Class-Specific Feature Sets in Classification *

Paul M. Baggenstoss
Naval Undersea Warfare Center
Newport RI, 02841
401-832-8240 (TEL)
401-841-7453 (FAX)
p.m.baggenstoss@ieee.org (EMAIL)
EDICS 6.1.6 / 3.5 / 6.17

March 12, 2002

Abstract

The commonly used feature-based classifier implements the maximum aposteriori probability (MAP) of the data class given the features. This requires the joint probability density function (PDF) of the features under each of the class hypotheses. Unfortunately, these PDF's are rarely known and must be estimated from training data. Poor performance results if the amount of training data is insufficient to estimate the high-dimensional feature PDF's. The class-specific theorem is presented in which the MAP decision rule is rewritten as a function of low-dimensional PDF's which may be estimated in practice from far smaller data sets. Necessary conditions include (a) that there exists a low-dimensional feature subset for each class that is a sufficient statistic for the underlying random parameters of each data class, and (b) that there exists at least one point in each parameter space that corresponds to a common PDF. We provide a proof of the theorem supported by an example using synthetic signals. Two orders of magnitude fewer training samples are required by the class-specific approach.

1 Introduction

Consider the classification problem in which a data sample $\mathbf{X}$ is to be classified into one of M classes. This is done optimally by the classifier known as the maximum aposteriori (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* $\mathbf{Z} = T(\mathbf{X})$. Many methods exist for choosing the features, however suppose for now that a *class-specific* strategy is used. One possible class-specific strategy 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]. Because some classes may be similar to each other, it is possible that the M feature sets are not all distinct. Let

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

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

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

*This work supported by the Office of Naval Research

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

The object of the feature selection process is that (2) is equivalent to (1). Thus, they are *sufficient* for the problem at hand. We will see in the theorem that follows, 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 $\mathbf{Z}$. 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 and effort. The complexity of the high dimensional space is such that it becomes impossible to estimate the probability density function (PDF) with a reasonable amount of training data and computational burden. The exponential increase in complexity of systems has been termed the *curse of dimensionality* by Richard Bellman [1]. In complex problems, $\mathbf{Z}$ may be need to contain as many as a hundred features to retain all necessary information. This dimensionality is entirely unmanageable. It is recognized by a number of researchers that attempting to estimate PDF's nonparametrically above 5 dimensions is difficult and above 20 dimensions is futile [2]. Dimensionality reduction is the subject of much research currently and over the past decades (some good overviews are available [3], [2], [4]). Various approaches include feature selection [5], [4], [3], projection pursuit [6], [7], and *independence grouping* [8]. Several other methods are based on projection of the feature vectors onto lower dimensional subspaces [9]. A significant improvement on this is the *subspace method* [10], [11], [12], in which the assumption is less strict in that each class may occupy a different subspace. Improvements on this allow optimization of error performance directly. [13].

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. A simple example that illustrates a situation where this assumption fails is when the classes are distributed in a 3-dimensional volume and arranged in concentric spheres. The classes are not separated when projected on any 1 or 2-dimensional linear subspace. However, statistics based on the radius of the data samples and would constitute a simple 1-dimensional space in which the data is perfectly separated.

Whatever approach one uses, if $\mathbf{Z}$ has a large dimension, and no low-dimensional linear or nonlinear function of the data can be found in which most of the useful information lies, the curse of dimensionality leaves one with the following fundamental choice: either (a) discard much of the useful information in an attempt to reduce the dimension or (b) obtain a crude PDF estimate in the high-dimensional space. In either case, poor performance may result. What we now show is that it is possible to drastically 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). But the *dimensionality bottleneck* cannot be circumvented unless certain *a priori* information is used. In the class-specific method of feature selection introduced above, the fact that that $\mathbf{z}_j$ corresponds to H_j is information that is discarded when $\mathbf{Z}$ is created and is not utilized in (2). We now show how to circumvent the dimensionality bottleneck by utilizing this lost information. This will require two fundamental ideas. The first idea involves defining some common class H_0 which is a subset of all classes. This is possible if all classes have random amplitudes and are embedded in additive noise. Then if H_0 is the noise-only class,

$$H_0 \in H_j, \quad j = 1, 2, \ldots, M.$$

The next idea is to connect the selection of $\mathbf{z}_j$ with the idea of sufficiency.

2 Main Theorem

Theorem 1 *Let there be M distinct PDF families $p(\mathbf{X}|H_j)$, $j = 1, 2, \ldots, M$ where H_j are the class hypotheses. For class each 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 there be a combined feature set $\mathbf{Z} = T(\mathbf{X})$ such that $\mathbf{z}_j \in \mathbf{Z}$, $j = 1, 2, \ldots, M$. Let the PDF $p(\mathbf{Z}|H_j)$ 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:

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

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: Clearly from (3), we have

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

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 $\boldsymbol{\theta}_j^0$.

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

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

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

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

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

Which is the same as (4) □

We therefore have shown that it is possible to reduce the dimensionality, yet end up with a classifier theoretically equivalent to the classifier based on the full-dimensional feature set. It is noted by Kay [14] that under the same assumptions necessary for the above theorem, (4) is equivalent to (1), thus (4) is fully equivalent to the MAP classifier based on $\mathbf{X}$.

While the reduction of the high-dimensional problem to a low-dimensional problem is significant enough, another significant idea emerges revolving around the idea of sufficiency. If $\{\mathbf{z}_j\}$ are sufficient (in the Neyman-Fisher sense) for the parameterizations of the corresponding class, and a common class H_0 can be found, then $\mathbf{Z}$ is sufficient for the classification problem at hand [14].

It is also important to note that while the parameter distributions $p(\boldsymbol{\theta}_j|H_j)$ are used in the proof, they are not required in practice. All that is required are estimates of the low-dimensional PDF's $p(\mathbf{z}_j|H_j)$.

3 Classifier Architecture

The formulation (4) suggests a detector/classifier architecture as shown in Figure 1. Each data class corresponds to a distinct and independent branch in the diagram (the third branch has been expanded for reasons which we will explain below). 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.

In the figure, we have shown a case when for a given class H_3, there may be a variety of sub-classes indexed by a parameter θ. It is possible to carry out a maximization over θ prior to normalization by $p(\mathbf{z}_j|H_0)$.

The common class H_0 does not need to be a real class. Technically, the only requirement is that the parameter sets of each class must include H_0 as a special case, thus the natural role of the noise-only hypothesis. We have found it useful that H_0 represent the condition that $\mathbf{X}$ be samples of *iid* Gaussian noise.

