Parametric Classification Techniques for Theater Ballistic Missile Defense

Geoffrey L. Silberman

One of the fundamental challenges in theater ballistic missile defense (TBMD) is ascertaining which element in the threat complex is the lethal object. To classify the lethal object and other objects in the complex, it is necessary to model how these objects will appear to TBMD sensors. This article describes a generic parametric approach to building classifier models. The process is illustrated with an example of building a classifier for an infrared sensor. The formulas for probability of classification error are derived. The probability of error for a proposed classification scheme is vital to assessing its efficacy in system trade studies.
(Keywords: Classifier, Empirical feature distribution, Probability of error.)

INTRODUCTION

One of the fundamental challenges in theater ballistic missile defense (TBMD) is ascertaining which element in the threat complex is the lethal object. This task is known as discrimination. In the Navy Phase II Cost and Operational Effectiveness Analysis (COEA), an information-theoretic architecture was proposed for synthesizing the discrimination contributions of the ensemble of theater sensors. At the heart of each sensor's—or the system's—ability to discriminate is the degree to which it is able to distinguish the types of objects that constitute the ballistic complex. The different object classes may include the lethal object or reentry vehicle (RV), an attitude control module (ACM), the spent booster tank, solid fuel fragments, and separation debris. It is useful to classify not only the lethal object but other elements of the ballistic complex as well, in order that inferences about the RV location can be drawn from the spatial extent of the complex and from temporal events.

To classify objects, it is necessary to characterize how each object class will appear to a given sensor as a function of engagement geometry, object dynamics, object properties, sensor noise, and observation period. An ensemble of time series signatures is generated for each object class by varying the parameters for those inputs that will be unknown during an engagement, while fixing those parameters for those inputs that will be known. From the time series data, summary measures of data, called "features," are abstracted that characterize each class and serve to distinguish between classes. The feature realization for each class is called a "class-dependent feature distribution." The set of class-dependent feature distributions constitutes the "training set" for the scenario. Good features will be tightly clustered for a given class (low "intra-class dispersion") with relatively large distance between class clusters (high "inter-class separability").

The purpose of this article is to provide a brief overview of the procedure for building a classifier. The process will be illustrated by an example of an infrared (IR) seeker classifier built for the Navy Phase II COEA. Formulas will be developed for the probability of error in classifying the RV. The probability of error for a given sensor is an expression of the confidence

in the classification decision. The confidence bounds allow the classification decisions from different sensors to be synthesized in a system-level discrimination function.

THE TBMD CLASSIFICATION PROBLEM

For a given interceptor engagement scenario, the TBMD system will designate a single object in the ballistic complex as the RV. There are two reasons for this: the identified theater ballistic threats possess a single RV, and the hit-to-kill lethality requirements of TBMD preclude successful intercept of more than one object.

The RV will be designated on the basis of where its feature realization falls with respect to the class-dependent feature distributions for the RV class and each of the non-RV classes. Typically, objects are classified by using an *m*-ary hypothesis test. However, the TBMD classification problem differs from traditional hypothesis testing in important ways. In hypothesis testing, each object is considered individually and independently. The decision thresholds for each class are established *a priori* based on desired probabilities of "miss" (calling an RV a non-RV) and "false alarm" (calling a non-RV an RV). These probabilities are established by class-dependent feature distributions. In any given engagement, a hypothesis test could identify a number of objects within the RV decision region. However, in TBMD the decisions on targets are not independent: if one object is designated the RV, the others cannot be. Thus, it is necessary to select the object that is most "RV-like" rather than choosing *n* RV candidates at the specified confidence level. How to choose this object, and the probability of error in doing so, is the subject of the remainder of this article.

BUILDING A CLASSIFIER

There are several steps to building a classifier:

1. Render a range of signatures for each of the classes of objects in the ballistic complex.
2. Abstract features from the data.
3. Model the class-dependent feature distributions.
4. Develop decision boundaries in the feature space.
5. Evaluate the probability of error.

In practice, the last four steps are approached iteratively as new features or combinations of features are tested to improve the expected performance. The first two steps will be described generically. Steps three and four will be addressed with the standard methods used for IR discrimination in the Navy Phase II COEA. The fifth step is the objective of this article.

Render Signatures

The class-dependent feature distributions are derived from many signature realizations. These realizations should capture the variability in all of the object classes that the TBMD discrimination system is likely to see. To encompass this variability the following steps are necessary:

1. Describe the engagement geometries.
2. Describe the threat.
   - Characterize any of the static properties (size, shape, emissivity, mass, inertia) that can be apprehended by the sensor.
   - Characterize the rotational dynamics (orientation of the angular momentum vector and initial aspect angles for post-boost objects).
   - Characterize the translational dynamics (magnitude and direction of velocity vectors).
   - Characterize temporal events (burnout, separation, deployment) and the type and number of resulting objects.
3. Describe the environment (Earth's atmosphere, gravity, solar and stellar effects, clutter, diurnal variation).
4. Decide upon the sensor and the sensor noise model.

Parameter Specification

Parameters that will be known during an engagement can be fixed at discrete values; others must be varied to produce a representative range of signatures. Thus, it is desirable that all of the above descriptions be stochastic. Moreover, it is clear that some of the parameters must be jointly distributed. For example, the speed of fuel particles ejected during a thrust termination event will be related to their size under energy considerations. A number of efforts are under way in Navy Theater-Wide TBMD to determine distributions, ranges, or fixed values for the input parameters. While parameter characterization is beyond the scope of this article, it should be clear that the degree of fidelity required in the parameter modeling should be driven, in part, by how each given parameter impinges upon the classification process.

The partition of the engagement parameters into "known" and "unknown" is crucial. The resulting classifiers will be functions of the known parameters. For example, time to intercept and aspect angle might be known during an engagement. Fixing values for the former at 15, 10, and 5 s and for the latter at 4° and 8° will result in a matrix of $3 \times 2 = 6$ classification models with six performance levels. Randomizing these two parameters over their respective ranges of expected values results in a single classifier with lower fidelity (and probably poorer performance) than any of the six more specialized classifiers. However, the

single classifier would be more robust to changes in the parameters. The choice of which training set—and classifier—to use in this instance depends on whether the time to intercept and/or aspect angle will be known during the engagement.

Signature Database: Measurement vs. Simulation

Once the parameters are specified, signatures can be generated. There are ways to populate the signature database: with actual field test measurements, and with simulation outputs. There are advantages and drawbacks to each.

Field test data are believable because they are observed measurements of the specified threat. However, they may be collected under conditions that cannot be extrapolated to tactical scenarios. They may be collected with instrumentation that does not characterize tactical sensors. Finally, they may be sparse with respect to many of the parameter values that could be observed during a tactical launch. A reliable classifier requires a large number of signatures on which to train. As an example, during an intelligence collection exercise, the test radars (Cobra Judy or any of the Kwajalein radars) will be positioned in viewing locations optimal with respect to the threat trajectory. They possess greater sensitivity and resolution than the tactical Aegis/SPY-1 system. Finally, although these radars collect enormous amounts of data, the data are for a single threat realization: the same threat could be launched with different lofts, deployment options, kinematics, and variations in the number and types of debris.

Signatures can be simulated for specified parameter values using kinematic, environmental, and sensor models. The chief shortcoming of simulation is that the model fidelity may be insufficient to capture real-world phenomenology. The advantage is flexibility. Simulations can span the required input parameters, extrapolating even to conditions that have not been observed. Simulations can produce data in sufficient quantity to build a high-confidence classifier.

The two approaches are typically reconciled by developing simulation models that are validated against test data for prescribed cases. The complementary synthesis enjoys credibility without sacrificing flexibility.

Example

To illustrate the steps in building a classifier, an example is drawn from the Navy Phase II COEA. The problem was to classify the T7 RV using the Light Exo-Atmospheric Projectile (LEAP) long-wave IR seeker. This example will focus on the primary discrimination problem: that of distinguishing between the RV and the ACM. For clarity in the presentation, only a few representative signatures will be shown. In a real application of the classification methodology, many sample points are required to realize statistical confidence.

The parameter specifications for the signature generation process are the following:

1. Engagement Geometry

   Defended area, ship 100 km downrange and 100 km crossrange left of projected impact point
2. Threat

   Static Properties. The size, emissivity, mass, and inertia were specified and/or classified. The shape of the RV is a conic. The shape of the ACM is a truncated conic. The geometrical axis of the RV is misaligned 1° with respect to its fundamental inertia axis. (This is a nominal manufacturing tolerance.) The axial symmetry of both objects is useful in determining the torque-free motion of the bodies after boost.

   Rotational Dynamics. The precession and spin periods for both objects were specified. The angular momentum vector for the ACM was randomized over 4π steradians whereas it was discretely parameterized about the trajectory path for the RV.

   Translational Dynamics, Temporal Events. The thrust profile of the unitary threat was known from intelligence, as were the bounds on the tip-off velocities for booster and ACM separation events.
3. Environment

   A number of environmental effects were modeled. Variables impacting temperature—and, therefore, in-band irradiance—were moved in concert to produce upper and lower bounds on the apparent target intensity.
4. Sensor

   The sensor in this example is the LEAP long-wave IR seeker. Both additive and multiplicative noises were modeled. Because the efficacy of candidate features is highly dependent upon the noise model, it is worth detailing the noise assumptions:

$$\hat{I}(t) = [1 + e_{\mathrm{Ex}}(t) + e_{\mathrm{NUC}}(r) + e_{\mathrm{Cal}}]I(t) + e_{\mathrm{NEI}}(t), \quad (1)$$

where $\hat{I}(t)$ = the measured intensity,

$e_{\mathrm{Ex}}(t)$ = residual error due to estimation/extraction (a function of time) $\approx N(0,0.075^2)$,

$e_{\mathrm{NUC}}(r)$ = residual error due to nonuniformity or imperfect knowledge of optical transfer function (a function of location on the sensor) $\approx N(0,0.001^2)$,

$e_{Cal}(r)$ = calibration error as a function of sensor location $\approx N(0,0.1^2)$, and

$e_{NEI}(t)$ = noise equivalent irradiance (NEI) due to shot noise.

The multiplicative noise sources are unitless; the additive noise is expressed in irradiance (W/cm^2).

Figure 1 shows five (of many) representative irradiance time-series waveforms received at the seeker. The irradiance exponents have been obfuscated for security reasons. For this scenario, the intercept occurs approximately 340 s after threat launch. The band is the LEAP long-wave IR band. The environmental parameters were such that the intensity is at the upper bound of what would likely be observed during the scenario. The data are taken at the 20-Hz seeker scan rate; hence, each waveform comprises $5 \times 20 = 100$ data points. Figure 2 shows the same waveforms with seeker noise added. The plots of the RV irradiance waveforms are on a different scale than the ACM irradiance waveforms. (Although they possess similar temperatures and emissivities, the ACM can be brighter than the RV due to its larger projected area.) The plots of each of the five waveforms for each object are spaced out to improve readability; the mean value of each of the four upper waveforms should be centered about that of the first to give the correct absolute irradiance.

