place within the duration of any pulsed signals. When this process is applied to the above example, the best-fit power ratio is found to be 6.1 dB.

IV. Summary

The CWT (using the Morlet wavelet) can be used to extract the relative amplitudes of two or more transient sinusoidal signals. Only the ratio of the CWT amplitudes at a single time-slice is required. However, this ratio must be multiplied by the square root of the inverse ratio of the scale factors (the reciprocal of the frequency) of the signals (2) in order to calculate the correct relative amplitudes. Two limitations on the direct application of the correction technique were presented. First, the effect of finite pulse width on the calculation of the amplitude ratio was presented. If the chosen time-slice location is inside the pulse and at least three scale factors from the closest pulse edge, (2) can still be used. Second, the error introduced by nearby (in frequency) signals in the ratio calculation was presented. Error-multiplier contours were calculated for different amplitude and scale ratios. Example calculations were presented in order to demonstrate the general technique and to obtain the correct amplitude ratio even if the worst-case errors are too large.

Acknowledgment

The author would like to thank Dr. C. P. Silva for his encouragement and review of this work.

References

[1] A. Grossmann and J. Morlet, "Decomposition of hardy functions into square integrable wavelets of constant shapes," *SIAM J. Math. Anal.*, vol. 15, pp. 723–736, 1984.
[2] G. Strang and T. Nguyen, *Wavelets and Filter Banks*. Wellesley, MA: Wellesley-Cambridge, 1996.
[3] S. G. Sánchez, N. G. Prelcic, and S. J. G. Galán, *Uvi_Wave Wavelet Toolbox for Matlab*. Vigo, Spain: Univ. de Vigo, 1996.

Class-Specific Feature Sets in Classification

Paul M. Baggenstoss

***Abstract*—In this correspondence, we present a new approach to the design of probabilistic classifiers that circumvents the dimensionality problem. Rather than working with a common high-dimensional feature set, the classifier is written in terms of likelihood ratios with respect to a common class using sufficient statistics chosen specifically for each class.**

***Index Terms*—Classification, class-specific features, sufficiency.**

Manuscript received March 26, 1997; revised April 29, 1999. This work supported by the Office of Naval Research. The associate editor coordinating the review of this paper and approving it for publication was Dr. Elias S. Manolakos.

The author is with the Naval Undersea Warfare Center, Newport, RI 02841 USA.

Publisher Item Identifier S 1053-587X(99)09207-7.

I. Introduction

Consider the problem of classifying a data sample $\mathbf{X}$ into one of M classes. This is done optimally by the classifier known as the maximum *a posteriori* (MAP) or Bayesian classifier

$$\arg \max_{j=1}^{M} p(H_j|\mathbf{X}) = \arg \max_{j=1}^{M} p(\mathbf{X}|H_j)\, p(H_j). \tag{1}$$

However, if the likelihood functions $p(\mathbf{X}|H_j)$ are not known, it is necessary to estimate them from *training data*. Dimensionality dictates that this is impractical or impossible unless $\mathbf{X}$ is reduced to a smaller set of statistics or *features* $\mathbf{Z} = T(\mathbf{X})$. One possible strategy for choosing features is to identify a set of statistics $\mathbf{z}_j$ corresponding to each class H_j that is sufficient or approximately sufficient to estimate the unknown state of the class .[1] For example, if H_j was a sinewave in white noise, $\mathbf{z}_j$ would be based on a Fourier analysis, but if H_j was a white noise signal, $\mathbf{z}_j$ would be based on power estimates. Because some classes may be similar to each other, it is possible that the feature sets are not distinct. Let

$$\mathbf{Z} = \bigcup_{i=1}^{M} \mathbf{z}_i$$

where set union notation is used to indicate that there are no redundant features in $\mathbf{Z}$. However, removing redundant features is not restrictive enough. A more restrictive but necessary requirement is that $p(\mathbf{Z}|H_j)$ exists for all j.[2] The classifier based on $\mathbf{Z}$ becomes