A method of system identification (or model selection) is implied by the structure as well. Suppose that a quantity of training data is available from some data source and that it is desired to fit the data to a model. Let H_1 and H_2 be two candidate models (i.e. two candidate feature sets $\mathbf{z}_1$, $\mathbf{z}_2$) that have each been trained on the same training data. In spite of the fact that H_1 and H_2 may have differing structures, have different model orders or complexities, and may be based on different feature sets it is possible to directly compare the two model outputs do determine which model is better. The optimality of this comparison will be in the sense of minimum probability of error when H_1 and H_2 are regarded as separate hypotheses. If one of the model outputs is greater on average than the other model output for the same population of testing data, then it will likely make a better model for classifying the signal class from other signal classes when a class-specific structure is used. Thus, we have a method of choosing a model for a signal class that is *independent of other signal classes*, yet finds the best model for classifying a signal *against other signal classes*. This is an entirely new classifier design paradigm.

While the class-specific architecture is not new [15], this is the first time it has been placed on any theoretical relationship to the MAP classifier. Theorem 1 shows clearly how the various branches of the structure are normalized and compared in order to achieve the optimal performance of the MAP classifier. It also shows that normalization by the likelihood of the common class H_0 is necessary to allow the outputs to be compared fairly. Without any further knowledge about the class likelihood functions, it represents the architecture with the smallest possible feature dimension that is still equivalent to the optimum Bayesian classifier.

While Theorem 1 requires very specific conditions to hold, specifically the sufficiency of the feature sets and the existence of a common class, it is reasonable to ask where approximations may be made. We have found from experience that while the sufficiency of the various statistics can be relaxed somewhat, and approximations to the various likelihood functions may be made, the likelihood functions under H_0 cannot be approximated without careful attention to the tails. In practice, $\mathbf{X}$ may vary significantly from H_0, especially at high SNR. Thus, it is necessary in many cases to use exact analytic expressions for $p(\mathbf{z}_j|H_0)$. This may seem to be an overly restrictive requirement at first. But, in most cases solutions can be found, especially if H_0 is chosen as *iid* Gaussian noise.

4 Practical Considerations

For real-world problems, the sufficiency of features can never be established. How, then can the method be used? The simple answer is that sufficiency is not really required in practice. Sufficiency is required to establish the *exact* relationship of the class-specific classifier to the optimum Bayesian classifier. If sufficiency is approximated, so is this relationship. Compare the class-specific (CS) approach with the full-dimensional (FD) approach. With CS, if the feature dimensions are low, one can have a good PDF approximation of approximate sufficient statistics. However, in the FD approach, one has the choice of a *very poor* PDF estimate of the full feature set, or a *good* PDF estimate of a *sorely inadequate* feature set.

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, a numerical problem arises for feature vectors which differ greatly from the noise-only hypothesis (i.e. high-SNR). Then, the denominator density $p(\mathbf{z}_j|H_0)$ will be outside its useful range in which it can approximate the density. We are left with these choices:

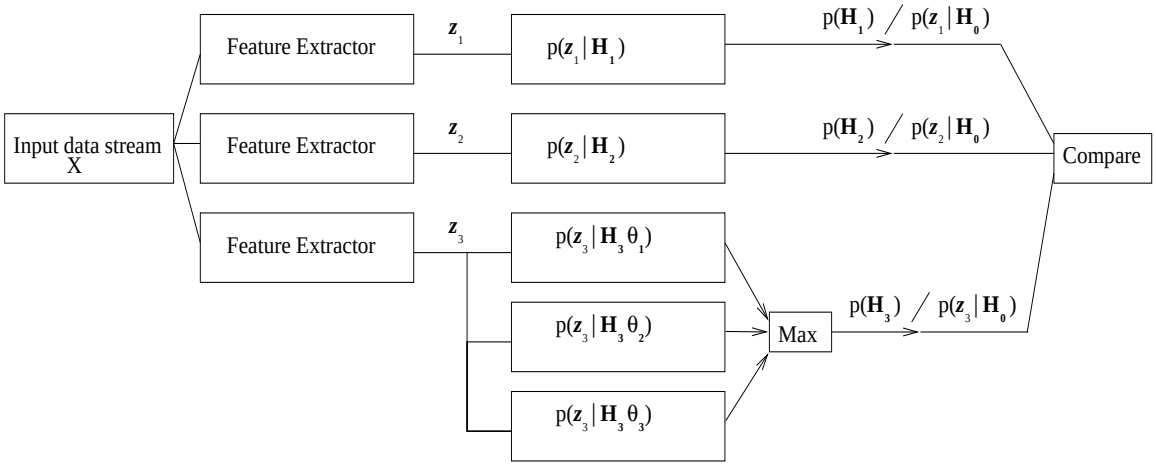

Figure 1: Detector/Classifier Architecture

1. Obtain theoretical densities under H_0 by deriving them analytically. This is aided by the fact that the number of features is (hopefully) small and that H_0 is straight-forward (i.e. iid Gaussian noise).

2. Obtain analytic expressions for the characteristic function, then apply the inverse Fourier transform numerically.

3. Use an asymptotic analysis of the tail behavior of $p(\mathbf{z}_j|H_0)$.

4. If the primary issue in causing $\mathbf{z}_j$ to be out on the tails of $p(\mathbf{z}_j|H_0)$ is due to signal strength, then the following decomposition is useful. Let $\mathbf{z}_j = \{z_j^1, z_j^2, \ldots, z_j^K\}$ and let z_j^1 be a measure of signal strength. Then

$$p(\mathbf{z}_j|H_0) = p(z_j^2, \ldots, z_j^K|H_0, z_j^1)p(z_j^1|H_0)$$

   It may be, as it is for example in ARMA or AR parameter estimates, that $z_j^2, \ldots, z_j^K$ are independent of z_j^1 under H_0. In this case, $p(\mathbf{z}_j|H_0) = p(z_j^2, \ldots, z_j^K|H_0)p(z_j^1|H_0)$ where the term $p(z_j^1|H_0)$ may be known analytically and bears the brunt of the tail behavior.

5 Example Problem

5.1 Signal Classes

The purpose of the example is to offer a controlled experiment using synthetic signals. These 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 readers 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.

The signals are produced with random signal amplitudes distributed from very weak to moderate. A significant number of incorrect classifications are expected, even for the limiting case. We consider 9 data classes denoted $H_1, \ldots, H_9$.