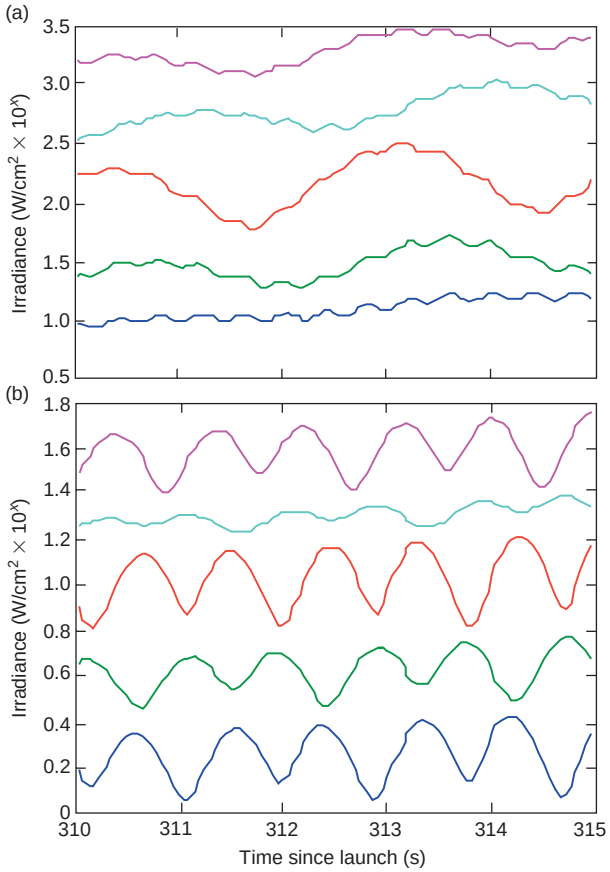

Figure 1. Noise-free irradiances at infrared seeker: long-wave, defended area intercept, upper bound temperature. (a) Reentry vehicle misaligned 1°; (b) nominal attitude control module.

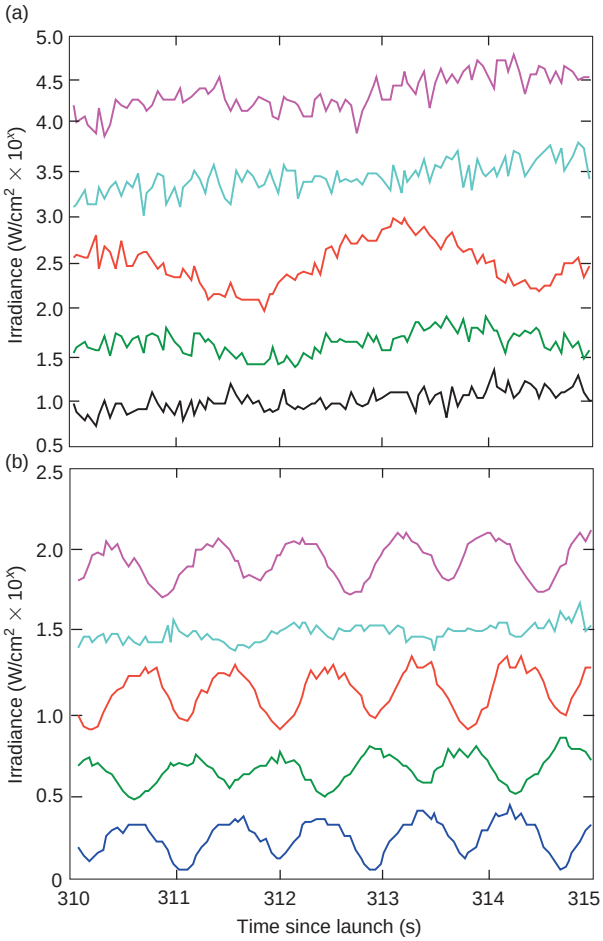

Figure 2. Irradiances with infrared seeker noise (multiplicative and additive): long-wave, defended area intercept, upper bound temperature. (a) Reentry vehicle misaligned 1°; (b) nominal attitude control module.

All four plots evince a positive trend over the 5-s interval due to decreasing range. In Fig. 1a, the RV exhibits two characteristic frequencies: that of the slower precession and faster spin. The spin is observable due to the axis misalignment. The ACM (Fig. 1b) shows one characteristic frequency—its tumble or precession—as it is not spin-stabilized. Neither object presents a pure sinusoid irradiance waveform to the seeker because the physical shapes of the objects modulate the harmonic oscillations of the projected area.

The effect of noise is manifest in a comparison of Figs. 1 and 2. The RV signatures are more corrupted than those of the ACM due to the NEI coupled with the relative dimness of the RV at this range. The primary noise contributer on the brighter ACM signatures is the multiplicative noise.

Abstract Features

The features summarize an epoch of data. They reduce the time series information down to a number or vector of numbers in a prescribed way. There is no pro forma approach to designing features because the underlying objects, processes, and measurement phenomenology are very complex. However, there are two general considerations for a good feature: it should capture the behavior of each class (or at least the class of greatest interest, the RV class), and it should distinguish between classes.

A secondary goal is that the feature be efficient; that is, it should reduce the data as much as possible. A feature is analogous to a sufficient statistic for data drawn from a parametric distribution. For example, 1 million data points that can be shown to be plausibly Gaussian (for example, by the Kolmogorov–Smirnov test) can be reduced to two numbers: the mean and variance. Thus, a feature can be viewed as a mapping of a high-dimensional space (the entire time series of signature data) to one of smaller dimension.

In choosing the features, it is necessary to understand how the object will appear to a particular sensor. The time series signatures will be a function of all the input parameters listed in the previous section entitled Render Signatures. The signatures will vary both with time and from signature to signature. The inter-signature or ensemble variation is due to different input parameter realizations. The temporal variation is both explicit (due to dynamics) and implicit (due to the dependence of the parameters, themselves, upon time). Both the temporal and the ensemble variation can be partitioned into effects due to engagement, environmental, object, and measurement parameters. The first set is somewhat controllable through concept of operations; the second set may be known during an engagement; the third set must be characterized by intelligence, measurement, simulation, and possibly real-time threat typing by other theater assets; and the fourth set is well known and optimized for a given sensor.

The feature should preserve what is unique to the classes of interest; that is, the classes should be invariant under a feature transformation. Some features may be better suited to distilling the time-series information from certain classes than others.

Example (continued)

The uncorrupted waveforms of Fig. 1 suggest features that may be used to classify the RV and the ACM. First, the five RV waveforms (Fig. 1a) have a lower intensity on average than those of the ACM (Fig. 1b). Second, the excursions in intensity with respect to the average value tend to be greater for the ACM than for the RV. Finally, the characteristic frequencies of each object are different: the RV displays two frequencies (around 0.5 and 2.5 Hz) whereas the ACM manifests a fundamental ($\approx$1 Hz) modulated by the object shape and viewing angle. These observations may be reduced to three respective features: mean intensity, scintillation or coefficient of variation (standard deviation divided by the mean), and frequency of maximum amplitude between 0.1 and 3 Hz (both signals will be dominated by the mean energy at 0 Hz). Mathematically,

$$x_1 = E[s(t)], \text{ where } s(t) \text{ is the time domain signal,}$$
$$x_2 = \sigma[s(t)]/\bar{s}(t),$$
$$x_3 = \max_{\omega \in (0.1-3\,\mathrm{Hz})} [\Im\{s(t)\}],$$
$$\bar{x} \equiv (x_1, x_2, x_3).$$

The feature vector $\bar{x}$ reduces the 100 data points of each 5-s epoch to three points. For the noise-free signals, it is clear that $\bar{x}$ will vary from signature to signature. The feature vector will also vary for a given 5-s signature depending on the phasing of the signal. For longer epochs that include more cycles of the signals, the intra-class variability due to phasing is ameliorated. However, to remove all of the signal variability, a feature would have to incorporate a precise estimate of the signal model, itself, as a matched filter. This is clearly impossible in a real-world setting.

For the signals with sensor noise applied, the ensemble variability will be increased, as will that due to phasing. Increasing the length of the measurement epoch will help to "average out" the measurement noise (or to increase the signal-to-noise ratio). However, even perfect measurements will exhibit the intrinsic signal variability discussed.

The spectral characteristics of the RV signatures are shown in Fig. 3a. The absolute amplitude is correct only for the bottom-most plot; the other graphs are spaced for readability. The precession frequency is visible as the fundamental around 0.5 Hz for the five signals. The transforms also show a spectral peak at the higher spin frequency of 2.5 Hz, as would be expected from the time-domain signals.

In pinpointing the location of spectral peaks, a longer observation period will improve the resolution because the frequency resolution is the inverse of the observation period (in this case $\Delta f = 1/(5\text{ s}) = 0.2$ Hz). This is especially important for the low-frequency artifacts: few cycles are observed, and windowing the signal spreads the energy into sidelobes that bleed over into the DC signal for shorter data windows. The obverse is also true: the sidelobes from the DC signal can obfuscate the low-frequency signals. The shorter the observation period, the farther out the sidelobes will spread. Hence, a prosaic but vital first step is

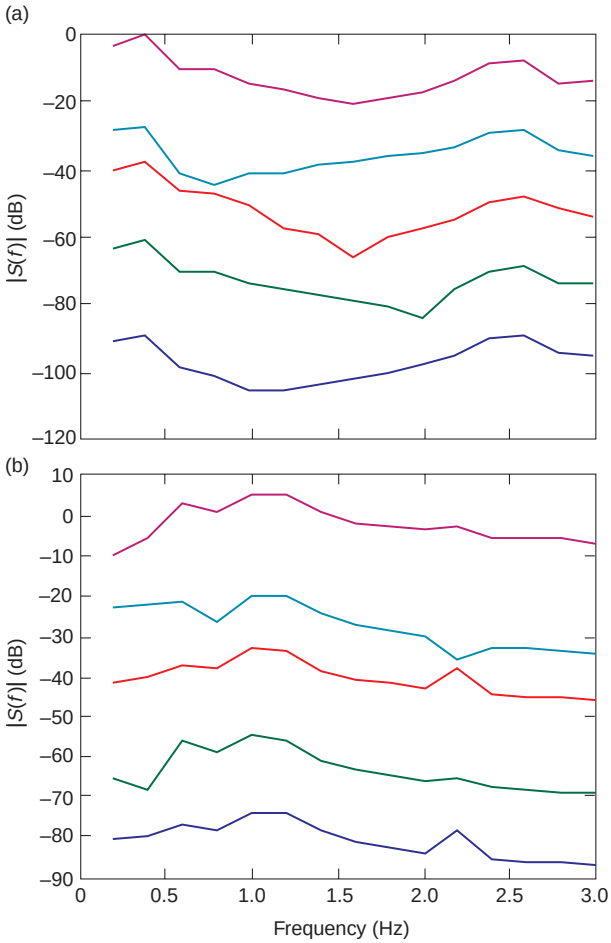

Figure 3. Spectra of (a) reentry vehicle and (b) nominal attitude control module time series, without sensor noise.