$$\arg \max_{j=1}^{M} p(\mathbf{Z}|H_j)\, p(H_j). \tag{2}$$

The object of the feature selection process is to insure that (2) is equivalent to (1). Thus, the features are *sufficient* for the problem at hand. We will see in the theorem that follows that there is a connection between the sufficiency of the feature set for the classification problem and the classic (Neyman–Fisher) sufficiency. In spite of the fact that the feature sets $\mathbf{z}_j$ are chosen in a *class-specific* manner and are possibly each of low dimension, implementation of (2) requires that the features be grouped together into a super-set $\mathbf{Z}$. However, dimensionality issues dictate that $\mathbf{Z}$ must be of low dimension (less than about 5 or 6) so that a good estimate of $p(\mathbf{Z}|H_j)$ may be obtained with a reasonable amount of training data. It is recognized by a number of researchers that attempting to estimate PDF's nonparametrically above five dimensions is difficult and above 20 dimensions is futile [1]. It is common for high-dimensional PDF estimators to perform very well as classifiers in many applications. However, this is due to the inherent separability of the classes in the high-dimensional space where any PDF estimator may perform as well as another. Further performance improvements are difficult without addressing the dimensionality problem.

Dimensionality reduction is the subject of much research currently and over the past decades (some good overviews are available [1]–[3]). Various approaches include feature selection [2]–[4], projection pursuit [5], [6], *independence grouping* [7], and *subspace methods* [8]–[12]. All these methods involve various approximations. In feature selection, the approximation is that most of the information concerning all data classes is contained in a few of the features. In projection-based methods, the assumption is that information is confined to linear subspaces.

[1] Sufficiency in this context will be defined more precisely in the theorem that follows.

[2] Thanks to S. Kay for suggesting this requirement.

We now describe a procedure for choosing class-specific feature sets $\mathbf{z}_j, j = 1, 2, \ldots M$ such that

a) a classifier can be constructed using only the joint PDF's of these class-specific feature sets taken separately;
b) this classifier is equivalent to the classifier constructed from the union of the features;
c) both classifiers are equivalent to the MAP classifier (1).

Thus, the features are sufficient for the classification problem as a whole. The following theorem proves the first two points.

II. Main Theorem

What we now show is that it is possible to reduce the maximum PDF dimension while at the same time retaining theoretical equivalence to the classifier constructed from the full feature set (2) and to the optimum MAP classifier (1). In the class-specific method of feature selection introduced above, the fact that $\mathbf{z}_j$ corresponds to H_j is information that is discarded when $\mathbf{Z}$ is created and is not utilized in (2).

Theorem 1: Let there be M distinct PDF families $p(\mathbf{X}|H_j), j = 1, 2, \cdots, M$, where H_j are the class hypotheses. For each class j, let $p(\mathbf{X}|H_j)$ be parameterized by a random parameter set $\boldsymbol{\theta}_j$; thus

$$p(\mathbf{X}|H_j) = \int_{\boldsymbol{\theta}_j} p(\mathbf{X}|\boldsymbol{\theta}_j, H_j) p(\boldsymbol{\theta}_j)\, d\boldsymbol{\theta}_j$$

for all j. For each class j, let there be a sufficient statistic for $\boldsymbol{\theta}_j$, $\mathbf{z}_j = T_j(\mathbf{X})$. Let the PDF of the combined feature set $p(\mathbf{Z}|H_j)$, where $\mathbf{Z} = \bigcup_{i=1}^{M} \mathbf{z}_i$ exist for all j. Let the span of $\boldsymbol{\theta}_j$ include a point $\boldsymbol{\theta}_j^0$ that results in an equivalent distribution for $\mathbf{X}$ regardless of j

$$p(\mathbf{X}|H_j, \boldsymbol{\theta}_j^0) = p(\mathbf{X}|H_0), \qquad j = 1, \cdots, M. \tag{3}$$

Then, the classifier based on the combined feature set (2) reduces to

$$\arg \max_j \frac{p(\mathbf{z}_j|H_j)}{p(\mathbf{z}_j|H_0)}\, p(H_j). \tag{4}$$

Proof: Note that from (3), we have

$$p(\mathbf{Z}|H_j, \boldsymbol{\theta}_j^0) = p(\mathbf{Z}|H_0), \qquad j = 1, 2, \cdots, M. \tag{5}$$