- Class H_0: Noise only

- Class H_1: Long Sinewave
- Class H_2: Medium Sinewave
- Class H_3: Short Sinewave
- Class H_4: Long Gaussian Signal
- Class H_5: Short Gaussian Signal
- Class H_6: Short Impulse Signal
- Class H_7: Long Impulse Signal
- Class H_8: Long Laplacian Distributed Noise
- Class H_9: Short Laplacian Distributed Noise

Examples of these signals are provided in Figure 2. Mathematical descriptions of signals used in the simulation are described in enough detail in the Appendix so that the reader may recreate the experiment. Sufficient (or approximately sufficient) statistics $\mathbf{z}_j$ are provided in Table 1. The distributions $p(\mathbf{z}_j|H_0)$, where H_0 is the condition that $\mathbf{X}$ is *iid* Gaussian noise of unit variance, are provided in Table 2.

5.2 Results

A total of 16384 samples from each of classes H_1 through H_9 were created. Each sample consisted of a statistically independent realization of a time series of length $N = 256$ generated under the corresponding hypothesis. For each hypothesis, the values of pertinent model parameters were selected at random as described in the Appendix. For each time series produced, the the statistics (features) $\mathbf{z}_1, \ldots, \mathbf{z}_9$ were computed.

As a check on the determination of theoretical PDF under H_0, data was also generated for pure Gaussian noise. Histograms of the H_0 distributions overlaid on the theoretical curves are provided in Figure 3. Notice that $\mathbf{z}_8$ and $\mathbf{z}_9$ are two-dimensional and a planar plot is needed.

The feature data was used in holdout trials to determine probability of correct classification (P_{cc}) as a function of the number of training samples (NTRAIN). For each value of NTRAIN, four independent trials were performed. For each trial, the data was divided randomly into training and testing portions. All the data not used in training was used in testing (i.e. for determining P_{cc}). The average of the four trials was plotted. The results of the experiment are provided in Figure 4 for three classifiers:

1. K-nearest neighbor classifier with $K = 3$. The nearest 3 training samples were located in each class training set and the distance to the farthest of these 3 was used as a classification statistic. Distance was computed using "data sphering", that is, the distance from a feature vector to a training sample of a particular class was computed in a coordinate system in which the features of that class were uncorrelated. The features were decorrelated using an estimate of the feature covariance obtained from the training data.

2. Full-dimensional (FD) classifier implementing equation (2). The distributions $p(\mathbf{Z}|H_j)$ were estimated from the training data using a Gaussian Mixture estimate obtained using the E-M algorithm [16],[17].

3. Class-specific (CS) classifier implementing equation (4). The distributions $p(\mathbf{z}_j|H_j)$ were estimated from the training data using a Gaussian Mixture estimate.

Two claims of this paper are supported by the graph. First that the lower dimensional formulation achieves maximum performance with fewer training samples. Second, that both formulations are equivalent (given sufficient data). The latter claim is supported by the asymptotic convergence to similar performance levels. Of course, the approximations used for classes H_8, H_9 could account for some sub-optimal behavior of the class-specific formulation. Due to practical limitations, the FD performance could not be evaluated at higher than 8192 training samples.

Further evidence that the two formulations are approximately equivalent is obtained from the confusion matrices of the FD and CS classifiers for 8192 and 128 training samples, respectively are provided in Tables 3,4.

$\mathbf{z}_1 = \log\left\{\left[\sum_{i=1}^{N} x_i \cos(\omega i)\right]^2 + \left[\sum_{i=1}^{N} x_i \sin(\omega i)\right]^2\right\}$
$\mathbf{z}_2 = \log\left\{\left[\sum_{i=1}^{N/2} x_i \cos(\omega i)\right]^2 + \left[\sum_{i=1}^{N/2} x_i \sin(\omega i)\right]^2\right\}$
$\mathbf{z}_3 = \log\left\{\left[\sum_{i=1}^{N/4} x_i \cos(\omega i)\right]^2 + \left[\sum_{i=1}^{N/4} x_i \sin(\omega i)\right]^2\right\}$
$\mathbf{z}_4 = \sum_{i=1}^{N} x_i^2$
$\mathbf{z}_5 = \sum_{i=1}^{N/2} x_i^2$
$\mathbf{z}_6 = \log(x_1^2)$
$\mathbf{z}_7 = \log(x_1^2 + x_2^2)$
$\mathbf{z}_8 = \begin{bmatrix} \sum_{i=1}^{N} \lvert x_i \rvert \\ \sum_{i=1}^{N} x_i^2 \end{bmatrix}$
$\mathbf{z}_9 = \begin{bmatrix} \sum_{i=1}^{N/2} \lvert x_i \rvert \\ \sum_{i=1}^{N/2} x_i^2 \end{bmatrix}$

Table 1: Class-Specific Statistics

$p(\mathbf{z}_1\|H_0) = \left(\frac{e^{\mathbf{z}_1}}{N\sigma^2}\right)\exp\left\{-\frac{e^{\mathbf{z}_1}}{N\sigma^2}\right\}$
$p(\mathbf{z}_2\|H_0) = \left(\frac{2e^{\mathbf{z}_2}}{N\sigma^2}\right)\exp\left\{-\frac{2e^{\mathbf{z}_2}}{N\sigma^2}\right\}$
$p(\mathbf{z}_3\|H_0) = \left(\frac{4e^{\mathbf{z}_3}}{N\sigma^2}\right)\exp\left\{-\frac{4e^{\mathbf{z}_3}}{N\sigma^2}\right\}$
$p(\mathbf{z}_4\|H_0) = \frac{1}{\sigma^2}\Gamma^{-1}(\frac{N}{2})\ 2^{-\frac{N}{2}}(\frac{\mathbf{z}_4}{\sigma^2})^{\frac{N}{2}-1}\ \exp\left\{-\frac{\mathbf{z}_4}{2\sigma^2}\right\}$
$p(\mathbf{z}_5\|H_0) = \frac{1}{\sigma^2}\Gamma^{-1}(\frac{N}{4})\ 2^{-\frac{N}{4}}(\frac{\mathbf{z}_5}{\sigma^2})^{\frac{N}{2}-1}\ \exp\left\{-\frac{\mathbf{z}_5}{2\sigma^2}\right\}$
$p(\mathbf{z}_6\|H_0) = \left(2\pi\sigma^2\right)^{-1/2}\exp\left\{-\frac{e^{\mathbf{z}_6}}{2\sigma^2}\right\}e^{\mathbf{z}_6/2}$
$p(\mathbf{z}_7\|H_0) = \left(4\pi\sigma^2\right)^{-1/2}\exp\left\{-\frac{e^{\mathbf{z}_7}}{4\sigma^2}\right\}e^{\mathbf{z}_7/2}$
Gaussian for $N \to \infty$: $E\left(\mathbf{z}_8\|H_0\right) = N\begin{bmatrix}\sqrt{\frac{2}{\pi}} \\ 1\end{bmatrix}$ $\mathrm{cov}\left(\mathbf{z}_8\|H_0\right) = N\begin{bmatrix}1-\frac{2}{\pi} & \sqrt{\frac{2}{\pi}} \\ \sqrt{\frac{2}{\pi}} & 2\end{bmatrix}$
Gaussian for $N \to \infty$: $E\left(\mathbf{z}_9\|H_0\right) = \frac{N}{2}\begin{bmatrix}\sqrt{\frac{2}{\pi}} \\ 1\end{bmatrix}$ $\mathrm{cov}\left(\mathbf{z}_9\|H_0\right) = \frac{N}{2}\begin{bmatrix}1-\frac{2}{\pi} & \sqrt{\frac{2}{\pi}} \\ \sqrt{\frac{2}{\pi}} & 2\end{bmatrix}$

