.)

(iii) There exists an open neighborhood $N_2(\theta_o)$ of θ_o such that $T^{-1} Q_T(\theta)$ converges to a nonstochastic function $Q(\theta)$ in probability uniformly in θ in $N_2(\theta_o)$, and $Q(\theta)$ attains a strict local maximum at θ_o.

Let Θ_T be set of roots of the equation

$$\frac{\partial Q_T}{\partial \theta} = 0 \tag{1A}$$

corresponding to the local maxima. If that set is empty, set Θ_T equals to {0}. Then, for any $\varepsilon > 0$,

$$\lim_{T \to \infty} P[\inf_{\theta \in \Theta_T} (\theta - \theta_0)'(\theta - \theta_0) > \varepsilon] = 0. \tag{2A}$$

In essence, Lemma 1A affirms that there is a consistent root of (1A). (For the proof, see[26]). Under certain conditions, a consistent root of (1A) is asymptotically Normal. The affirmation is shown in Theorem 1A, where asymptotic convergence is denoted by $A \to B$.

Theorem 1A.[26] *Assumptions:*

(i) All the assumptions of Lemma 1.

(ii) $\frac{\partial^2 Q_T}{\partial \theta \partial \theta'}$ exists and is continuous in an open, convex neighborhood of θ_o.

(iii) $\frac{\partial^2 Q_T}{\partial \theta \partial \theta'}$ converges to a finite nonsingular matrix $S(\theta_0) = \lim E[T^{-1}(\frac{\partial^2 Q_T}{\partial \theta \partial \theta'})_{\theta_0}]$ in probability for any sequence θ_T^* such that $\theta_T^* = \theta_0$.

[26] Amemiya, 1985.

(iv)

$$\sqrt{T}\,(\frac{\partial Q_T}{\partial \theta})_{\theta_0} \;\rightarrow\; N[0, V(\theta_0)],\ \textit{where} \tag{3A}$$

$$V(\theta_0) = \lim E[T^{-1}(\frac{\partial Q_T}{\partial \theta})_{\theta_0} \times (\frac{\partial Q_T}{\partial \theta'})_{\theta_0}]. \tag{4A}$$

Let $\{\hat{\theta}_T\}$ *be a sequence obtained by choosing one element from* Θ_T *defined in Lemma 1 such that* $\hat{\theta}_T \rightarrow \theta_0$.

Then

$$\sqrt{T}\left(\hat{\theta}_T - \theta_0\right) \rightarrow N(0,\Sigma),\ \textit{where} \tag{5A}$$

$$\Sigma = S(\theta_0)^{-1} V(\theta_0) S(\theta_0)^{-1}.- \tag{6A}$$

For the proof, see also.[26]

The semiparametric model's MLE solution satisfies the assumptions of Lemma 1A, including of course (1A) via (11) and (12). Therefore, by Lemma 1, (α^*,β^*) is consistent and, as we shall see by Theorem 1A, it converges asymptotically to a Normal distribution.

Under H_0: $\beta=0$ $(g_1=g_0)$, I shall use the following notation for the moments of t (the combined sample vector) with respect to the reference distribution g_0:

$$\begin{aligned} E(t^k) &\equiv \int t^k g_0(t)dt\,, \\ Var\ (t) &\equiv E(t^2) - E^2(t) \end{aligned} \tag{7A}$$

Let (α_0,β_0) be the true value of (α,β) under model (3)-(4) and assume $\rho = n_1/n_0$ remains constant as both n_1 and n_0 go to *infinity*. Define $\nabla \equiv \left(\frac{\partial}{\partial\alpha}, \frac{\partial}{\partial\beta}\right)$ and notice from (11) and (12) that $E\left[\nabla l\left(\alpha_0, \beta_0\right)\right] = 0$. Under the null hypothesis (H_0: $\beta = 0$ [$g_1 = g_0$]), using (4), (9), (11), (12) and (7A) one can verify that

$$\begin{aligned} -\frac{1}{n}\frac{\partial^2 l(\alpha_0,\beta_0)}{\partial\alpha\partial\beta} &\rightarrow K_1 \int \frac{t\exp(\alpha_0+\beta_0 t)}{1+\rho\exp(\alpha_0+\beta_0 t)}\, g_0(t)\, dt \\ &= K_2 \int t \cdot g_0(t)[\exp(\alpha_0+\beta_0 t) g_0(t)]\, dt \\ &= \frac{\rho}{1+\rho} E(t), \end{aligned} \tag{8A}$$

[26] Amemiya, 1985.