We may write

$$\begin{aligned} p(\mathbf{Z}|H_j) &= \int p(\mathbf{Z}|H_j, \boldsymbol{\theta}_j) p(\boldsymbol{\theta}_j|H_j)\, d\boldsymbol{\theta}_j \\ &= \int p(\mathbf{Z}^j|\mathbf{z}_j, H_j, \boldsymbol{\theta}_j)\, p(\mathbf{z}_j|H_j, \boldsymbol{\theta}_j)\, p(\boldsymbol{\theta}_j|H_j)\, d\boldsymbol{\theta}_j \end{aligned}$$

where $\mathbf{Z}^j$ is the result of removing $\mathbf{z}_j$ from $\mathbf{Z}$ defined by

$$\begin{aligned} \mathbf{Z}^j \bigcap \mathbf{z}_j &= \emptyset \\ \mathbf{Z}^j \bigcup \mathbf{z}_j &= \mathbf{Z}. \end{aligned}$$

We now make use of the fact that $p(\mathbf{Z}^j|\mathbf{z}_j, H_j, \boldsymbol{\theta}_j)$ is independent of $\boldsymbol{\theta}_j$ due to sufficiency, and we may evaluate it at any value of $\boldsymbol{\theta}_j$; we choose $\boldsymbol{\theta}_j^0$.

$$\begin{aligned} p(\mathbf{Z}|H_j) &= p(\mathbf{Z}^j|\mathbf{z}_j, H_j, \boldsymbol{\theta}_j^0) \\ &\quad \cdot \int p(\mathbf{z}_j|H_j, \boldsymbol{\theta}_j)\, p(\boldsymbol{\theta}_j|H_j)\, d\boldsymbol{\theta}_j \\ &= p(\mathbf{Z}^j|\mathbf{z}_j, H_j, \boldsymbol{\theta}_j^0)\, p(\mathbf{z}_j|H_j) \end{aligned}$$

Now, $p(\mathbf{Z}^j|\mathbf{z}_j, H_j, \boldsymbol{\theta}_j^0)$ may be expanded to

$$p(\mathbf{Z}^j|\mathbf{z}_j, H_j, \boldsymbol{\theta}_j^0) = \frac{p(\mathbf{Z}|H_j, \boldsymbol{\theta}_j^0)}{p(\mathbf{z}_j|H_j, \boldsymbol{\theta}_j^0)}.$$

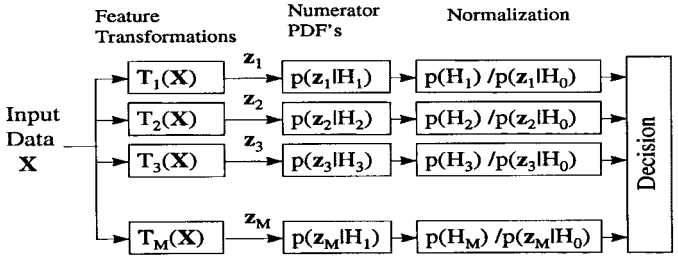

Fig. 1. Detector/classifier architecture.

Now, $p(\mathbf{Z}|H_j, \boldsymbol{\theta}_j^0)$ is independent of j as a result of (3), and thus

$$p(\mathbf{Z}|H_j) = \frac{p(\mathbf{z}_j|H_j)}{p(\mathbf{z}_j|H_0)}\, p(\mathbf{Z}|H_0)$$

where we write the conditioning $\{H_j, \boldsymbol{\theta}_j^0\}$ as H_0. Now, plugging into (2) and dividing out $p(\mathbf{Z}|H_0)$, which does not depend on j, we get

$$\begin{aligned} &\arg \max_{j=1}^{M} p(\mathbf{Z}|H_j)\, p(H_j) \\ &\quad = \arg \max_{j=1}^{M} \frac{p(\mathbf{z}_j|H_j)}{p(\mathbf{z}_j|H_0)}\, p(H_j) \end{aligned} \tag{6}$$

which is the same as (4). □

Relationship to MAP Classifier: The equivalence of the full-dimensional feature-based classifier (2) and the class-specific formulation (4) leads us to ask whether the two classifiers are equivalent to the MAP classifier itself (1). The answer is yes. To see this, we begin by dividing the MAP classifier by the density of $\mathbf{X}$ under the common class H_0. We have

$$\arg \max_j \frac{p(\mathbf{X}|H_j)}{p(\mathbf{X}|H_0)}\, p(H_j). \tag{7}$$

This leads to the M-ary classifier for uniform Bayesian cost function [15]. The M-ary classifier is implemented in practice by choosing H_0 so that it is possible to analytically simplify the likelihood ratios for each j to processors such as matched filters, etc. Since this is possible only in some simple cases, it has not found much use in general classification problems.