Table 2: Distributions of Class-Specific Statistics

	Declared as class:								
	1	2	3	4	5	6	7	8	9
Class:									
1	70	8	7	3	2	3	2	0	0
2	8	57	15	5	3	5	3	0	0
3	6	12	53	7	6	7	5	0	0
4	3	5	7	59	9	4	3	5	1
5	3	7	11	12	35	5	5	0	18
6	3	4	7	5	3	63	10	0	0
7	3	3	6	3	3	11	67	0	0
8	0	0	0	4	0	0	0	94	0
9	0	0	0	1	9	0	0	0	88

Table 3: Confusion Matrix for Full-Dim (FD) classifier at 8192 training samples in percent.

	Declared as class:								
	1	2	3	4	5	6	7	8	9
Class:									
1	68	6	6	5	6	4	2	0	0
2	5	58	11	5	8	5	3	0	0
3	4	8	52	9	11	8	5	0	0
4	2	3	4	60	12	4	2	8	0
5	3	3	6	14	35	6	3	0	27
6	1	2	7	5	7	65	8	0	0
7	1	3	5	4	6	11	64	0	2
8	0	0	0	1	0	0	0	94	4
9	0	0	0	1	13	0	0	0	84

Table 4: Confusion Matrix for Class-Specific (CS) classifier at 128 training samples in percent.

More evidence to concerning the equivalence of the two classifiers is obtained as follows. It is clear that for any two classes A,B,

$$\frac{p(\mathbf{Z}|H_A)}{p(\mathbf{Z}|H_B)} = \frac{p(\mathbf{z}_A|H_A)/p(\mathbf{z}_A|H_0)}{p(\mathbf{z}_B|H_B)/p(\mathbf{z}_B|H_0)} \tag{7}$$

Thus, the ratios between the classifier outputs for any two classes is the same for the FD or CS classifiers. This equivalence may be tested experimentally by plotting the right-hand side (RHS) of the above equality on one axis and the left-hand side (LHS) on the other axis. Refer now to Figure 5 which plots the LHS of (7) vs. the RHS side for hypotheses 4 and 1. Notice that as the amount of training data increases, equality is approximated. In Figure 6, the same graph is plotted for hypotheses 8,1. Notice that the trend to equality is less dramatic. This may be explained by the approximate form of $p(\mathbf{z}_8|H_0)$ that was used in the simulation.

6 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 9-class problem. More that 2 orders of magnitude more training data is required by the traditional approach. These improvements were obtained in spite of the use of approximations to sufficient statistics and to their distributions.

7 Appendix

In this section we provide details of the sufficient statistics required for each hypothesis.

7.1 Class H_0: Noise only

The noise-only class is characterized by pure *iid* Gaussian noise.

$$p(\mathbf{X}|H_0) = (2\pi\sigma^2)^{-N/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N} x_i^2\right\}$$

7.2 Class H_1: Long Sinewave

Class H_1 is a sinewave of random positive amplitude a and uniformly distributed random phase in Gaussian *iid* noise of known variance σ^2. Let $a = 10^{b/20}$ where b is uniformly distributed on [-20,0]. We have

$$p(\mathbf{X}|a,\theta) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}[x_i - a\cos(\omega i + \theta)]^2\right\}$$

The LR test statistic given a

$$\frac{\int_\theta p(\mathbf{X}|a,\theta)p(\theta)d\theta}{p(\mathbf{X}|a=0)}$$

may be reduced [18] to a monotonic function of

$$q^2 = c^2 + s^2$$

where

$$c \triangleq \sum_{i=1}^{N} x_i \cos(\omega i) \qquad s \triangleq \sum_{i=1}^{N} x_i \sin(\omega i)$$

Thus, q is a sufficient statistic for a. We choose

$$\mathbf{z}_1 = \log(q^2)$$

Note that under H_0 the term c and s each are Gaussian with zero mean and variance $\frac{N\sigma^2}{2}$. We see, then that $\frac{2}{N\sigma^2}q^2 = \frac{2}{N\sigma^2}(c^2+s^2)$ is distributed $\chi^2(2)$. Let $u \sim \chi^2(r)$, then

$$p(u|H_0) = \Gamma^{-1}(r/2)\ 2^{-r/2}\ u^{r/2-1}\ e^{-u/2}$$

Let $y = ku$, then

$$p(y|H_0) = \frac{1}{k}\ \Gamma^{-1}(r/2)\ 2^{-r/2}\ (y/k)^{r/2-1}\ e^{-\frac{y}{2k}}.$$

Now let $z = \log(y)$, then

$$p(z|H_0) = \frac{1}{k}\ \Gamma^{-1}(r/2)\ 2^{-r/2}(k)^{1-r/2}\ e^{zr/2} \exp\left\{-\frac{e^z}{2k}\right\} \tag{8}$$

The distribution of $\mathbf{z}_1 = \log(q^2)$ is obtained by letting $r = 2, k = \frac{N\sigma^2}{2}$, thus

$$p(\mathbf{z}_1|H_0) = \left(\frac{e^{\mathbf{z}_1}}{N\sigma^2}\right) \exp\left\{-\frac{e^{\mathbf{z}_1}}{N\sigma^2}\right\}.$$

