MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Classification in Likelihood Spaces

Rita Singh, Bhiksha Raj

TR2004-089 December 2004

Abstract

In classification methods that explicitly model class-conditional probability distributions, the true distributions are often not known. These are estimated from the data available, to approximate the true distributions. Errors in classificaiton that arise due to this approximation can be reduced to some extent if the estimated distributions are used merely to project data into a space of likelihoods and classification is performed in that space suing discriminant functions. In this article, we discuss the rationale behind this, and also the general properties of likelihood projections. We demonstrate the utility of likelihood projections in improving classification performance through experiments carried out on a standard image database and a standard speech database.

Technometrics

Classification in Likelihood Spaces

Rita Singh

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
(*rsingh@cs.cmu.edu*)

Bhiksha Raj

Mitsubishi Electric Research Labs
Cambridge, MA 02139
(*bhiksha@merl.com*)

In classification methods that explicitly model class-conditional probability distributions, the true distributions are often not known. These are estimated from the data available, to approximate the true distributions. Errors in classification that arise due to this approximation can be reduced to some extent if the estimated distributions are used merely to project data into a space of like lihoods and classification is performed in that space using discriminant functions. In this article, we discuss the rationale behind this, and also the general properties of likelihood projections. We demonstrate the utility of likelihood projections in improving classification performance through experiments carried out on a standard image database and a standard speech database.

KEY WORDS: Discriminant-based classifiers; Distribution-based classifiers; Likelihood projections.

1. INTRODUCTION

Pattern classification methods can be broadly categorized into two groups, those that explicitly require class-conditional probability values of the data being classified and those that do not. The former category is sometimes referred to as the sampling approach, whereas the latter category is called the diagnostic paradigm (Dawid 1976; McLachlan 1992). Methods in the former category require explicit representations of the probability distributions of classes. These distributions are usually estimated either using nonparametric kernel methods, such as Parzen windows (Parzen 1962), or parametric methods that assume specific parametric forms for the distributions, such as Gaussian mixtures (McLachlan and Peel 2000). Class-conditional probabilities are used to estimate a posteriori class probabilities, which form the basis for classification (Duda, Hart, and Stork 2000). In this article we refer to these methods as *distribution-based* methods. The latter category of methods (i.e., methods that do not require explicit computation of class-conditional probability values), typically compute functions, called *discriminant functions*, of the data being classified and classify the data based on the values taken by these functions. The functions used may be diverse, ranging from simple linear functions of the data (Highleyman 1962) to complex structures such as classification and regression trees (Breiman, Friedman, Olshen, and Stone 1984) and need bear no direct relation to the a posteriori probabilities of the classes. We refer to such methods as *discriminant-based* methods in this article.

The dichotomy between the two categories of methods is, however, not complete. Methods that use explicit representations of class probability distributions are effectively based on discriminant functions. For instance, the classification rule of a distribution-based two-class classifier is based on the comparison of the ratio of the a posteriori probabilities of the classes against a threshold. In this case, this ratio is the discriminant function. Multiclass classification can be expressed similarly as the successive application of a series of such two-class discriminants. In this article, however, we maintain the categorization of classification methods as we have described them, because it imparts conceptual clarity to the subject matter of the article.

Distribution-based classifiers are widely used for classification tasks in diverse disciplines and are particularly useful in classifying real-valued data (Brown and Prescott 2000; Durbin, Eddy, Krogh, and Mitchison 1999; Mantegna and Stanley 2000; Wilks 1995). However, the performance of these classifiers is dependent on obtaining good estimates of the class-conditional distributions of the various classes. Although it is relatively easy to determine the best set of parameters for a given parametric model of distributions, determining the most appropriate parametric form is frequently a difficult problem. Inaccurate models can lead to reduced classification accuracies.