In order to arrive at (4), we use the property of likelihood ratios that they are invariant when written in terms of a sufficient statistic [16]. We note that the sufficiency of $\mathbf{z}_j$ for the underlying parameters set $\boldsymbol{\theta}_j$ means that $\mathbf{z}_j$ is sufficient for the binary test H_j versus H_0. In this way, we arrive at (4) immediately.[3]

A. Discussion

1) Classifier Architecture: The formulation (4) suggests a detector/classifier architecture as shown in Fig. 1. Each data class corresponds to a distinct and independent branch in the diagram. The output of each branch is a detection statistic for distinguishing the corresponding signal class from H_0. The modularity of the processor is has obvious advantages. As long as the same H_0 is used, each branch can be independently designed, trained, and implemented by separate computational hardware. As new signal classes are added to the classifier, it only means adding new branches to the structure—existing branches remain unchanged.

2) Previous Work: The likelihood ratios are known to be sufficient for optimal classification (see, for example, Lehmann [13]). Furthermore, the use of likelihood ratios referenced to a "dummy" hypothesis (H_0) has been used in classification (see Van Trees [15]).

[3]Thanks to S. Kay for this argument and other useful comments.

Yet, the replacement of $\mathbf{X}$ by sufficient statistics individually chosen for each hypothesis, and of various dimension, appears to be new. Architectures have previously been proposed with a class-specific structure, as in Fig. 1 [14] but is the first time it has been placed on any theoretical relationship to the MAP classifier.

3) The Common Class, H_0: The common class H_0 does not need to be a realistic class. Technically, the only requirement is that the parameter sets of each class must include H_0 as a special case; thus, we have the natural role of the noise-only hypothesis. For reasons explained in the next section, we have found it useful that H_0 represents the condition that $\mathbf{X}$ be samples of *iid* Gaussian noise.

4) Establishing Sufficiency: In many real-world problems, the PDF of $\mathbf{X}$ is never known. Thus, the sufficiency of features can never be established theoretically. Therefore, how can the technique be used? The simple answer is that sufficiency does not need to hold exactly in practice. If sufficiency is approximated, so is the relationship of the resulting classifier to the optimal MAP classifier. The sufficiency question is a separate problem that we do not address. However, we have found it useful to require that the features provide enough information so that the original data can be "recreated to acceptable fidelity." The meaning of this depends on the application. For speech recognition, this would mean that the spoken word is still intelligible. In a lie detector, however, it would be a more stringent requirement because the emotional state of the speaker would need to be preserved.

5) Numerical Issues in Estimating the Densities: To utilize (4), it is necessary to obtain estimates of $p(\mathbf{z}_j|H_k)$ for both $k = 0$ and $k = j$. For $k = j$, it is clear that exemplars of $\mathbf{z}_j$ from a training data set may be used to train a density estimate, for example, using Gaussian mixtures via the EM algorithm. Likewise, for $k = 0$, a large number of exemplars may be created under the noise-only assumption by simulation. However, in applications where the input data differs greatly from H_0 (i.e. high-SNR), the denominator densities $p(\mathbf{z}_j|H_0)$ must be evaluated in the tail areas where the approximation is poor. We observe all the denominators in (4) going to zero simultaneously. Thus, it is necessary in many cases to use exact analytic expressions for $\log\ p(\mathbf{z}_j|H_0)$. It is surprising and counterintuitive that meaningful results can be obtained in the far-tail regions. However, the densities in the tails contain all the required normalization factors, and as long as the expressions for $\log\ p(\mathbf{z}_j|H_0)$ are accurate, there are no errors introduced. It also seems to be an overly restrictive requirement that analytic expressions need to be obtained under the H_0 assumption. However, if H_0 is defined as *iid* Gaussian noise, the problem is greatly simplified. This problem of tail approximation breathes new life into an old statistical problem. We have already obtained and tabulated exact results for a large variety of features including order statistics and autocorrelation estimates.