7.3 Class H_2: Medium Sinewave

Class H_2 is the same as H_1 except the signal covers only the first half of $\mathbf{X}$. The desire is to create a class highly correlated with H_1. Let $a = 10^{b/20}$ where b is uniformly distributed on [-20,0]. Let

$$q^2 = c^2 + s^2$$

where

$$c \triangleq \sum_{i=1}^{N/2} x_i \cos(\omega i) \qquad s \triangleq \sum_{i=1}^{N/2} x_i \sin(\omega i)$$

$$\mathbf{z}_2 = \log(q^2)$$

Note that under H_0, c and s are each Gaussian with zero mean and variance $\frac{N\sigma^2}{4}$. Following the derivation for $p(\mathbf{z}_1)$, we have

$$p(\mathbf{z}_2|H_0) = \left(\frac{2e^{\mathbf{z}_2}}{N\sigma^2}\right) \exp\left\{-\frac{2e^{\mathbf{z}_2}}{N\sigma^2}\right\}$$

7.4 Class H_3: Short Sinewave

Class H_3 is the same as H_1 except the signal covers only the first quarter of $\mathbf{X}$. Let $a = 10^{b/20}$ where b is uniformly distributed on [-20,0]. Let

$$q^2 = c^2 + s^2$$

where

$$c \triangleq \sum_{i=1}^{N/4} x_i \cos(\omega i) \qquad s \triangleq \sum_{i=1}^{N/4} x_i \sin(\omega i)$$

$$\mathbf{z}_3 = \log(q^2)$$

Note that under H_0, c and s are each Gaussian with zero mean and variance $\frac{N\sigma^2}{8}$. Following the derivation for $p(\mathbf{z}_1)$, we have

$$p(\mathbf{z}_3|H_0) = \left(\frac{4e^{\mathbf{z}_3}}{N\sigma^2}\right) \exp\left\{-\frac{4e^{\mathbf{z}_3}}{N\sigma^2}\right\}$$

7.5 Class H_4: Long Gaussian Signal

Class H_4 is Gaussian signal in Gaussian noise. Let σ_s be the signal variance. We have

$$p(\mathbf{X}|H_4) = \left[2\pi(\sigma^2+\sigma_s^2)\right]^{-N/2} \exp\left\{-\frac{1}{2(\sigma^2+\sigma_s^2)}\sum_{i=1}^{N} x_i^2\right\}$$

Let $\sigma_t = \sqrt{\sigma^2+\sigma_s^2}$. In the simulation, realizations of σ_t were produced according to $\sigma_t = \sigma + 10^{b/20}$ where b is uniformly distributed on [-30,-10]. As a sufficient statistic for σ_s^2, we choose

$$\mathbf{z}_4 = \sum_{i=1}^{N} x_i^2$$

Noting that $\mathbf{z}_4/\sigma^2$ is $\chi^2(N)$

$$p(\mathbf{z}_4|H_0) = \frac{1}{\sigma^2}\Gamma^{-1}(N/2)\ 2^{-N/2}(\frac{\mathbf{z}_4}{\sigma^2})^{N/2-1} \exp\left\{-\frac{\mathbf{z}_4}{2\sigma^2}\right\}$$

7.6 Class H_5: Short Gaussian Signal

Class H_5 is the same as H_4 but occupies only the first half of $\mathbf{X}$. As before, $\sigma_t = \sqrt{\sigma^2+\sigma_s^2}$. Realizations of σ_t were produced according to $\sigma_t = \sigma + 10^{b/20}$ where b is uniformly distributed on [-16,4]. As a sufficient statistic for σ_s^2, we choose

$$\mathbf{z}_5 = \sum_{i=1}^{N/2} x_i^2$$

Noting that $\mathbf{z}_5/\sigma^2$ is $\chi^2(N/2)$

$$p(\mathbf{z}_5|H_0) = \frac{1}{\sigma^2}\Gamma^{-1}(N/4)\ 2^{-N/4}(\mathbf{z}_5/\sigma^2)^{N/2-1} \exp\left\{-\frac{\mathbf{z}_5}{2\sigma^2}\right\}$$

7.7 Class H_6: Short Impulse Signal

Class H_6 is an impulsive signal occurring on the first data sample x_1. Let

$$\begin{aligned} x_i &= as + n_i, \quad i = 1 \\ x_i &= n_i, \quad i = 2, 3, \ldots, N \end{aligned}$$

where n_i are *iid* Gaussian random variables variables with mean zero and variance σ^2, s equals -1 or 1 with equal probability, and $a = 10^{b/20}$ where b is uniformly distributed on [-2,18]. We choose as a sufficient statistic

$$\mathbf{z}_6 = \log(x_1^2)$$

The argument of the log times $1/\sigma^2$ is distributed $\chi^2(1)$, thus from (8),

$$\begin{aligned} p(\mathbf{z}_6|H_0) &= \Gamma^{-1}(1/2)\tfrac{1}{\sqrt{2\sigma^2}}\ e^{\mathbf{z}_6/2}\ \exp\left\{-\tfrac{e^{\mathbf{z}_6}}{2\sigma^2}\right\} \\ &= \left(2\pi\sigma^2\right)^{-1/2} \exp\left\{-\tfrac{e^{\mathbf{z}_6}}{2\sigma^2}\right\} e^{\mathbf{z}_6/2} \end{aligned}$$

7.8 Class H_7: Long Impulse Signal

Class H_7 is similar to H_6 but occupies the first 2 data samples. Let

$$\begin{aligned} x_i &= as + n_i, \quad i = 1, 2 \\ x_i &= n_i, \quad i = 3, 4, \ldots, N \end{aligned}$$

where n_i are *iid* Gaussian random variables variables with mean zero and variance σ^2, s equals -1 or 1 with equal probability, and $a = 10^{b/20}$ where b is uniformly distributed on [-4,16]. Let

$$\mathbf{z}_7 = \log((x_1 + x_2)^2)$$

The argument of the log times $\frac{1}{2\sigma^2}$ is distributed $\chi^2(1)$, thus from (8),

$$\begin{aligned} p(\mathbf{z}_7|H_0) &= \Gamma^{-1}(1/2)\tfrac{1}{\sqrt{4\sigma^2}}\, e^{\mathbf{z}_7/2}\, \exp\left\{-\tfrac{e^{\mathbf{z}_7}}{4\sigma^2}\right\} \\ &= \left(4\pi\sigma^2\right)^{-1/2} \exp\left\{-\tfrac{e^{\mathbf{z}_7}}{4\sigma^2}\right\} e^{\mathbf{z}_7/2} \end{aligned}$$