where K_1 and K_2 are constants involving (n_1,n_0) and $\rho/(1+\rho) = n_1/n$ (where $n = n_1+n_0$). Using similar argument to arrive at (8A) and applying the *weak law of large numbers* (WLLN) (see, for instance,[27]), one can use assumption *(iii)* in Theorem 1A to recognize that

$$-\frac{1}{n}\nabla\nabla' l(\alpha_0,\beta_0) \to S = \frac{\rho}{1+\rho}\begin{pmatrix} 1 & E(t) \\ E(t) & E(t^2) \end{pmatrix} \tag{9A}$$

in probability as $n \to \infty$. It follows that S is nonsingular and its inverse is

$$S^{-1} = \frac{1}{E(t^2)-E^2(t)}\begin{pmatrix} E(t^2) & -E(t) \\ -E(t) & 1 \end{pmatrix}\frac{1+\rho}{\rho}. \tag{10A}$$

Our interest is only in the parameter β, so, let S_β denote the lower-right component of the expanded version of S^{-1} and use (7A) to obtain

$$S_\beta = \frac{1}{E(t^2)-E^2(t)}\frac{1+\rho}{\rho} = \frac{1}{Var(t)}\frac{1+\rho}{\rho}. \tag{11A}$$

Using also the application of the *central limiting theorem* (CLT) in Theorem 1A *(iv)* and the fact that

$$E\left[\nabla l(\alpha_0,\beta_0)\right] = 0\ , \tag{12A}$$

from (11) and (12), one can write

$$\sqrt{n}\left[\nabla l(\alpha_0,\beta_0)\right] \to N[0,V(\alpha_0,\beta_0)], \tag{13A}$$

where

$$V(\alpha_0,\beta_0) = \frac{\rho}{1+\rho}\begin{bmatrix} 1 & E(t) \\ E(t) & E(t^2) \end{bmatrix} - \rho\begin{bmatrix} 1 \\ E(t) \end{bmatrix}\begin{bmatrix} 1 & E(t) \end{bmatrix}. \tag{14A}$$

$V(\alpha_0,\beta_0)$ is a direct result from (4A), see, for instance[16]. Using the conclusion of Theorem 1A, or (5A)-(6A), in terms of S_β in (11A) and the lower-right component of the expanded version of $V(\alpha_0,\beta_0)$ in (14A), one can verify that

$$\sqrt{n}\beta^* \to N\left(0, \frac{\rho^{-1}(1+\rho)^2}{Var(t)}\right) \tag{15A}$$

[16] Qin, 1997.
[27] Lehmann, 1991.

and having the left side of (15A) normalized by the asymptotic variance and then squared, one can conclude (see convergence results, for instance, in[22]) that the resulting random variable

$$\chi = n\rho(1+\rho)^{-2}\beta^{*2}V^{*}(t) \rightarrow \chi_1^2 \tag{16A}$$

converges to a chi square distribution with *1* degree of freedom, where $V^{*}(t)$ estimates *Var(t).* A multivariate solution is presented in[17].

[17] Fokianos, 2001.
[22] Casella, 1990.

Appendix B

Proof of Proposition 1: If hypothesis H_0: $\left(\tilde{\beta} = 0; \zeta_1 = \zeta_2 = 1\right)$ is true in Proposition 1, then $\sigma^2 = \sigma_1^2 = \sigma_0^2$ and, using the independent assumptions of x_1 and x_0, and CLT, it follows that

$$\frac{\hat{\tilde{\beta}}}{\sqrt{\frac{1}{n0} + \frac{1}{n_1}}\sigma_0} = \frac{\hat{\tilde{\beta}}}{\sqrt{\frac{1}{n0} + \frac{1}{n_1}}\sigma_1} \xrightarrow[\substack{n_0 \to \infty \\ n_1 \to \infty}]{} N(0,1), \tag{1B}$$

$\hat{\tilde{\beta}}$ as defined in Proposition 1; in addition, the following estimators of σ_1^2 and σ_0^2 are known to be consistent [see, for example,[26]]:

$$S_1^2 = \frac{\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2}{n_1 - 1}, \text{ and } S_0^2 = \frac{\sum_{i=1}^{n_0}(x_{0i} - \bar{x}_0)^2}{n_0 - 1}. \tag{2B}$$

Using both samples x_1 and x_0, let the following be another estimator of σ_0^2 (or σ_1^2), given that under H_0 $\sigma_0^2 = \sigma_1^2$,

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_0 - 1)S_0^2}{(n_1 - 1) + (n_0 - 1)}. \tag{3B}$$

The estimator S^2 is unbiased under H_0, as its expected value $E[S^2]$ is equal to σ_0^2 and σ_1^2:

$$\begin{aligned} E\left(S^2\right) &= \frac{(n_1 - 1)E\left(S_1^2\right) + (n_0 - 1)E\left(S_0^2\right)}{(n_1 - 1) + (n_0 - 1)} \\ &= \frac{(n_1 - 1)\sigma_1^2 + (n_0 - 1)\sigma_0^2}{(n_1 - 1) + (n_0 - 1)} \\ &= \sigma_0^2. \end{aligned} \tag{4B}$$