III. Example Problem

The purpose of the example is to illustrate the application of the class-specific method in a controlled experiment using synthetic signals. A given set of features will be used in both a conventional and a class-specific arrangement. The signals were not chosen to represent any real-world problem in particular. They were chosen 1) to provide clear sufficient statistics with known distributions under H_0 and 2) to provide a difficult classification environment with some similar signal types at a wide range of signal strengths. Sufficient information is provided so that the experiment may be reproduced and we may compare the results with other methods. Because the signals are synthetic, an unlimited number of samples may be produced. This allows the asymptotic (large sample) classification performance to be approximated in the limit.

A. Signal Models

Let the input data to the classifier be a sample of a time-series of N samples denoted $\mathbf{X} = \{x_1, x_2, \cdots, x_N\}$. Consider the following three signal classes as possible statistical models for $\mathbf{X}$.

1) H_1:

$$x_t \sim \mathcal{N}(\alpha, 1), \qquad \alpha \neq 0$$

2) H_2:

$$x_t \sim \mathcal{N}(0, 1 + \sigma^2), \qquad \sigma^2 > 0$$

3) H_3:

$$x_t \sim \begin{cases} \mathcal{N}(0, 1 + \beta^2), \beta^2 > 0, & t = 1 \\ \mathcal{N}(0, 1), & t > 1 \end{cases}$$

where we use the shorthand notation $\mathcal{N}(\mu, \sigma^2)$ to represent *iid* Gaussian noise of mean μ and variance σ^2. Let the parameters α, β^2, and σ^2 be random variables that are fixed for the duration of $\mathbf{X}$ and whose probability distributions are not known.

Consider the following CS features:

$$\begin{aligned} z_1 &= \sum_t x_t \\ z_2 &= \sum_t x_t^2 \\ z_3 &= \log(x_1^2). \end{aligned} \tag{8}$$

It should be obvious that we have set up this problem so that the "common class" H_0 is given by $x_t \sim \mathcal{N}(0, 1)$, all t. In the next section, we see that these features are indeed sufficient statistics for the corresponding classes.

B. Sufficient Statistics and Their Densities Under H_0

In this section, we derive the sufficient statistics for hypotheses H_1 through H_3 in the example. For H_1, we write the likelihood ratio as a function of the unknown parameter α.

$$\begin{aligned} \frac{p(\mathbf{X}|H_1)}{p(\mathbf{X}|H_0)} &= \frac{\prod_{t=1}^{N} (2\pi)^{-1/2}\ \exp\{-(x_t - \alpha)^2/2\}}{\prod_{t=1}^{N} (2\pi)^{-1/2}\ \exp\{-x_t^2/2\}} \\ &= \exp\left\{\alpha \sum_t x_t - N\alpha^2/2\right\}. \end{aligned}$$

It is clear that the likelihood ratio is a function of $z_1 = \Sigma_t\ x_t$. Thus, no matter what the distribution of α, the likelihood ratio test will depend on the data only through z_1. Therefore, z_1 is an SS for the problem of testing H_1 against H_0. The distribution of z_1 under H_0 is $\mathcal{N}(0, N)$:

$$\log\ p(z_1|H_0) = -0.5\ \log(2\pi N) - \frac{z_1^2}{2N}.$$

TABLE I
TABLE OF FEATURES AND PDF'S UNDER H_0 FOR THE EXAMPLE

Hypothesis	Feature	Distribution
H_1	$z_1 = \sum_t x_t$	$\log p(z_1\|H_0) = -0.5\log(2\pi N) - \frac{z_1^2}{2N}$
H_2	$z_2 = \sum_t x_t^2$	$\log p(z_2\|H_0) = -\log\Gamma(N/2) - N/2\log 2 + (N/2-1)\log z_2 - z_2/2$
H_3	$z_3 = \log(x_1^2)$	$\log p(z_3\|H_0) = -1/2\log 2\pi + z_3/2 - \exp(z_3)/2$