7.9 Class H_8: Long Laplacian Distributed Noise

This is a heavy-tailed type of random signal. Under signal only, x_i is distributed

$$p(x_i) = \frac{1}{2\rho} e^{-|x_i|/\rho}.$$

When Gaussian noise of variance σ^2 is added, the distribution is [thanks to Tod Luginbuhl for this formula]

$$p(x_i|\rho^2) = \frac{1}{2\rho} e^{\frac{\sigma^2}{2\rho^2}} \left\{ e^{x_i/\rho} \frac{\operatorname{erf}\left(-\frac{x_i}{\sigma\sqrt{2}} - \frac{\sigma}{\rho\sqrt{2}}\right) + 1}{2} - e^{-x_i/\rho} \frac{\operatorname{erf}\left(-\frac{x_i}{\sigma\sqrt{2}} + \frac{\sigma}{\rho\sqrt{2}}\right) - 1}{2} \right\}$$

No sufficient statistic for ρ is immediately obvious from this distribution. We choose as a pair of statistics

$$\mathbf{z}_8 = \begin{bmatrix} \sum_{i=1}^N |x_i| \\ \sum_{i=1}^N x_i^2 \end{bmatrix}$$

This will provide an example where sufficiency is only approximate. The two-dimensional characteristic function of the distribution of $\mathbf{z}_8$ under H_0 is

$$\begin{aligned} \Phi_8(\omega_1, \omega_2) &= \left(\int_x \int_y p_x(x) \delta(y - \sqrt{x}) e^{-j\omega_1 x} e^{-j\omega_1 y} dx dy \right)^N \\ &= \left(\int_0^\infty p_x(x) e^{-j\omega_1 x} e^{-j\omega_2 \sqrt{x}} dx \right)^N \end{aligned}$$

where x is a χ^2 random variable, y is the absolute value of a normal random variable (χ-distributed), and the $p_x(x)$ is the χ^2 distribution

$$p_x(x) = \Gamma^{-1}(1/2)\, 2^{-1/2}\, x^{-1/2}\, e^{-x/2}$$

Solution is possible by inverse Fourier transform, however it is not a simple task, especially when tail behavior is needed. We note, however, that by applying the Central Limit theorem for large N, the distribution of $\mathbf{z}_8$ is approximately Gaussian with mean and covariance

$$E(\mathbf{z}_8|H_0) = N \begin{bmatrix} \sqrt{\frac{2}{\pi}} \\ 1 \end{bmatrix}$$

$$\mathrm{cov}\left(\mathbf{z}_8|H_0\right) = N\begin{bmatrix} 1-\frac{2}{\pi} & \sqrt{\frac{2}{\pi}} \\ \sqrt{\frac{2}{\pi}} & 2 \end{bmatrix}$$

Using the asymptotic distribution will test the robustness of the overall technique, however it is not recommended in general to use the Central Limit theorem for the H_0 density because accurate tail behavior is needed.

7.10 Class H_9: Short Laplacian Distributed Noise

This is identical to H_8 except the signal covers the first $N/2$ samples. In a similar manner,

$$\mathbf{z}_9 = \begin{bmatrix} \sum_{i=1}^{N/2} |x_i| \\ \sum_{i=1}^{N/2} x_i^2 \end{bmatrix}$$

The distribution of $\mathbf{z}_9$ under H_0 is approximately Gaussian with mean and covariance

$$E\left(\mathbf{z}_9|H_0\right) = \frac{N}{2}\begin{bmatrix} \sqrt{\frac{2}{\pi}} \\ 1 \end{bmatrix}$$

$$\mathrm{cov}\left(\mathbf{z}_9|H_0\right) = \frac{N}{2}\begin{bmatrix} 1-\frac{2}{\pi} & \sqrt{\frac{2}{\pi}} \\ \sqrt{\frac{2}{\pi}} & 2 \end{bmatrix}$$

References

[1] R. E. Bellman, *Adaptive Control Processes*. Priceton, New Jersey, USA: Princeton Univ. Press, 1961.

[2] D. W. Scott, *Multivariate Density Estimation*. Wiley, 1992.

[3] S. Aeberhard, D. Coomans, and O. de Vel, "Comparative analysis of statistical pattern recognition methods in high dimensional settings," *Pattern Recognition*, vol. 27, no. 8, pp. 1065–1077, 1994.

[4] S. J. Raudys and A. K. Jain, "Small sample size effects in statistical pattern recognition: Recommendations for practitioners," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 13, no. 3, pp. 252–264, 1991.

[5] Duda and Hart, *Pattern Classification and Scene Analysis*. Wiley, 1973.

[6] N. Intrator, *Feature Extraction Using an Exploratory Projection Pursuit Neural Network*. PhD thesis, Brown University, 1991.

[7] P. J. Huber, "Projection pursuit," *Annals of Statistics*, vol. 13, no. 2, pp. 435–475, 1985.

[8] P. M. Baggenstoss, "Structural learning for classification of high dimensional data," in *Proceedings of the 1997 International Conference on Intelligent Systems and Semiotics*, pp. 124–129, National Institute of Standards and Technology, 1997.

[9] A. Finch, "A neural network for dimension reduction and application to image segmentation," in *Proceedings of the 1994 International Conference on Artificial Neural Networks (ICANN-94)*, pp. 252–264, 1994.

[10] H. Watanabe, *Knowing and Guessing*. New York: John Wiley, 1969.

[11] T. Kohonen, G. Németh, K.-J. Bry, M. Jalanko, and H. Riittinen, "Spectral classification of phonemes by learning subspaces," in *Proc. ICASSP 79*, pp. 97–100, 1979.

[12] E. Oja, *Subspace Methods of Pattern Recognition*. Research Studies Press, 1983.

[13] H. Watanabe and S. Katagiri, "Discriminative subspace method for minimum error pattern recognition," in *Proc. 1995 IEEE Workshop on Neural Networks for Signal Processing*, pp. 77–86, 1995.

[14] S. Kay, "Sufficiency, classification, and the class-specific feature theorem," *IEEE Trans. Information Theory*, vol. 46, pp. 1654–1658, July 2000.

[15] H. Watanabe, T. Yamaguchi, and S. Katagiri, "Discriminative metric design for robust pattern recognition," *IEEE Trans. Signal Processing*, vol. 45, no. 11, pp. 2655–2661, 1997.

[16] R. L. Streit, "A neural network for optimum Neyman-Pearson classification," in *Proc. International Joint Conference on Neural Networks*, (San Diego, California), pp. 685–690, June 1990.

[17] L. I. Perlovsky, "A model-based neural network for transient signal processing," *Neural Networks*, vol. 7, no. 3, pp. 565–572, 1994.

[18] A. D. Whalen, *Detection of Signals in Noise*. Academic Press, 1971.