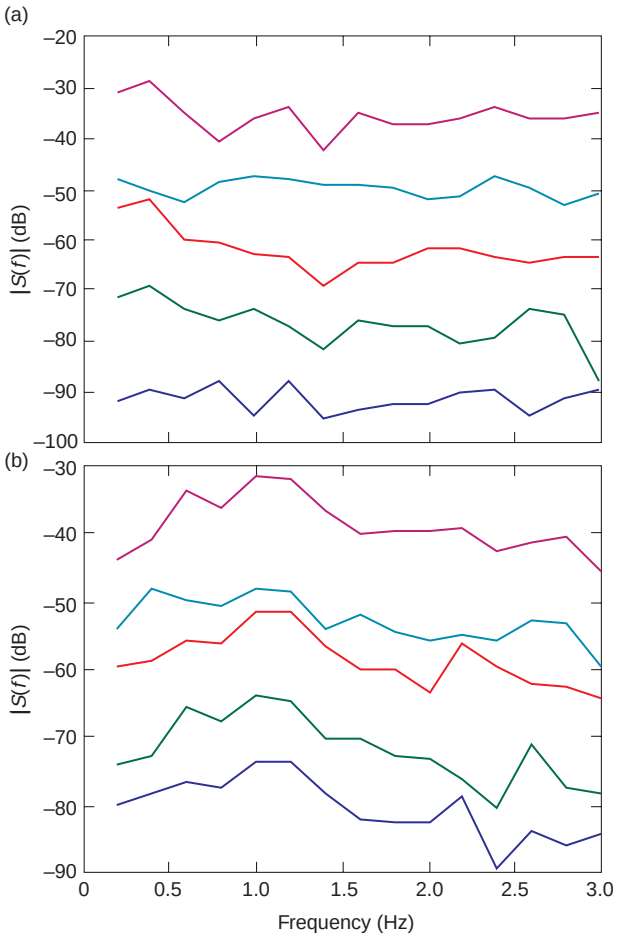

Figure 4. Spectra of (a) reentry vehicle and (b) nominal attitude control module time series, with sensor noise.

removing at least the mean and possibly linear trends before transforming the time-domain signal.

Another method would be to model the spectrum by estimating the autoregressive coefficients of the time sequence using robust deconvolution. This technique provides very sharp resolution at the expense of spurious peaks for overestimated model order.

The spectrum of the ACM in Fig. 3b shows the anticipated spectral peak around 1 Hz for all five waveforms, along with the first harmonic at twice the frequency for four of the five. The harmonic shows up because the ACM time signals are not pure sinusoids: the simple harmonic motion of the body is multiplied by the envelop function of the body's projected area. Multiplication in the time domain is equivalent to periodic convolution in the frequency domain, hence, the harmonic.

The spectra of the noisy signals are shown in Fig. 4. The low-frequency fundamental appears to be lost in two of the five signals in the RV ensemble due to additive noise. The more intense ACM signals suffer less degradation, and the fundamental is preserved.

Model Class-Dependent Feature Distributions

During training with either real or simulated data, the class to which a given time series of data belongs is known. It is necessary to characterize what each class of data "looks like" in the feature space. These class-dependent feature distributions will determine the ability of the system to discriminate between objects of different classes. The distributions will be conditioned upon parameters that will be known during the engagement as well as upon class. Conditioning on a known parameter removes the variability due to that parameter and improves the inter-class separability. Modifying the notation in Ref. 1, the class-dependent feature distributions can be written $p(\bar{x} \mid \omega_i, \bar{\theta})$, where

$\bar{x} \in X \equiv \{\bar{x}_l\}_{l=1}^{N}$ is the set of particular features chosen from the set of all features X,

$\omega_i \in \Omega \equiv \{\omega_i\}_{i=1}^{K}$ is a particular class from the set of all K classes of objects in the complex, and

$\bar{\theta}$ is a set of parameters that are known during the engagement.

Parametric vs. Nonparametric Models

Because of the difficulty in obtaining closed-form expressions, the class-dependent feature distributions $\{p(\bar{x} \mid \omega_i, \bar{\theta})\}_{i=1}^{K}$ will be obtained empirically from simulation or test data or a combination thereof. There is a vast literature to address the problem of choosing a distribution to summarize the data in an empirical distribution function (edf). The way in which the feature distribution is modeled is inextricably interwoven with how the decision boundaries for each class are constructed (see the section entitled Decision Boundaries). This is because the classification problem is observing the features for a number of objects in the seeker field of view (FOV), then deciding which distributions (or which class) those features most likely came from.

Approaches to modeling the class-dependent feature distributions can be grouped into two broad categories: parametric and nonparametric.

Parametric models involve fitting the edfs to distributions of a certain functional form. The parameters of the distribution are estimated subject to certain criteria; for example, best least squares fit to the data, or maximum likelihood. By far the most popular parametric model is the Gaussian. A Gaussian parametric assumption fits the scatter of feature points for a given class with a mean and a covariance. In the M-dimensional feature space, the model will be a hyper-elliptical cluster about the mean.

The approach taken in the Phase II COEA was to model each class with a Gaussian in the feature space. The Gaussian parameters were simply the sample statistics. This is illustrated for the example by modeling the edfs in [feature (3,2)]-space in Fig. 5c with the ellipses in Fig. 6. These are the third and second features described previously.

The advantages to parametric models include the following:

1. The notion of distance—and, hence, inter-class separability—is well defined. This is discussed in the section entitled Measuring Feature Efficacy.
2. The scatter characteristics are concisely summarized in the parameters of the distribution. In the same way that a feature was a sufficient statistic for the distinguishing elements of the time series data in the section entitled Abstract Features, the parameters are sufficient statistics for the feature realizations.
3. The distributions are mathematically tractable, often with closed-form expressions or tabulated data for the cumulative distribution function, which is useful in ascertaining error probabilities.

The chief drawback to parametric modeling is loss of fidelity. A single parametric model does not usually capture the scatter in an edf. (In the example of Fig. 6, the ellipses do not reflect the asymmetry of the scatter.) There are a wealth of techniques for assessing the goodness of fit for data to given distributions. A poor fit can be remedied somewhat by introducing a mixture model to refine the fidelity for a given class. However, improved fidelity comes at the expense of computational complexity and storage requirements. The classification models must be stored in look-up tables and evaluated in real time by the interceptor's onboard processor.

No systematic attempt will be made to describe nonparametric approaches to building classifier models. The trade-off between parametric and nonparametric models can be broadly characterized in terms of robustness versus fidelity. As they are not beholden to any particular assumptions on the class-dependent feature edfs, nonparametric models can capture more of the variability of the given classes in the feature space. This comes at the expense of perhaps capturing too much of the variability. With their ability to draw almost arbitrarily convoluted boundaries between the classes of objects in the feature space, neural network and logistic regression models must not be overtrained on the feature data. Overtraining is said to occur when the decision boundaries are fit too specifically to the training set. This makes the classifier performance susceptible to noise or small perturbations in the training data. One way to guard against overtraining is a jackknife training protocol: the classifier is trained against subsets of the data. A properly trained classifier should be robust to selection of the subset.

Measuring Feature Efficacy

For specified classes ω_i (say, the target RV class) and ω_j (the non-RV or ACM class), and fixed input parameters $\bar{\theta}$, a feature vector $\bar{x}$ can be used to make a decision between classes using the likelihood ratio: $p(\bar{x} \mid \omega_i, \bar{\theta}) / p(\bar{x} \mid \omega_j, \bar{\theta})$. If this value is large for a given object, the object probably belongs to class ω_i; if it is small, it is probably of class ω_j. Since a monotonic function will preserve inequality, the decision statistic between classes ω_i and ω_j is often taken to be $d_{ij}(\bar{x}) \equiv \log[p(\bar{x} \mid \omega_i) / p(\bar{x} \mid \omega_j)]$, where the input parameters $\bar{\theta}$ are understood.

For the exponential family of distributions (which includes the Gaussian), this transformation allows the distribution of d_{ij} to be computed analytically. The distribution for d_{ij}, whether empirical or analytic, in turn allows for the computation of probability of error for classification. For independent binary classification of an object into ω_i against ω_j, a hypothesis test is used. For a set threshold T, the probabilities of false alarm