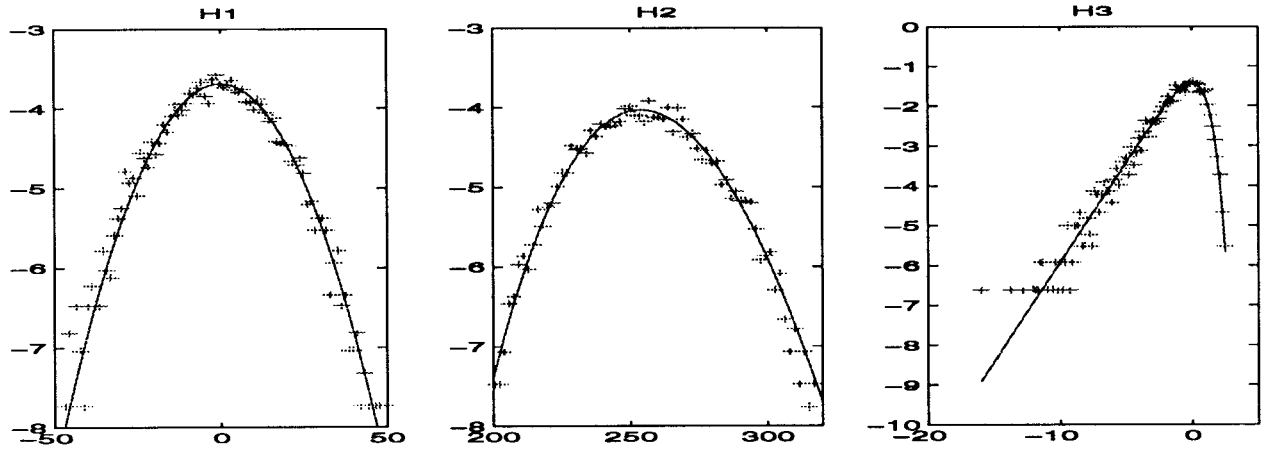

Fig. 2. Log-histograms of features $z_1, z_2,$ and z_3 for Gaussian input data plotted on the theoretical curves of log-PDF.

For H_2, we write the likelihood ratio as a function of the unknown parameter σ^2.

$$\frac{p(\mathbf{X}|H_2)}{p(\mathbf{X}|H_0)} = \frac{\prod_{t=1}^{N}(2\pi[1+\sigma^2])^{-1/2}\ \exp\left\{-\dfrac{x_t^2}{2(1+\sigma^2)}\right\}}{\prod_{t=1}^{N}(2\pi)^{-1/2}\ \exp\{-x_t^2/2\}}$$
$$= \prod_{t=1}^{N}(1+\sigma^2)^{-1/2}\ \exp\left\{-\frac{x_t^2}{2(1+\sigma^2)} + x_t^2/2\right\}$$
$$= (1+\sigma^2)^{-N/2}\ \exp\left\{\frac{\sum_t x_t^2}{2}\left(\frac{\sigma^2}{1+\sigma^2}\right)\right\}.$$

It is clear that the likelihood ratio is a function of $z_2 = \Sigma_t\ x_t^2$. Thus, no matter what the distribution of σ^2, the likelihood ratio test will depend on the data only through z_2. Therefore, z_2 is an SS for the problem of testing H_2 against H_0. The distribution of z_2 under H_0 is Chi-squared with N degrees of freedom

$$\log\ p(z_2|H_0) = -\log\ \Gamma(N/2) - N/2\log 2 + (N/2-1)\ \log z_2 - z_2/2.$$

For H_3, we write the likelihood ratio as a function of the unknown parameter β.

$$\frac{p(\mathbf{X}|H_3)}{p(\mathbf{X}|H_0)} = \frac{(2\pi\beta^2)^{-1/2}\ \exp\{-x_1^2/(2\beta^2)\}}{(2\pi)^{-1/2}\exp\{-x_1^2/2\}} \cdot \frac{\prod_{t=2}^{N}(2\pi)^{-1/2}\ \exp\{-x_t^2/2\}}{\prod_{t=2}^{N}(2\pi)^{-1/2}\ \exp\{-x_t^2/2\}}$$
$$= \frac{(2\pi\beta^2)^{-1/2}\ \exp\{-x_1^2/(2\beta^2)\}}{(2\pi)^{-1/2}\ \exp\{-x_1^2/2\}}$$
$$= (\beta^2)^{-1/2}\ \exp\{-x_1^2/(2\beta^2) + x_1^2/2\}.$$

It is clear that the likelihood ratio is a function of $z_3 = \log(x_1^2)$ (taking the log is unnecessary but results in a better-behaved distribution). Thus, no matter what the distribution of σ^2, the likelihood ratio test will depend on the data only through z_3. Therefore, z_3 is an SS for the problem of testing H_3 against H_0. The distribution of z_3 under H_0 is log of Chi-squared with one degree of freedom:

$$\log\ p(z_3|H_0) = -1/2\ \log\ 2\pi + z_3/2 - \exp(z_3)/2.$$

We have shown in this section that z_1 through z_3 are indeed SS for the corresponding unknown parameters (and for the hyothesis tests H_j versus H_0). The features and their PDF's under the H_0 hypothesis are summarized in Table I.