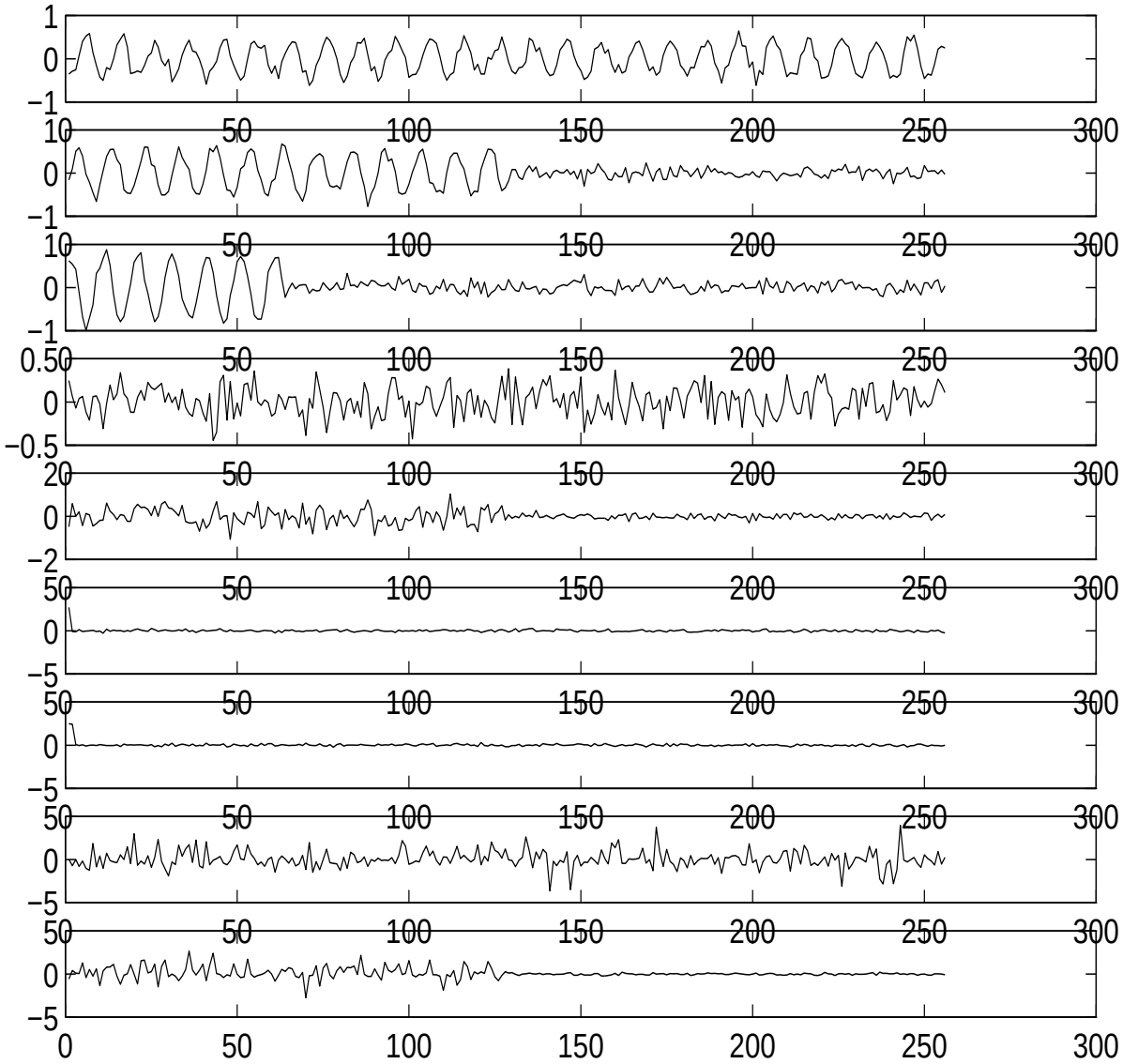

Figure 2: Examples of the nine signal types. Signal-to-Noise (SNR) has been increased for clarity. Actual SNR varies.