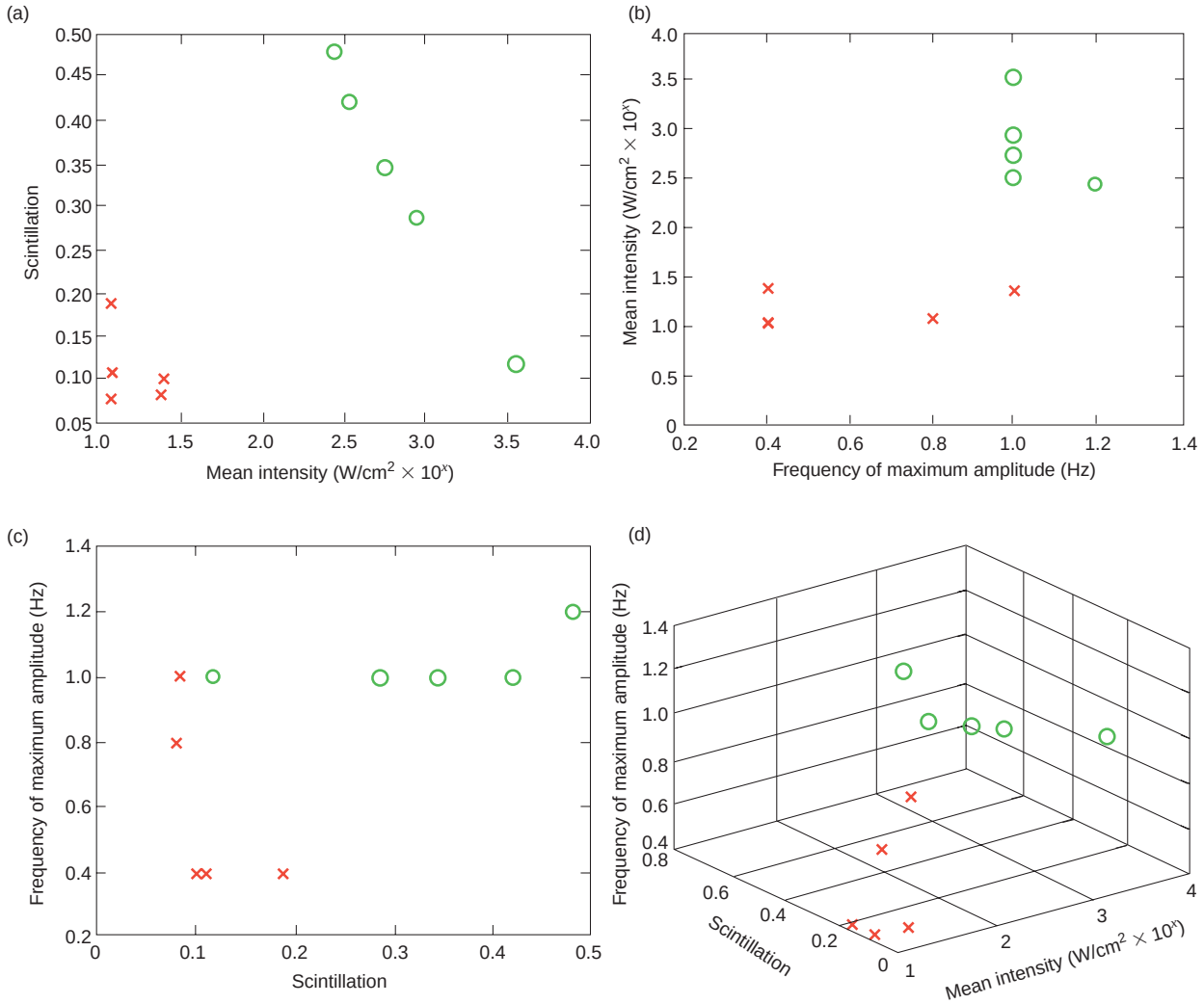

Figure 5. Feature realizations for the reentry vehicle (×) and attitude control module (o) time series signals. (a) (feature 2) vs. (feature 1); (b) (feature 1) vs. (feature 3); (c) (feature 3) vs. (feature 2); and (d) (feature 3) vs. (feature 1,2).

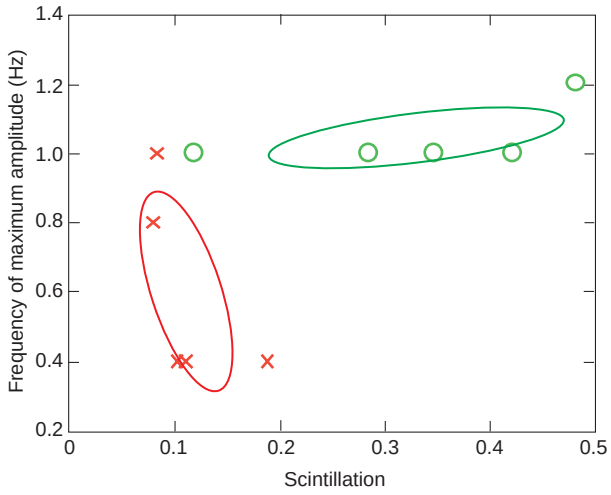

Figure 6. Gaussian models for the reentry vehicle and attitude control module classes in [feature (3,2)]-space.

(calling a nontarget a target) and miss (calling a target a nontarget) are given as follows:

$$P_{\mathrm{FA},ij}(T) = P(d_{ij} < T \mid \omega_j) = \int_{\{\bar{x}:d_{ij}(\bar{x})<T\}} p(\bar{x} \mid \omega_j)\,d\bar{x}\ ,$$

$$P_{\mathrm{Miss},ij}(T) = P(d_{ij} < T \mid \omega_i) = \int_{\{\bar{x}:d_{ij}(\bar{x})<T\}} p(\bar{x} \mid \omega_i)\,d\bar{x}\ . \quad (2)$$

The implicitly defined limits of this integral can be difficult to evaluate for continuous probability distributions. In practice, however, $p(\bar{x} \mid \omega_i)$ is an edf. In the discrete case, these expressions are straightforward. For I N-dimensional realizations of $\bar{x}$ for the ω_i class and J realizations for the ω_j class,

$$P_{\mathrm{FA},ij}(T) = \frac{1}{J}\left(\sum_{k=1}^{J} \delta[d_{ij}(\bar{x}_k \mid \omega_j) < T] \right),$$
$$P_{\mathrm{Miss},ij}(T) = \frac{1}{I}\left(\sum_{k=1}^{I} \delta[d_{ij}(\bar{x}_k \mid \omega_i) < T] \right), \quad (3)$$

where

$\delta(\bullet)$ = 1 if the expression is true
= 0 if the expression is false.

As stated before, the measure of quality for a feature (or set of features) is twofold: the degree to which it tightly clusters given classes (especially the class of interest, the RV class), and the degree to which it separates classes. Addressing the latter criterion, a natural measure is the expected difference between the decision statistic d_{ij} for objects of class i versus those of class j. The difference between the conditional means is written as

$$\begin{aligned} D_{ij} &= E[d_{ij}(\bar{x}) \mid \omega_i] - E[d_{ij}(\bar{x}) \mid \omega_j] \\ &= E_{\omega_i}[d_{ij}(\bar{x})] - E_{\omega_j}[d_{ij}(\bar{x})] \\ &= \int [p(\bar{x} \mid \omega_i) - p(\bar{x} \mid \omega_j)] \log \frac{p(\bar{x} \mid \omega_i)}{p(\bar{x} \mid \omega_j)} d\bar{x} . \end{aligned} \quad (4)$$

This is the "information divergence" discussed in Ref. 1.

The intra-class dispersion criterion can be captured by computing the variance of the decision statistic for a given class:

$$\begin{aligned} V_i &= Var[d_{ij}(\bar{x}) \mid \omega_i] \\ &= \int \{d_{ij}(\bar{x}) - E[d_{ij}(\bar{x}) \mid \omega_i]\}^2 p(\bar{x} \mid \omega_i) d\bar{x} . \end{aligned}$$

In practice, it is straightforward to choose the best feature set for one, two, or three dimensions simply by looking at scatter plots of the different classes rendered in candidate feature sets. However, it is desirable to have a constructive approach for higher dimensions. With the separation and dispersion criteria, potential features can be evaluated. The best scalar feature $x_k \in \{x_l\}_{l=1}^{N}$ for distinguishing between two classes ω_i and ω_j is given by

$$x_k = \sup_l \left(\frac{D_{ij}(x_l)}{\sqrt{V_i(x_l) + V_j(x_l)}} \right). \quad (5)$$

It seems intuitive that adding features can only improve inter-class separability as long as the features added individually provide some separability. This can be shown formally using the information divergence. Less obvious, however, is what the choice of the best set of features should be for a specified feature set dimension. The best M-dimensional feature vector is seldom the set of the first M of N rank-ordered possible features. This is because good features will often be correlated, or contain the same information. There will be $\binom{N}{M}$ possible feature vectors. Choosing the best feature in a procedure analogous to Eq. 5 is then a problem of combinatorial computational complexity. This problem can be relieved somewhat by performing an eigenvalue analysis of $p(\bar{x}_{(N)} \mid \omega_i), p(\bar{x}_{(N)} \mid \omega_j)$ to determine the principal values and principal vectors of the class-dependent feature distributions. The objective is to find independent linear combinations of the available features $\{\bar{x}_l\}_{l=1}^{N}$ that maximize inter-class separability while minimizing intra-class dispersion. The subscript (N) above denotes an N-dimensional feature vector.

Capturing Uncertainty in Feature Models

The feature distributions will comprise all that is known about the given class for the specified parameters. Each distribution will constitute a number of discrete realizations in the feature space, from either actual or simulated data, or both. The variability in the realizations will be due to both the ensemble and the temporal variation in the underlying signatures. This variation should be fully captured by independently sampling enough realizations: there is no "confidence interval" either for a given realization or for the feature distribution for a whole. All of the uncertainty should be contained in the sampling for the parameters used to produce the signatures. For example, if the value for ambient launch temperature is suspect, it should be varied during signature production. Variability cannot be "added" to the feature distribution after the fact. This is because of the transformations the parameters undergo:

$$\{\text{parameters}\} \rightarrow \{\text{signatures}\} \rightarrow \{\text{features}\}.$$

Each one of these transformations is, in general, nonlinear, so that even if the input parameter distributions are well characterized, a closed-form expression of the transformed distribution is not attainable. If it were possible, there would be no reason to simulate the process: the class-dependent feature distributions $p(\bar{x} \mid \omega_i)$ discrimination models could be built analytically.

Independent Sampling

An important question in building the class-dependent feature distributions is what constitutes independent samples of the feature vector. Double-counting the sample points will result in a feature distribution that is more tightly clustered than it should be. This gives undue confidence about the separation achieved between object classes. On the other hand, undersampling fails to take advantage of the information contained in the signature training sets.

The causes of variability in the feature realizations can be divided into those due to parameters underlying the time series signatures and those due to the phasing of each individual signature. This latter temporal variation can be further parsed into the deterministic variation due to which epoch of the noise-free signature is chosen, and the random variation due to additive and multiplicative noise.

Achieving independence with respect to the parameter set is straightforward: the underlying parameters are simply sampled randomly and independently prior to generating the time series signal. For example, to model the IR signatures of solid fuel chunks, each signature was generated from a random angular rate and angular momentum vector orientation.

Correctly reproducing the sample statistics due to temporal variability is more problematic. For wide-sense stationary Wiener or Markov processes driven by white noise, the typical criterion is when the autocorrelation $R_{xx}(\tau) \equiv E[X(t)X(t+\tau)]$ has decayed to $1/e$ of its maximum value. (Thus, a narrow autocorrelation function signifies a process that is relatively uncorrelated in time.) For ascertaining feature independence, this is problematic for two reasons. First, even if the autocorrelation of the input signal $R_{ss}(\tau)$ is known, the feature autocorrelation cannot be determined due to the nonlinearity of the feature processing. This means that the sampling interval τ would have to be determined empirically, perhaps for every signature. Second, the signals are not random; they are periodic. The torque-free motion of rigid bodies in space produces IR intensity sinusoids that are modulated by the shape of the objects. Therefore, the autocorrelation of the signal will evince a number of periodic peaks.

These two confounding factors mean that it is best to deal with temporal variability in the feature space. The rule of thumb to use is that clustering of sample points about a given signature realization should not be evident in the feature space. Another way to put this is that the variability due to temporal sampling should replicate that due to parameter sampling. This criterion can be formalized using cluster analysis, but is probably best attacked empirically by looking at scatter plots of the feature data.

Example (continued)

Each of the five time series signatures for the ACM and the RV were reduced to the three features described above: mean intensity, scintillation, and frequency of maximum amplitude. The three features for each class are plotted in Fig. 5.

The features are plotted two at a time in Figs. 5a–5c, and all three are plotted in Fig. 5d. The efficacy of the univariate features can be surmised by projecting the two-dimensional features to the feature axis.

Figures 5a and 5b show that the mean intensity is a good feature: it separates the two classes with no overlap. Furthermore, the separation distance is large with respect to spread for each individual class. The ability of this feature to separate the two types of objects could be anticipated by realizing that the two objects are at nearly the same temperature, whereas the ACM is considerably larger.