C. Testing the Models under H_0

A crucial step that must be taken prior to proceeding with any CS development is the validation of the H_0 PDF's. Fig. 2 shows the result of comparing histograms of $z_1, z_2,$ and z_3 with theoretical PDF curves. There is an excellent match with the formulas in Table I. It is practically impossible to test the tail probabilities, but validation near the PDF maximum is necessary.

D. Data Generation

Data was generated with $N = 64$ under each class hypothesis using random parameter values.

1) For H_1, α was distributed uniformly in decibels in the interval $[-30, -10]$. The conversion from decibels is $\alpha = 10^{\mathrm{dB}/20}$.
2) For H_2, σ^2 was distributed uniformly in decibels in the interval $[-20, 0]$. The conversion from decibels is $\sigma^2 = 10^{\mathrm{dB}/10}$.
3) For H_3, β^2 was distributed uniformly in decibels in the interval [10, 30]. The conversion from decibels is $\beta^2 = 10^{\mathrm{dB}/10}$.

Due to space limitations, we have not provided histograms of the distributions of the features. However, the distributions of the parameters $\alpha, \sigma^2,$ and β^2 are chosen specifically to provide a significant overlap between the feature densities under H_j and H_0 and to make the classification problem a difficult one.

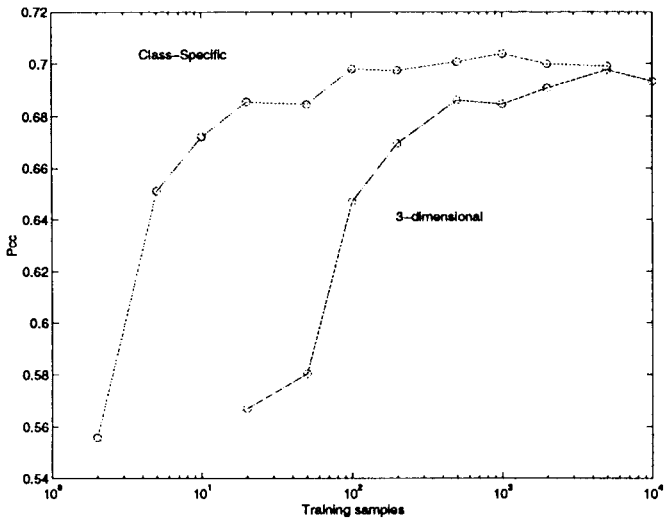

Fig. 3. Probability of correct classification (P_{cc}) as a function of the number of training samples from each class. The upper trace is for the class-specific method. The lower trace is for the traditional method (3-D PDF). Each estimate of P_{cc} is an average of ten independent trials using 500 testing samples from each class in each trial.

E. PDF Estimation

The distributions $p(z_j|H_j)$ for $j = 1$, 2, and 3 were estimated from simulated data using Gaussian mixture approximation [17]. Similarly, for the traditional method, the features were combined into a single feature set set $\mathbf{z} = \{z_1, z_2, z_3\}$, whose PDF was estimated using Gaussian mixtures under each hypothesis. While a 3-D PDF estimation will hardly pose a problem in most situations, feature dimensions up to 50 or 100 are often called for in complex problems. Still, a 3-D PDF requires a healthy amount of training data to accurately characterize. We will see that it needs much more data than the 1-D PDF's of the CS method.