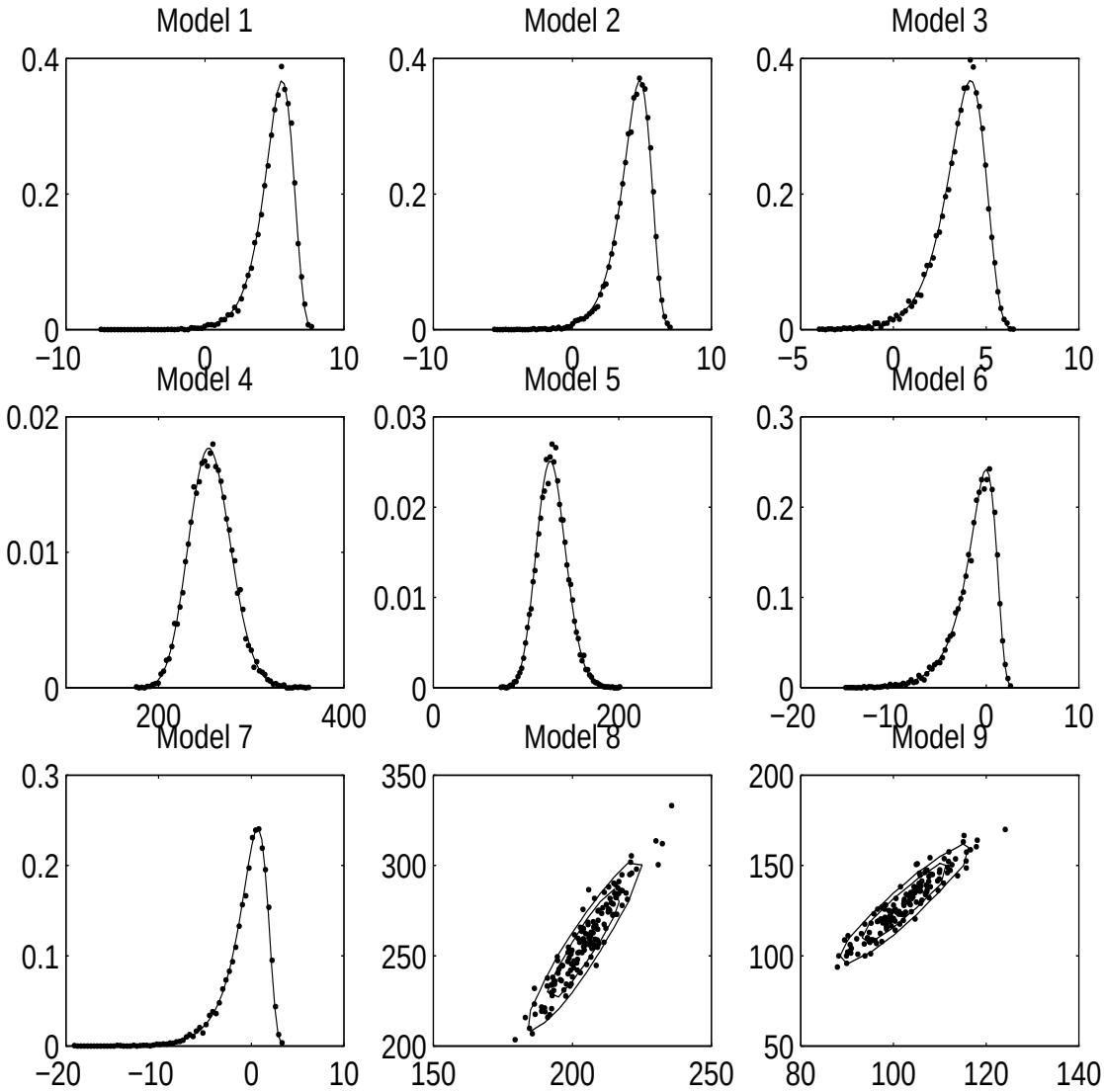

Figure 3: Histograms of statistics under H_0 with theoretical distributions.

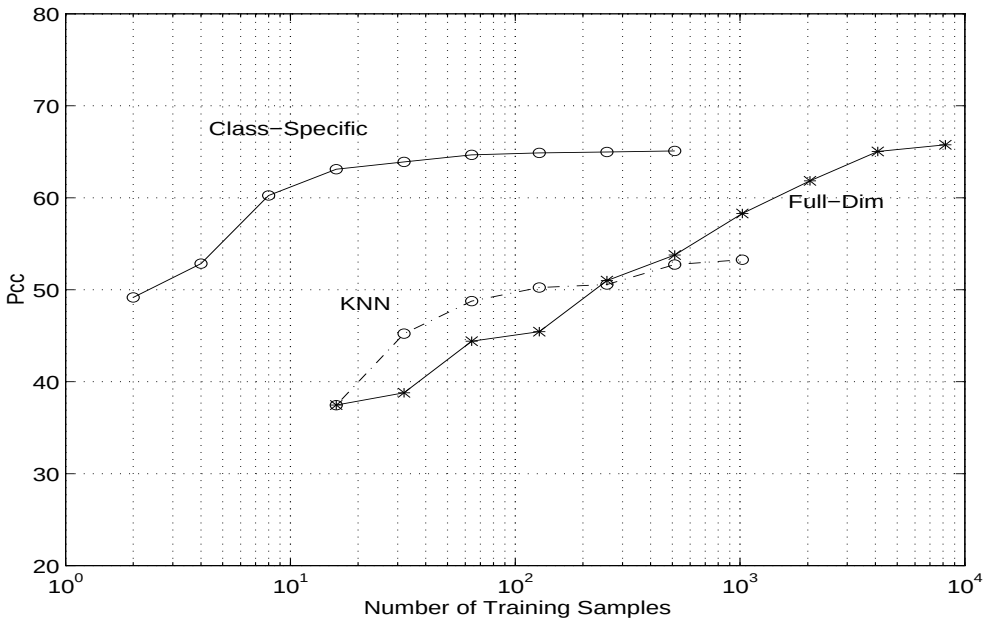

Figure 4: Percent correct vs. number of training samples for three classifiers. Each data point is the average of 4 independent trials.

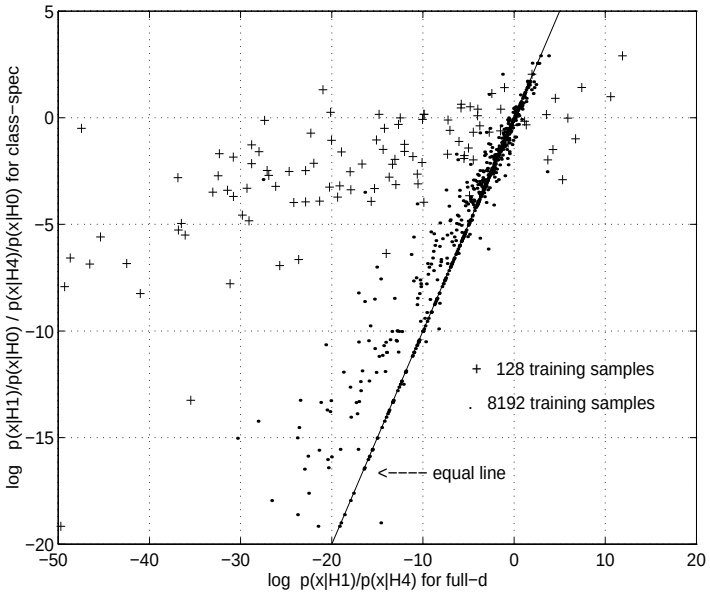

Figure 5: Comparison of classifier statistics: Full-D vs. Class-Specific, difference of outputs for classes 4 and 1 hypotheses, Gaussian noise data

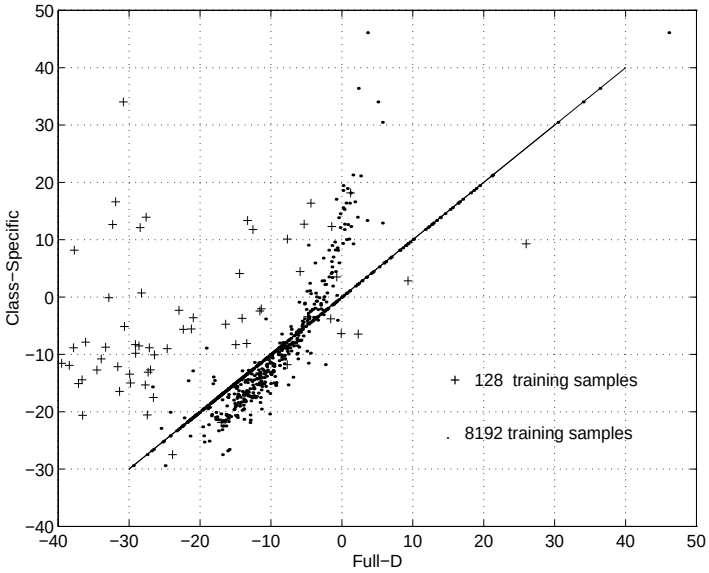

Figure 6: Comparison of classifier statistics: Full-D vs. Class-Specific, difference of outputs for classes 8 and 4 hypotheses, Gaussian noise data