True because S_0^2 and S_1^2 are consistent estimators and, under H_0, $\sigma_0^2 = \sigma_1^2$. I want to prove now a WLLN for S^2 to verify that S^2 is also a consistent estimator. Using *Chebychev's* inequality[28] under H_0:

$$P(|S^2 - \sigma_0^2| \geq \varepsilon) \leq \frac{E(S^2 - \sigma_0^2)^2}{\varepsilon^2} = \frac{Var(S^2)}{\varepsilon^2} \tag{5B}$$

and, thus, a sufficient condition that S^2 converges in probability to σ_0^2, or σ_1^2, is that $Var\left(S^2\right) \xrightarrow[(n_0, n_1) \to \infty]{} 0$.

[26] Amemiya, 1985.
[28] Serfling, 1980.

Note that $Var(S^2)$ can be expressed as

$$Var(S^2) = \left[\frac{(n_1-1)}{(n_1-1)+(n_0-1)}\right]^2 Var(S_1^2) + \left[\frac{(n_0-1)}{(n_1-1)+(n_0-1)}\right]^2 Var(S_0^2) \tag{6B}$$

and, as S_0^2 and S_1^2 are both consistent estimators, their variances must converge to zero,

$$Var(S_1^2) \xrightarrow[n_1\to\infty]{} 0; \quad Var(S_0^2) \xrightarrow[n_0\to\infty]{} 0, \tag{7B}$$

also

$$\left[\frac{(n_k-1)}{(n_1-1)+(n_0-1)}\right]^2 \xrightarrow[(n_1,n_0)\to\infty]{} 0; \; k = 0,1. \tag{8B}$$

Then $Var(S^2) \xrightarrow[(n_1,n_0)\to\infty]{} 0$ and, therefore, under H_0:

$$\frac{\sigma_0^2}{S^2} = \frac{\sigma_1^2}{S^2} \to 1, \text{ as } (n_0, n_1) \to \infty. \tag{9B}$$

Using the same argument to arrive at (9B), we can also show that under H_0:

$$\frac{S_t^2}{\sigma_0^2} = \frac{S_t^2}{\sigma_1^2} \to \zeta_1 = \zeta_2 = 1, \text{ as } n = (n_1 + n_0) \to \infty, \tag{10B}$$

where S_t^2 is a consistent estimator of σ^2 under H_0, or

$$S_t^2 = (n-1)^{-1} \sum_{i=1}^{n} (t_i - \bar{t})^2, \; \bar{t} = n^{-1} \sum_{i=1}^{n} t_i. \tag{11B}$$

Furthermore, since S_t^2 is the sample variance of t (the combined vector) under H_0, one can readily verify that

$$\frac{S_t}{\sigma_0} = \frac{S_t}{\sigma_1} \to 1, \text{ as } n = (n_1 + n_0) \to \infty \tag{12B}$$

(see, for example,[26] Chapter 5).

To finalize the proof, consider Theorem 1B below.

[26] Amemiya, 1985.

Theorem 1B (Slutsky). *Let X_n tend to X in distribution and Y_n tend to c in probability, where c is a finite constant. Then*

(i) $X_n + Y_n$ tend to X+c in distribution;

(ii) $X_n Y_n$ tend to cX in distribution;

(iii) X_n/Y_n tend to X/c in distribution, if c is not zero.

See proof in.[28]

Using (1B), (9B), (12B) and the *Slutsky* Theorem, one can conclude that

$$\left[\left(\frac{\hat{\tilde{\beta}}}{\sqrt{\frac{1}{n0}+\frac{1}{n_1}}\sigma_0}\right)\left(\frac{\sigma_0^2}{S^2}\right)\right]\left(\frac{S_t}{\sigma_0}\right)\xrightarrow[\substack{n_0\to\infty\\ n_1\to\infty}]{} N(0,1) \tag{13B}$$

and that by squaring (13B) and using convergence results from,[22] one can also conclude that

$$\tilde{\chi} = \left(\frac{\hat{\tilde{\beta}}^2}{\frac{1}{n0}+\frac{1}{n_1}}\right)\frac{S_t^2}{S^4}\xrightarrow[\substack{n_0\to\infty\\ n_1\to\infty}]{} \chi_1^2, \tag{14B}$$

which can be readily reformatted into (23) using the definitions given in Proposition 1 and in this proof. Equation (14B) concludes the proof.

[22] Casella, 1990.
[28] Serfling, 1980.

INTENTIONALLY LEFT BLANK.