Scintillation is also a good feature, with only one of the ACM realizations straying into the RV area in Figs. 5a and 5c. This ACM realization fortuitously illustrates the power of multidimensional feature models: it is well-separated in the two-dimensional space. The efficacy of the scintillation feature arises from the spin stabilization of the RV in contrast to the tumbling ACM. Another noteworthy aspect of Fig. 5a is the nonlinear correlation especially evident in the ACM class. This is because the scintillation is inversely proportional to the mean intensity ($x_2 \propto 1/x_1$). The almost perfect correlation of the ACM class is uncommon; more typical is the clustering of the RV class.

Frequency of maximum amplitude is shown to be a useful feature in Figs. 5b and 5c. Only one of the RV class realizations penetrates the ACM boundaries. However, the intra-class dispersion of the RV class does not give tremendous confidence. The two upper samples could be misclassified as ACM-like with respect to this feature. A greater number of samples might reveal these two points to be outliers. However, the spectra of Figs. 4a and 4b show that the RV frequency components are more susceptible to additive sensor noise than those of the ACM because the RV is relatively dimmer (due to its smaller size). Hence, the spectral feature will be of greatest use in concert with other features as shown in Figs. 5b, 5c, and 5d. It should be clear that the signal-to-noise ratio (SNR) will impinge upon the features. However, the dependency is implicit in the feature scatter. On the other hand, SNR is explicitly functionalized in the track initiation logic upon which the feature epochs are predicated.

As discussed earlier, the feature edfs for the Phase II COEA were reduced Gaussian parametric distributions using the sample statistics of each class. The resulting models for the classes projected onto the

(feature 3)-(feature 2) plane are plotted in Fig. 6. These are the 1-sigma ellipses.

Decision Boundaries

Decision boundaries can be used in two ways. For a hypothesis test for classifying an object into one of two possible classes, the decision boundary is set according to the desired probabilities of false alarm and leakage (or miss). The boundary is static.

As discussed in the section entitled The TBMD Classification Problem, the RV classification problem is of a different sort: a number of objects must be classified, and only one of them can be the RV. The object selected will be that which is most RV-like. This requirement translates formally to finding the object in the feature space that is on the decision contour closest to the RV class.

Nonparametric techniques are well-suited to developing static boundaries with low, fixed probabilities of error. This is because information in the feature edfs is not lost in reducing the data to parametric distributions. Another way to see this is to realize that complex hyper-surfaces can be fashioned with simple functions to separate the feature realizations of each class. The literature on neural networks and regression on various functions is voluminous.

Parametric techniques have an advantage in choosing the decision contour closest to a given class in that the notion of "close" is well-defined: it is simply the probability that a particular realization belongs to the class. For the Gaussian assumption, each class is represented by a hyper-ellipse. The closeness to a given class is established by the k-sigma ellipse upon which a point lies. For the two-class problem, a given point in the feature space will lie on the k_1-sigma ellipse of the first class and the k_2-sigma ellipse of the second class. A decision contour will then constitute the set of all such points for k_1, k_2 fixed. It is established by the log likelihood:

$$
\begin{aligned}
d_{ij}(\bar{x}) &\equiv \log\left[\frac{p(\bar{x}\,|\,\omega_i)}{p(\bar{x}\,|\,\omega_j)}\right] \\
&= -(\bar{x}-\bar{\mu}_i)^T \sum_i^{-1} (\bar{x}-\bar{\mu}_i) + (\bar{x}-\bar{\mu}_j)^T \\
&\quad \times \sum_j^{-1} (\bar{x}-\bar{\mu}_j)
\end{aligned}
\qquad (6)
$$

for the Gaussian assumption.

It is clear from Eq. 6 that the Gaussian parametric assumption induces a quadratic decision space. The decision contours will be ellipsoids, hyperbolas—and in degenerate cases, parabolas and lines—in the feature space. Contours are sketched for a two-dimensional feature space in Fig. 7 for class ellipses in various relative locations and orientations. Here the suboptimality of the Gaussian assumption becomes apparent: the only curves that can be drawn to separate the two-class edfs are quadratic curves. This suboptimality results in larger probabilities of error. However, the Gaussian assumption does not engender any uncertainty about these resulting probabilities of error; they can be computed exactly, as will be shown in the next section. The only "error" in the probabilities of error is due to failure of the feature edfs to capture the true variability of the threat in the feature space.

Example (continued)

The feature realizations and models of Fig. 6 are reprised in Fig. 8 with the iso-contour of each realization sketched. As the contours make clear, the probability of error is zero for this simple training set. All of the RVs are more RV-like than the most RV-like ACM. The converse obtains as well.

Probability of Error

This section develops the mathematics necessary to compute the probability of classification error. The probability of error is the yardstick by which any proposed system design—comprising signal processing, feature set, and decision algorithm—must be measured. Although the notation can be cumbersome, the concepts are clarified in the associated figures. The vital insights into what drives classification error are evident in the two-class problem described thus far and continued below. The sections on the multi-class problem are included for completeness and can be skimmed or omitted.

It is clear that the probability of error in classification will be contingent upon the inter-class variability and the intra-class dispersion. For the problem of classifying an object into one of two classes, the operating curve is the plot of $P_{\mathrm{FA}}(T)$ against $P_{\mathrm{Miss}}(T)$. For the edfs $p(\bar{x}\,|\,\omega_i), p(\bar{x}\,|\,\omega_j)$, this will be a plot of $I + J$ points of Eq. 3, where T is computed for each of the $\bar{x}$ realizations for each class. This curve is reduced to a scalar by finding the point T_{EER} such that $P_{\mathrm{FA}}(T_{\mathrm{EER}}) = P_{\mathrm{Miss}}(T_{\mathrm{EER}}) \equiv P_{\mathrm{EER}}$. This point is the equal error rate (EER) for the two distributions. The equal error rate is often used interchangeably with the k-factor, although the equivalence requires stringent parametric assumptions. If $p(\bar{x}\,|\,\omega_i), p(\bar{x}\,|\,\omega_j)$ are Gaussian with means $\bar{\mu}_i, \bar{\mu}_j$ and identical variances $\Sigma_i = \Sigma_j = \Sigma$, then the information divergence (Eq. 4) can be written as

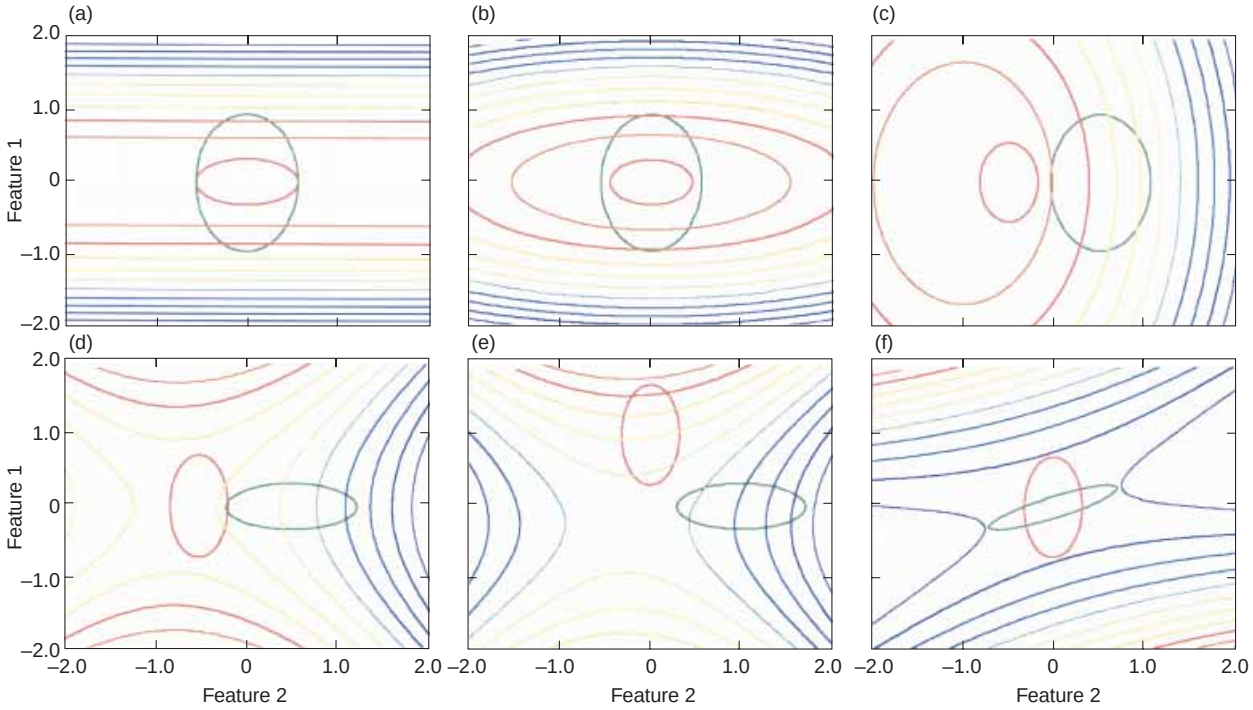

Figure 7. Binary decision boundaries for two-dimensional Gaussian models.

$$
\begin{aligned}
D_{ij} = |\, & E_{\omega_i}[-(\bar{x}-\bar{\mu}_i)^T \Sigma^{-1}(\bar{x}-\bar{\mu}_i) \\
& + (\bar{x}-\bar{\mu}_j)^T \Sigma^{-1}(\bar{x}-\bar{\mu}_j)] \\
& - E_{\omega_i}[-(\bar{x}-\bar{\mu}_i)^T \Sigma^{-1}(\bar{x}-\bar{\mu}_i) \\
& + (\bar{x}-\bar{\mu}_j)^T \Sigma^{-1}(\bar{x}-\bar{\mu}_j)]| \\
= |\, & E[-X^2(M) + X^2(M,\rho_j)] \\
& - E[-X^2(M,\rho_i) + X^2(M)]| \\
= |\, & -M + (M+\rho_j) - (M+\rho_i) + M| \\
= |\, & \rho_j - \rho_i\,| .
\end{aligned}
$$

Here M is the dimension of the feature vector and ρ_i, ρ_j are the noncentrality parameters for the chi-squared distribution: $\rho_i = \bar{\mu}_i^T \Sigma^{-1} \bar{\mu}_i$. Then

$$
\begin{aligned}
D_{ij} &= (\bar{\mu}_j - \bar{\mu}_i)^T \Sigma^{-1} (\bar{\mu}_j - \bar{\mu}_i) \\
&\equiv K_{ij}^2 .
\end{aligned}
$$

For scalar x this reduces to

$$
K_{ij} = \frac{|\mu_j - \mu_i|}{\sigma},
$$

as noted in Ref. 1. The k-factor K_{ij} can be looked up in a standard table of unit normal p-values. The associated p-value is the equal error rate.