F. Classification Performance

To compare the class-specific method with the traditional method, the following experiment was carried out. A fixed amount of training data, say, N_{train} samples, from each class hypothesis was created. From this data, features z_1 through z_3 were computed. Gaussian mixture approximations of the the PDF's $p(z_1|H_1), p(z_2|H_2), p(z_3|H_3)$ were computed using this data. These PDF's were used in a class-specific classifier using the theoretical denominator PDF's. In addition, the conventional joint PDF's $p(\mathbf{z}|H_1), p(\mathbf{z}|H_2)$, and $p(\mathbf{z}|H_3)$ were estimated and used in a traditional classifier arrangement. A fixed amount of new data, say, N_{test}, for each class was then created for the purpose of measuring the total probability of correct classification (P_{cc}). As N_{train} was varied from as low as 2 samples to as high as 10 000 samples in approximate powers of 2, P_{cc} was determined always using $N_{\text{test}} = 500$. To calculate P_{cc}, the total number of correct decisions in each trial was divided by $3N_{\text{test}}$. The result is plotted in Fig. 3. The figure clearly shows that at least a factor of 10 more data is required by the traditional method for the same level of performance. Even for three fairly well-behaved features, several thousand training samples are needed for optimum performance. About 100 samples are required for minimal performance. This clearly shows the effect of dimensionality on classification performance. For the class-specific method, five samples are needed for minimal performance and 100 for optimum performance. Two claims of this correspondence are supported by the graph: first, that the lower dimensional formulation achieves maximum performance with fewer training samples and second, that both formulations are equivalent (given sufficient data). The latter claim is supported by the asymptotic convergence to similar performance levels.

IV. Conclusions

An exact expression has been derived that provides a way of breaking down the traditional Bayesian minimum error M-ary classifier into low-dimensional distributions. It requires 1) a (small) set of sufficient statistics for each signal class and 2) a common (noise-only) class. The benefit of the class-specific formulation over the optimum Bayesian classifier is clearly demonstrated in a synthetic three-class problem. More that an order of magnitude more training data is required by the traditional approach.

References

[1] D. W. Scott, *Multivariate Density Evaluation*. New York: Wiley, 1992.
[2] S. Aeberhard, D. Coomans, and O. de Vel, "Comparative analysis of statistical pattern recognition methods in high dimensional settings," *Pattern Recognit.*, vol. 27, no. 8, pp. 1065–1077, 1994.
[3] S. J. Raudys and A. K. Jain, "Small sample size effects in statistical pattern recognition: Recommendations for practitioners," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 13, pp. 252–264, Mar. 1991.
[4] Duda and Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.
[5] N. Intrator, "Feature extraction using an exploratory projection pursuit neural network," Ph.D. dissertation, Brown Univ., Providence, RI, 1991.
[6] P. J. Huber, "Projection pursuit," *Ann. Stat.*, vol. 13, no. 2, pp. 435–475, 1985.
[7] P. M. Baggenstoss, "Structural learning for classification of high dimensional data," in *Proc. Int. Conf. Intell. Syst. Semiotics*, 1997, pp. 124–129.
[8] A. Finch, "A neural network for dimension reduction and application to image segmentation," in *Proc. Int. Conf. Artificial Neural Networks (ICANN)*, 1994, pp. 252–264.
[9] H. Watanabe, *Knowing and Guessing*. New York: Wiley, 1969.
[10] T. Kohonen, G. Németh, K.-J. Bry, M. Jalanko, and H. Riittinen, "Spectral classification of phonemes by learning subspaces," in *Proc. ICASSP*, 1979, pp. 97–100.
[11] E. Oja, *Subspace Methods of Pattern Recognition*. New York: Research Studies, 1983.
[12] H. Watanabe and S. Katagiri, "Discriminative subspace method for minimum error pattern recognition," in *Proc. IEEE Workshop Neural Networks Signal Process.*, 1995, pp. 77–86.
[13] E. H. Lehmann, *Theory of Point Estimation*. New York: Wiley, 1983.
[14] H. Watanabe, T. Yamaguchi, and S. Katagiri, "Discriminative metric design for robust pattern recognition," *IEEE Trans. Signal Processing*, vol. 45, pp. 2655–2661, Nov. 1997.
[15] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part 1, Detection, Estimation, and Linear Modulation Theory* New York: Wiley, 1968.
[16] M. Kendall and A. Stuart, *The Advanced Theory of Statistics*, fourth ed. London, U.K., Charles Griffin, 1979, vol 2.
[17] D. M. Titterington, A. F. M. Smith, and U. E. Makov, *Statistical Analysis Of Finite Mixture Distributions*. New York: Wiley, 1985.