Distribution List

Admnstr
Defns Techl Info Ctr
ATTN DTIC-OCP (Electronic copy)
8725 John J Kingman Rd Ste 0944
FT Belvoir VA 22060-6218

DARPA
ATTN IXO S Welby
ATTN R Hummell
3701 N Fairfax Dr
Arlington VA 22203-1714

Ofc of the Secy of Defns
ATTN ODDRE (R&AT)
The Pentagon
Washington DC 20301-3080

Army Rsrch Physics Div
ATTN AMSRD-ARL-RO-MM R Launer
PO Box 12211
Research Triangle Park NC 27709-2211

AVCOM
ATTN AMSAM-RD-WS-PL W Davenport
Bldg 7804
Redstone Arsenal AL 35898

US Army TRADOC
Battle Lab Integration & Techl Dirctrt
ATTN ATCD-B
10 Whistler Lane
FT Monroe VA 23651-5850

CECOM NVESD
ATTN AMSRD-CER-NV-D J Ratches
10221 Burbeck Rd Ste 430
FT Belvoir VA 22060-5806

Dir for MANPRINT Ofc of the
Deputy Chief of Staff for Prsnnl
ATTN J Hiller
The Pentagon Rm 2C733
Washington DC 20301-0300

US Army Aberdeen Test Center
ATTN CSTE-DT-AT-WC-A Carlen
ATTN CSTE-DTC-AT-TC-N D L Jennings
400 Colleran Road
Aberdeen Proving Ground MD 21005-5059

US Army ARDEC
ATTN AMSTA-AR-TD
Bldg 1
Picatinny Arsenal NJ 07806-5000

US Army Aviation & Mis Lab
ATTN AMSRD-AMR-SG-IP H F Anderson
Redstone Arsenal AL 35809

US Army Avn & Mis Cmnd
ATTN AMSAM-RD-SG-IP R Sims
Redstone Arsenal AL 35898

Commanding General
US Army Avn & Mis Cmnd
ATTN AMSAM-RD W C McCorkle
Redstone Arsenal AL 35898-5000

US Army CERDEC, NVESD
ATTN AMSRD-CER-NV J Hilger
ATTN AMSRD-CER-NV P Perconti
ATTN AMSRD-CER-NV R Driggers
FT Belvoir VA 22060-5806

US Army Natick RDEC Acting Techl Dir
ATTN SBCN-TP P Brandler
Kansas Street Bldg78
Natick MA 01760-5056

US Army PM NV/RSTA
ATTN SFAE-IEW&S-NV D Ferrett
10221 Burbeck Rd
FT Belvoir VA 22060-5806

US Army RDECOM ARDEC
ATTN AMSRDL-AAR-QES P Willson
Radiographic Laboratory, B.908
Picatinney Arsenal NJ 07806-5000

US Army Soldier & Biological Chem Ctr
ATTN AMSSB-RRT-DP B Loerop
Edgewood Chem & Biological Ctr Bldg E-5544
Aberdeen Proving Ground MD 21010-5424

US Army Tank-Automtv Cmnd RDEC
ATTN AMSTA-TR-R G Gerhart
Warren MI 48397-5000

US Army Topographic Engrg Ctr
ATTN CEERD-RR-S R Rand
7701 Telegraph Rd
Alexandria VA 22315

Commander
USAISEC
ATTN AMSEL-TD Blau
Building 61801
FT Huachuca AZ 85613-5300

AFRL/SNAA
ATTN M Jarratt
2241 avionics Circle Area B, Bldg 620
Wright Patterson AFB OH 45433-7321

CMTCO
ATTN MAJ A Suzuki
1030 S Highway A1A
Patrick AFB FL 23925-3002

SMDC
ATTN SMDC-TC-TD-YF A Aberle
PO Box 1500
Redstone Arsenal AL 35898

SITAC
ATTN H Stiles
ATTN K White
ATTN R Downie
11981 Lee Jackson Memorial Hwy Suite 500
Fairfax VA 22033-3309

Director
US Army Rsrch Lab
ATTN AMSRD-ARL-RO-D JCI Chang
ATTN AMSRD-ARL-RO-EL W Sander
PO Box 12211
Research Triangle Park NC 27709-2211

US Army Rsrch Office
ATTN AMSRD-ARL-RO-PP R Hammond
PO box 12211
Research Triangle Park NC

US Army Rsrch Lab
ATTN AMSRD-ARL-CI-OK-T Techl Pub
(2 copies)
ATTN AMSRD-ARL-CI-OK-TL Techl Lib
(2 copies)
ATTN AMSRD-ARL-D A Grum
ATTN AMSRD-ARL-D J M Miller
ATTN AMSRD-ARL-SE J Pellegrino
ATTN AMSRD-ARL-SE J Rocchio
ATTN AMSRD-ARL-SE-S J Eicke
ATTN AMSRD-ARL-SE-SE D Rosario
(5 copies)
ATTN AMSRL-SE-SE N Nasrabadi
ATTN AMSRL-SE-SE P Gillespie
ATTN IMNE-AD-IM-DR Mail & Records Mgmt
Adelphi MD 20783-1197