Shortcomings of the k*-Factor and Equal Error Rate*

The k-factor is a qualitative measure of error in a binary classification process. It can only be used quantitatively if the class-dependent feature distributions

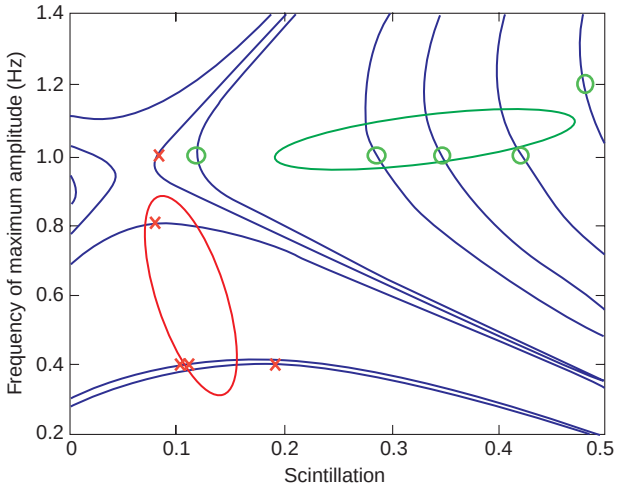

Figure 8. Decision boundaries for the reentry vehicle and attitude control module classes in [feature (3,2)]-space.

are Gaussian with equal variances. This is rare. Stipulating this assumption allows the equal error rate to be looked up. However, the equal error rate is the probability of error only for a binary hypothesis test with the threshold T set at the equal error level. As discussed in the section entitled The TBMD Classification Problem, the TBMD classification problem is not a binary hypothesis test with a preset threshold. Rather, of all the objects in the FOV, the classifier will target that object whose feature vector lies on the decision threshold closest to the RV class.

Classifying 1 RV vs. 1 Non-RV

Assume there are two objects in the FOV, one from the RV class $\omega_i, i \neq j$, and one from a non-RV class ω_j. Denote the RV feature observation $\overline{X}_i$ and the non-RV feature observation $\overline{X}_j$. Both are random variables. A classification error occurs if $d_{ij}(\overline{X}_i) < d_{ij}(\overline{X}_j)$. The probability of error is the probability that this occurs. This probability is written as

$$\begin{aligned} P_{\text{Error}} &= \iint\limits_{\{(x_i, x_j): d_{ij}(\overline{x}_i) < d_{ij}(\overline{x}_j)\}} p(\overline{x}_i, \overline{x}_j \mid \omega_i, \omega_j) d\overline{x}_i d\overline{x}_j \\ &= \iint\limits_{\{\bullet\}} p(\overline{x}_i \mid \omega_i) p(\overline{x}_j \mid \omega_j) d\overline{x}_i d\overline{x}_j . \end{aligned} \tag{7}$$

Here the joint distribution can be written as the product of the marginals under the assumption of independence. In reality, the motions of one object in the ballistic complex will not be independent of the motions of others. For example, the spin rate with which the ACM deploys the RV will affect its own kinematics by Newton's laws and the residual fuel remaining for other maneuvers. However, determining the parameter relationships to seed Monte Carlo simulations to produce joint distributions is beyond the current state of intelligence.

Equation 7 is an integration in $2M$-dimensional feature space. The implicit limits of integration render it very difficult to evaluate either for empirical or for analytic distribution functions. The probability of error integration must be transformed to a tractable domain. Denote the distribution of $d_{ij}(\overline{X}_i)$ as $p_{d_{ij}}(d \mid \omega_i)$, and that of $d_{ij}(\overline{X}_j)$ as $p_{d_{ij}}(d \mid \omega_j)$. If the feature distribution $p_{d_{ij}}(d \mid \omega_j)$. is specified analytically, in principle it can be transformed to yield the corresponding decision distribution $p_{d_{ij}}(d \mid \omega_i)$. In practice, these analytic transformations often involve a number of convolutions that result in integrals with obscure closed-form solutions or none.

The utility of the $p_{d_{ij}}(d \mid \omega_i), p_{d_{ij}}(d \mid \omega_j)$ distributions is twofold: they have scalar domains, and the decision region for either class is one-sided. If the transformation is feasible, Eq. 7 can be rewritten with straightforward limits of integration:

$$P_{\text{Error}} = \int_{-\infty}^{\infty} \int_{-\infty}^{d_j} p_{d_{ij}}(d_i) p_{d_{ij}}(d_j) \partial d_i \partial d_j . \tag{8}$$

Here $p_{d_{ij}}(d \mid \omega_i)$ has been rewritten as $p_{d_{ij}}(d_i)$, and the total differential d has been replaced by ∂ to avoid disagreeable notation. The difficulty in integrating Eq. 7 has been relieved in Eq. 8 at the expense of transforming $p(\overline{x} \mid \omega_i)$ to obtain $p_{d_{ij}}(d \mid \omega_i)$. However, if $p(\overline{x} \mid \omega_i)$ is an edf, $p_{d_{ij}}(d \mid \omega_i)$ can be obtained very simply by applying the d_{ij} operator to each element of $p(\overline{x} \mid \omega_i)$. Then Eq. 8 can be rewritten as

$$P_{\text{Error}} = \frac{1}{IJ} \left(\sum_{k_2 = 1}^{J} \sum_{k_1 = 1}^{I} \delta[d_{ij}(\overline{x}_{k_1} \mid \omega_i) < d_{ij}(\overline{x}_{k_2} \mid \omega_j)] \right), \tag{9}$$

where $\delta(\bullet)$was defined in Eq. 3. This is effectively a Monte Carlo integration.

The form of the probability of error in Eqs. 8 and 9 suggests a relationship to the probabilities of miss and false alarm in Eqs. 2 and 3. Rewriting Eq. 8,

$$\begin{aligned} P_{\text{Error}} &= \int_{-\infty}^{\infty} p_{d_{ij}}(d_j) \int_{-\infty}^{d_j} p_{d_{ij}}(d_i) \partial d_i \partial d_j \\ &= \int_{-\infty}^{\infty} p_{d_{ij}}(d_j) P_{\text{Miss}}(d_j) \partial d_j \\ &= E_{\omega_j} \{ P_{\text{Miss}}(d_j) \} . \end{aligned} \tag{10}$$

The d_j in the last line is a random variable; it is left lower case to avoid confusion with the information divergence. Equation 10 shows that the probability of error for picking the RV against a single non-RV is simply the expected probability of miss given the decision distribution for the non-RV, $p_{d_{ij}}(d \mid \omega_j)$. This seems intuitive. The converse can also be shown:

$$P_{\text{Error}} = E_{\omega_i} \{ P_{\text{FA}}(d_i) \} . \tag{11}$$

The probability of error can now be related to the equal error rate P_{EER}. Equation 10 can be rewritten

$$
\begin{aligned}
P_{\text{Error}} &= \int_{-\infty}^{T} p_{d_{ij}}(d_j) \int_{-\infty}^{d_j} p_{d_{ij}}(d_i)\partial d_i \partial d_j \\
&\quad + \int_{T}^{\infty} p_{d_{ij}}(d_j) \int_{-\infty}^{d_j} p_{d_{ij}}(d_i)\partial d_i \partial d_j \\
&= \int_{-\infty}^{T} p_{d_{ij}}(d_j) \int_{-\infty}^{d_j} p_{d_{ij}}(d_i)\partial d_i \partial d_j \\
&\quad + \left(\begin{array}{l} \int_{-\infty}^{T} p_{d_{ij}}(d_i) \int_{T}^{\infty} p_{d_{ij}}(d_j)\partial d_i \partial d_j \\ + \int_{T}^{\infty} p_{d_{ij}}(d_i) \int_{d_i}^{\infty} p_{d_{ij}}(d_j)\partial d_j \partial d_i \end{array} \right) \\
&= \int_{-\infty}^{T} p_{d_{ij}}(d_j) P_{\text{Miss}}(d_j)\partial d_j \\
&\quad + P_{\text{Miss}}(T)P_{FA}(T) + \int_{T}^{\infty} p_{d_{ij}}(d_i) P_{\text{FA}}(d_i)\partial d_i \\
&< P_{\text{Miss}}(T) \int_{-\infty}^{T} p_{d_{ij}}(d_j)\partial d_j + P_{\text{Miss}}(T)P_{\text{FA}}(T) \\
&\quad + P_{\text{FA}}(T) \int_{T}^{\infty} p_{d_{ij}}(d_i)\partial d_i \\
&= P_{\text{Miss}}(T)[1 - P_{\text{FA}}(T)] + P_{\text{Miss}}(T)P_{\text{FA}}(T) \\
&\quad + P_{\text{FA}}(T)[1 - P_{\text{Miss}}(T)] \\
&= P_{\text{Miss}}(T) + P_{\text{FA}}(T) - P_{\text{Miss}}(T)P_{\text{FA}}(T).
\end{aligned}
\tag{12}
$$

This is true for any decision distributions $p_{d_{ij}}(d_i)$, $p_{d_{ij}}(d_j)$ and any decision threshold T. Choosing $T = T_{\text{EER}}$, Eq. 12 becomes

$$
\begin{aligned}
P_{\text{Error}} &< P_{\text{Miss}}(T_{\text{EER}}) + P_{\text{FA}}(T_{\text{EER}}) \\
&\quad - P_{\text{Miss}}(T_{\text{EER}})P_{\text{FA}}(T_{\text{EER}}) \\
&< 2P_{\text{EER}}.
\end{aligned}
\tag{13}
$$

If the decision distributions are unimodal, Eq. 12 can be shown to be

$$
\begin{aligned}
&P_{\text{Error}} < \frac{1}{2}[P_{\text{Miss}}(T) + P_{\text{FA}}(T) - P_{\text{Miss}}(T)P_{\text{FA}}(T)] \\
&\Rightarrow P_{\text{Error}} < P_{\text{EER}}.
\end{aligned}
\tag{14}
$$

This is an important finding:

The equal error rate will be an upper bound on the probability of error for a binary classification decision between one object from each class.

This analysis has assumed that each object is equally probable (that is, there is one of each object.) Typically the equal error rate will be a loose bound; the probability of error will usually be much smaller. The bound becomes tighter as the feature distributions for the two classes move closer together, that is, as the equal error rate approaches 0.5.

Example

Figure 8 shows that for the given distributions there is no probability of error in misclassifying members of either class. This is because the most RV-like of the ACM realizations (each denoted by a circle) has a lower RV likelihood than the least RV-like of the RV realizations (denoted by ×) as shown by the likelihood contours drawn through each realization. Therefore, to illustrate how probabilities of error are derived, the example proceeds with the feature models from the sketch in Fig. 7f. Data were generated from the models used in Fig. 7: 6 data points from the RV class and 10 from the ACM class. The sample statistics for each class were calculated. These models were then used to form the decision statistic for the classification. As in Fig. 8, the iso-contours of the decision statistic are drawn through each point in Fig. 9. As in Fig. 7, each contour comprises two halves: one half is the image of the other reflected about an axis of symmetry. Thus, the contour lines that appear to have no associated feature realization are in fact reflections at the same value of the likelihood ratio.

At each feature realization of either class, the probability of leakage (or miss) is the percentage of RV realizations that are more ACM-like. This is the number of RV realizations that lie on the ACM side of the given realization divided by the total number of RV

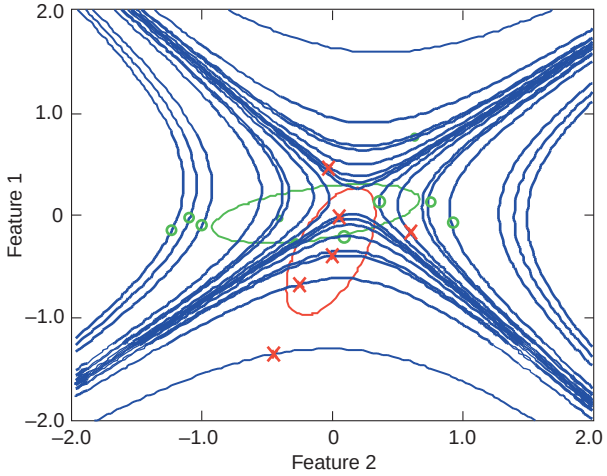

Figure 9. Generating probabilities of error for 1-vs.-1 classification: feature realizations and decision contours.

realizations (six). Similarly, the probability of false alarm is the percentage of ACM realizations that are more RV-like at the decision threshold of the given realization. These are the discrete probabilities of Eq. 3.

Figure 9 illustrates three points about the significance of the number of realizations chosen to each class-dependent feature distribution.

1. The points characterize the distributions. The greater the number of independent samples, the more reliably the distribution is characterized. The models used to generate the samples in Fig. 9 were those in Fig. 7f, yet the sample statistics clearly do not reproduce the models due to the limited sample size. However, the ACM sample statistics, drawn from 10 points, better reflect the underlying model than the RV sample statistics, which are based on only 6 points.
2. The number of points chosen to characterize each class has nothing to do with the scope of the classification engagement. Figure 9 does not depict a 6-vs.-10 engagement. However, the edfs can be used to compute the probabilities of error for 1-vs.-many engagements as described below.
3. The probabilities of error computed are exact for the edfs shown. There is no way to adjust these probabilities to account for the fact that the edfs might not reproduce the true distribution as noted in the first point. The edfs comprise all that is known about how the threat objects will vary in the feature space.

Figure 10 plots the probability of false alarm against the probability of leakage for each of the 16 realizations. This is the operating curve for the two distributions. The equal error rate is the point on the operating curve where the two probabilities are equal. An equal error rate of 0.33 would seem to indicate that the probability of error in selecting the RV would be little better than a coin flip. However, the true probability of error for classifying one of two objects in the FOV as the RV is much lower as shown in Eq. 14 and computed in Eq. 9.

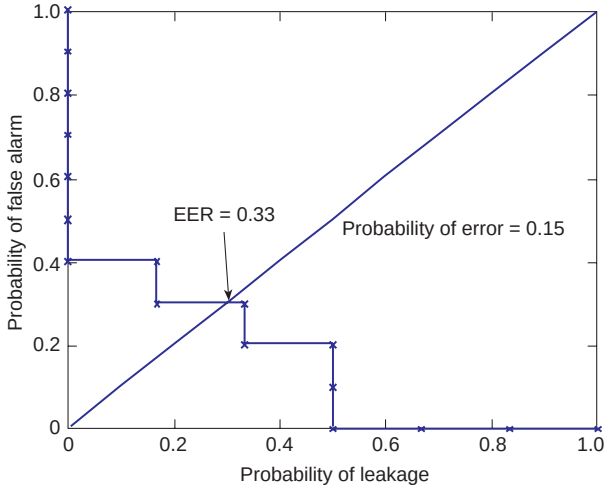

Figure 10. Generating probabilities of error for 1-vs.-1 classification: operating curve and equal error rate (EER).

Classifying 1 RV vs. q Non-RVs

Now assume there are $q + 1$ objects in the FOV, one from the RV class ω_i and q from a non-RV class ω_j. As before, denote the RV feature observation $\overline{X}_i$ and the non-RV feature observation

$$\left\{\overline{X}_j^k\right\}_{k=1}^{q}.$$

A classification error occurs if $d_{ij}(\overline{X}_i) < d_{ij}(\overline{X}_j^k)$ for any k—if any one of the non-RVs looks more RV-like than the RV itself. Thus, it suffices to look at the feature vector from the ω_j class with the largest decision metric. Let

$$d_{ij}^{(q)}(\overline{X}_j) \equiv \sup_{k \in [1,\, q]} \{d_{ij}(\overline{X}_j^k)\},$$

the qth order statistic of the decision metric. The distribution of $d_{ij}(\overline{X}_j^k)$ is known: $p_{d_{ij}}(d \mid \omega_j) \equiv p_{d_{ij}}(d_j)$. For continuous d_j, the distribution of the largest of q independent observations is given by

$$p_{d_{ij}^{(q)}}(d_j) = \binom{q}{1} p_{d_{ij}}(d_j)[P_{d_{ij}}(d_j)]^{q-1}. \qquad (15)$$

Here $P_{d_{ij}}$ is the cumulative distribution function, or the integral of $p_{d_{ij}}$. Mnemonically, the probability that the largest $d_{ij}(\overline{X}_j^k)$ is equal to d_j is the probability that any one is equal to this value times the probability that the remaining $q - 1$ are smaller. For the (1 vs. q) case, Eq. 8 becomes

$$P_{\text{Error}} = \int_{-\infty}^{\infty} \int_{-\infty}^{d_j} p_{d_{ij}}(d_i) p_{d_{ij}^{(q)}}(d_j) \partial d_i \partial d_j. \qquad (16)$$

The class-dependent feature edfs will induce a discrete decision space for d_j as shown in Fig. 9. However, Eq. 15 cannot be used in the discrete case as the underlying probability argument does not obtain. To

produce the analog for Eq. 16, a discrete version of Eq. 15 is necessary. The order statistics of the discrete distribution can be obtained with the following development. Suppose a scalar edf comprises M discrete points $\{d_k\}_{k=1}^{M}$. Let $\{d_{[k]}\}_{k=1}^{M}$ denote the ordered set of possible points and $d^{(q)}$ represent the largest of q independent observations.

$$\begin{aligned}
\Pr(d^{(q)} = d_{[k]}) &= \Pr(\{1 \text{ or more of the } q\ d's = d_{[k]}\} \\
&\quad \cap \{\text{all remaining } d's < d_{[k]}\}) \\
&= [p_d(d_{[k]})]^1 [P_d(d_{[k-1]})]^{q-1} \\
&\quad + [p_d(d_{[k]})]^2 [P_d(d_{[k-1]})]^{q-2} + \cdots \\
&\quad + [p_d(d_{[k]})]^q \\
&= \sum_{i=1}^{q} [p_d(d_{[k]})]^i [P_d(d_{[k-1]})]^{q-i} \\
&= \left(\sum_{i=0}^{q} p_d(d_{[k]})]^i [P_d(d_{[k-1]})]^{q-i} \right) \\
&\quad - [P_d(d_{[k-1]})]^q \\
&= [p_d(d_{[k]}) + P_d(d_{[k-1]})]^q \\
&\quad - [P_d(d_{[k-1]})]^q \\
&= [P_d(d_{[k]})]^q - [P_d(d_{[k-1]})]^q .
\end{aligned} \tag{17}$$

This last equation states that the probability that the largest of q independent observations is equal to the kth largest possible value is equal to the probability that all q values are less than or equal to the kth largest minus the probability that all q values are less than or equal to the $(k-1)$th largest. With the rank ordering, Eq. 17 can be rewritten very simply:

$$P_{d^{(q)}}(d_{[k]}) = \left[\frac{k}{M}\right]^q - \left[\frac{k-1}{M}\right]^q . \tag{18}$$

Returning to the classification problem, if the observations $\{d_{ij}(\bar{X}_{k_2} \mid \omega_j)\}_{k_2=1}^{J}$ are rank sorted so that $d_{ij}(\bar{X}_{(k_2)} \mid \omega_j)$ is the k_2th largest decision metric for the class ω_j, then the discrete edf formula of Eq. 9 is reprised:

$$\begin{aligned}
P_{\text{Error}} = \frac{1}{IJ^q} \Bigg(& \sum_{k_2=1}^{J} \left[k_2^q - (k_2 - 1)^q \right] \\
& \times \sum_{k_1=1}^{I} \delta[d_{ij}(\bar{x}_{k_1} \mid \omega_i) < d_{ij}(\bar{x}_{k_2} \mid \omega_j)] \Bigg) .
\end{aligned} \tag{19}$$

For $q = 1$, this equation simplifies to Eq. 9, as it should.

Example (continued)

The example of the section entitled Classifying 1 RV vs. 1 Non-RV continues with the assumption that now q non-RV objects are drawn from the non-RV distribution denoted by the circles in Fig. 9. The probability of classification error for correctly selecting 1 RV against q non-RVs is plotted against q in Fig. 11. The probability of error asymptotically approaches 0.5. This might seem counterintuitive in that one might expect mistaken classification with virtual certainty as the number of non-RV objects becomes large. The distribution of the non-RVs in the feature space in Fig. 9 explains this result. As the number of non-RVs increases, the distribution of most RV-like non-RV approaches a Dirac delta function on the most RV-like of the possible non-RV feature realizations. (This is analogous to the distribution of largest of a sequence of dice rolls. As the sequence grows longer, the largest value will be a 6 with increasing likelihood.) However, 3 of the 6 RV realizations are more RV-like than the most RV-like of the non-RVs. Put another way, half of the RVs are outside of the non-RV boundary. Hence, the probability of error approaches 0.5, rather than 1, for large numbers of non-RVs. This is an important observation:

> *The probability of error in classifying 1 RV against* q *non-RVs may be considerably lower than 1 even for large* q.

It can be argued that this rather benign circumstance is due to the small sample of non-RV feature realizations: a larger sample would certainly have a worst-case non-RV sample that was more RV-like than all but the most RV-like RVs. However, the process for introducing extrema is the same as that for nominal realizations: the simulation. The class-dependent feature edfs contain all that is known about how the

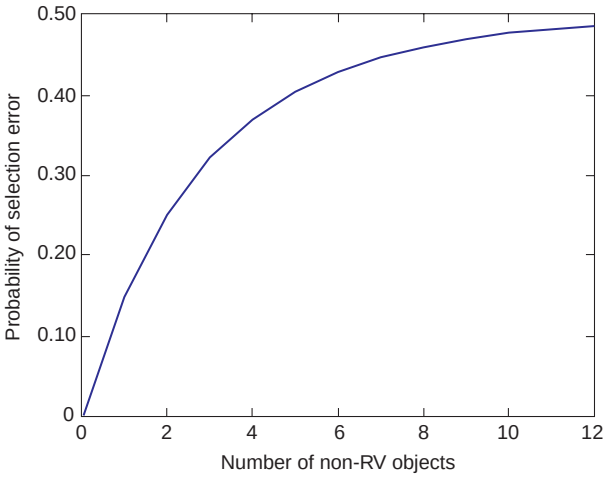

Figure 11. Probability of error in classifying 1 reentry vehicle (RV) against q non-RVs.

objects will appear in the feature space. There is no other formal method for making inferences about performance.

In tandem, Figs. 9 and 11 illustrate the superior performance (lower error rate) of the likelihood ratio approach to classification to the popular maximum likelihood approach. The maximum likelihood approach chooses as the RV the object with the feature realization (Fig. 9) closest to the center of the RV ellipse. As the number of non-RVs in the ballistic complex increases, the non-RV realization (circle) near the center of the RV cluster becomes increasingly likely. Although the iso-ellipses for the RV class are not sketched in Fig. 9, it is clear that only one of the six RV realizations is closer to the center than this non-RV circle. Therefore, the asymptotic probability of error will be 5/6, considerably worse than the 1/2 shown in Fig. 11.

Classifying 1 RV vs. n Unique Non-RVs

Suppose now there are $n + 1$ objects in the FOV: 1 RV and n non-RVs, each from a unique class. Then there are $n + 1$ models of the class-dependent feature distributions, and n possible pairwise decisions between the RV class and another class. Let ω_1 denote the RV class and $\omega_2, \omega_3, \cdots, \omega_{n+1}$ denote the non-RV classes. Designate the feature realizations from the respective classes $\overline{X}_1, \overline{X}_2, \cdots, \overline{X}_{n+1}$. The objective is to choose the feature realization that is most RV-like with respect to the n possible likelihood ratios. To assess the error in making this choice, the characteristics of the selection algorithm must be specified. The probability of error is the integral over the class-dependent distributions in the feature space where an object other than the RV will be determined to the RV. An error integral analogous to Eq. 7 must be specified. As in the section entitled Classifying 1 RV vs. 1 Non-RV, the integration is facilitated by circumventing the implicit limits of integration with the transformation of the feature distributions to the decision space.

In Shannon's Theorem it is shown that the capacity (or error rate) of a noisy channel can be determined without specifying the signal encoding to achieve that capacity. A similar situation obtains for assigning the RV in a polychotomous decision space: conditions can be imposed on feasible but undetermined assignment algorithms sufficient to determine the error rate of those algorithms. In the section entitled Classifying 1 RV vs. 1 non-RV, the algorithm for choosing 1 RV against 1 non-RV was straightforward and the error integral proceeded directly. Some care must be taken in specifying the characteristics of the analog for choosing one RV against n non-RVs from n unique classes.

It is desired to classify only the RV. (There is added benefit to assigning each of the n non-RV feature realizations to their respective classes in that the spatial context of the identified objects provides additional discrimination information. This assignment proceeds in exactly the same fashion as the assignment of the RV feature realization; therefore, it suffices to specify the algorithm for assigning a feature realization to the RV class.) The n pairwise decision families are the log-likelihoods $\{d_{1i}\}_{i=2}^{n+1}$, where d_{ij} was defined in the section entitled Classifying 1 RV vs. 1 Non-RV. The set of decision families will be used to cull the feature realization $\overline{X}_1$ associated with the RV class ω_1 from the set of realizations $\{\overline{X}_i\}_{i=1}^{n+1}$. The algorithm should correctly choose $\overline{X}_1$ if $d_{1i}(\overline{X}_1) > d_{1i}(\overline{X}_i) \forall i \neq 1$. With this requirement, an error will occur if $d_{1i}(\overline{X}_i) > d_{1i}(\overline{X}_1)$ for any $i \in [2, n+1]$. Denote each possible error E_i. Then

$$
\begin{aligned}
P_{\text{Error}} &= P(E_2 \cup E_3 \cup \cdots \cup E_{n+1}) \\
&= \sum_{k=2}^{n+1} P(E_k) - \sum_{k_1 < k_2} P(E_{k_1} \cap E_{k_2}) \\
&\quad + \cdots + (-1)^{p+1} \sum_{k_1 < k_2 < \cdots < k_p} P(E_{k_1} \cap E_{k_2} \cap \cdots \cap E_{k_p}) \\
&\quad + \cdots + (-1)^{n+1} P(E_2 \cap E_3 \cap \cdots \cap E_{n+1}).
\end{aligned}
$$

In the preceding equation, $p < n + 1$. Since each of the feature realizations is independent, the pairwise errors are independent so that

$$
\begin{aligned}
P_{\text{Error}} &= \sum_{k=2}^{n+1} P(E_k) - \sum_{k_1 < k_2} P(E_{k_1}) P(E_{k_2}) \\
&\quad + \cdots + (-1)^{p+1} \sum_{k_1 < k_2 < \cdots < k_p} P(E_{k_1}) P(E_{k_2}) \cdots P(E_{k_p}) \\
&\quad + \cdots + (-1)^{n+1} P(E_2) P(E_3) \cdots P(E_{n+1}).
\end{aligned} \tag{20}
$$

Here each error $P(E_k)$ is the one vs. one error derived in the section entitled Classifying 1 RV vs. 1 Non-RV.

Classifying 1 RV vs. $q_1, q_2, \ldots, q_n$ *Non-RVs from* n *Unique Classes*

Extending the problem from the previous section, allow q_i objects to be drawn from each non-RV class ω_{i+1}. (Recall that ω_1 denotes the RV class, of which there is a single member.) The algorithm is required to select the RV if the RV is more RV-like than each member of every class with respect to the appropriate likelihood ratio metric. This is written as

$$d_{1i}(\bar{X}_1) > d_{1i}^{(q_i)}(\bar{X}_i) \forall i.$$

Here $d_{1i}^{(q_i)}(\bar{X}_i)$ is the largest of the observed decision metrics for the class ω_i as defined earlier. Then if the error E_i is redefined from the previous section to be the event $d_{1i}(\bar{X}_i) < d_{1i}^{(q_i)}(\bar{X}_i)$, Eq. 20 suffices to specify the error, where $P(E_k)$ is now the assignment error for 1 RV vs. q non-RVs computed in Eqs. 16 and 19.

SUMMARY

This article traced the application of likelihood classification techniques to the problem of selecting the RV in a TBMD discrimination problem. The process was illustrated with an example using simulated IR data. The probability of error for a dichotomous decision between two objects was derived and related to the familiar quantities, equal error rate and k-factor. This probability of error formula was extended to a decision against a number of non-RVs from a single class. In tandem, these results yielded the probabilities of error for choosing the RV from an arbitrary number of objects drawn from an arbitrary number of classes.

REFERENCE

[1] Fry, R. L., *The Advantages of Computing the A Posteriori Discrimination Probability from Feature Statistics and A Priori Discrimination Information*, JHU/APL Technical Memorandum A1F(1)97-U-006 (25 Feb 1997).

BIBLIOGRAPHY

Fry, R. L., "Maximized Mutual Information Using Macrocanonical Probability Distributions," in *Proc. 1994 IEEE/IMS Workshop on Information Theory and Statistics*, p. 63, Arlington, VA (1994).

Fry, R. L., "Observer-Participant Models of Neural Processing," *IEEE Trans. Neural Networks* 6, 918–928 (July 1995).

Fry, R. L., "Neural Mechanics," in *Progress in Neural Information Processing: Proc. Int. Conf. Neural Inf. Processing*, pp. 158–164, Hong Kong (1996).

Fry, R. L., *The Use of Sufficient Information for Interceptor Guidance Against a Tactical Ballistic Missile Threat*, JHU/APL Technical Memorandum A1F(1)97-U-086 (16 Sep 1997).

Fry, R. L., and Sova, R. M. "A Logical Basis for Neural Network Design," in *Theory and Application of Neural Networks*, pp. 259–307, Academic Press, New York (1998).

Kitzman, K. V., *Quantitative Basis for Two-Color IR Band Selection: Implementation and Example Calculation*, JHU/APL Technical Memorandum A1F(1)97-U-088 (15 Sep 1997).

Kitzman, K. V., *Single-Waveband J-Factor Analysis*, JHU/APL Technical Memorandum A1F(1)97-U-131 (3 Dec 1997).

Kitzman, K. V., and Baker, J. P., *Monte Carlo Integration Applied to Class-Dependent Irradiance Distributions*, JHU/APL Technical Memorandum A1F(1)97-U-109 (Oct 1997).

Lafrance, P., *Sequential Discrimination for TBMD*, JHU/APL Technical Memorandum F2F-96-U-1-016 (22 Jul 1996).

Lafrance, P., Bohse, M. E., Budman, C. A., Jackson, A. D., Spriesterbach, T. P., and Telford, J. K., *Navy TBMD COEA Phase II Radar Discrimination Product*, JHU/APL Technical Memorandum SSD/PM-97-0536 (18 Aug 1997).

Resch, C. L., *Preliminary Results of Thrust Termination Debris Modeling and Discrimination*, JHU/APL Technical Memorandum RSI-97-021 (26 Feb 1997).

Resch, C. L., "Exo-atmospheric Discrimination of Thrust Termination Debris and Missile Segments," *Johns Hopkins APL Tech. Dig.* (this issue).

"Transduction and Transmission: Logical Processes of Information Exchange," in *Proc. 1998 IEEE Int. Symp. on Inf. Theory*, in press.

ACKNOWLEDGMENTS: I am grateful to Robert L. Fry for his inspired leadership throughout the Navy Phase II COEA and for his thorough and insightful review. The technical staff of Photon Research Associates (Ralph Waters, Richard Moore, Joseph Filice, Ben McGlammery, and Anthony Sommese) illustrated the classifier modeling process with their superlative work on the COEA. John P. Baker, James C. Stamper, Pierre Lafrance, and James M. Bielefeld patiently endured a number of questions. Forest C. Deal's elucidation of the arcana of discrete order statistics was positively pellucid. The Editorial Board reviewer of the *Technical Digest* made several suggestions that improved the readability and organization of this article.

THE AUTHOR

GEOFFREY L. SILBERMAN has a B.S. from Virginia Tech and an M. S. from Princeton University. He joined APL in 1991 and is a member of the technical staff. He is currently working on Navy Theater-Wide Theater Ballistic Missile Defense. He was the principal investigator on APL's Detection In-Vehicle of Impaired Driving project. He has also worked on building stochastic models of inertial navigation systems to estimate accuracy for the Trident II missile. His e-mail address is geoff.silberman@jhuapl.edu.