This article demonstrates how the performance of distribution-based classifiers can be improved under this scenario, by classifying in a different space into which the data are projected. In the rest of the article we refer to the space in which the original data reside as the *data space*. Instead of treating class-conditional probability distributions as facilitators for the estimation of a posteriori class probabilities to be used for Bayesian minimum error or minimum risk classification, we now treat them as facilitators for nonlinear projections, which we call *likelihood projections*, into a *likelihood space*. The coordinates of this space are the class-conditional likelihoods of the data for the various classes. In this space, the Bayesian classifier between any two classes in the data space can be viewed as a simple linear discriminant of unit slope with respect to the axes representing the two classes. The key advantage to be derived from working in the likelihood space is that we are no longer restricted to considering only this linear discriminant. Classification can now be based on any suitable classifier that operates on the projected data. When the projecting distributions are the true distributions of the classes, the optimal classifier in the likelihood space is guaranteed to result in error rates identical to those obtained by classifying the data in the original space. When the projecting distributions are not the true distributions, the optimal classification accuracy in the likelihood space is still guaranteed to be no worse than that obtainable with the projecting distributions in the data space. On the other hand, classification accuracy in the likelihood space can be higher

TECHNOMETRICS, ???? 0, VOL. 0, NO. 0
DOI 10.1198/00

than that in the data space in this situation. This feature of likelihood projections permits us to use them to compensate, to some extent, for errors in the modeling of class distributions in the original data space.

The use of secondary projections of data for improved classification has been widely considered in the field of kernel-based classification methods (Burges 1998; Cortes and Vapnik 1995; Schölkopf et al. 1999). Several density function have also been used as kernels in these methods (e.g., Schölkopf et al. 1997; Tresp 2001). Most of these methods, however, are specific to binary classification (Vapnik 1998), and although they can be restructured to perform multiclass classification (e.g., Lee, Lin, and Wahba 2001; Weston and Watkins 1998), their performance is frequently not as good as that obtainable with other multiclass classifiers. Although likelihood projections and likelihood spaces can be related to kernel methods, the treatment in this article is different in that it does not propose specific densities or projections to go with specific classifiers. The statement that we attempt to make in this article is that when a distribution-based classifier is the classifier of choice, then, rather than using it directly to classify in the data space, using the class-conditional distributions to project the data into its likelihood space and performing classification therein is a relatively better option. Furthermore, we do not impose any specific form on the classifiers to be used in the likelihood space. The approach proposed here is only a simple incremental step from distribution-based classification, but it can result in significant improvements in classification accuracy. The simplicity of the approach should make it appealing in any situation where distribution-based classification is to be performed for real-valued data. Many of the consequences or properties of likelihood projections are not immediately obvious. These have been discussed in greater detail in this article and may serve to throw some light on empirically observed results in various fields. For instance, researchers in the field of computer speech recognition have observed large improvements in recognition accuracy when classification of speech sounds is performed in the space of a posteriori class probabilities (Hermansky, Ellis, and Sharma 2000). These have largely been unexplained so far.

At the outset, we would like to point out that the concept of likelihood spaces is equally applicable to both discrete-valued and continuous-valued data. For this reason, we use the term "probability distribution," or simply "distribution," generically to represent both probability densities for the case of continuous valued data and probability distributions for discrete-valued data. Where the treatment is specific to continuous valued data, we use the term "probability density," or "density." In Section 2 we discuss likelihood projections and some key issues related to classification in likelihood spaces. In Section 3 we describe experiments that support our statements. Finally, we present our conclusions in Section 4.

2. LIKELIHOOD–BASED PROJECTIONS

Consider an N-class classification problem, where data must be classified as belonging to one of N classes $C_1, C_2, \ldots, C_N$. Let $P_{\mathbf{X}}(X|C_1), P_{\mathbf{X}}(X|C_2), \ldots, P_{\mathbf{X}}(X|C_N)$ represent the true distributions of the data from each of the N classes. In this notation the subscripted "$\mathbf{X}$" represents the random vector and the X within the parentheses represents a specific instance of the random vector $P_{\mathbf{X}}(X|C_i)$ thus represents the probability that the random vector $\mathbf{X}$ takes the value X, given that it belongs to class C_i. Let $\widetilde{P}_{\mathbf{X}}(X|C_1), \widetilde{P}_{\mathbf{X}}(X|C_2), \ldots, \widetilde{P}_{\mathbf{X}}(X|C_N)$ be the estimates of the true distributions that have been obtained for a distribution-based classifier. Such estimates could have been obtained by, for example, assuming a parametric form for the distributions and estimating their parameters from some training data using a likelihood maximization algorithm such as expectation maximization (Dempster, Laird, and Rubin 1977).

We define the *likelihood projection* of a vector X as the operation $L_N(X)$, resulting in an N-dimensional *likelihood vector*, $\mathbf{Y}_X$, as

$$\begin{aligned}\mathbf{Y}_X &= L_N(X)\\ &= [\,\log(\widetilde{P}_{\mathbf{X}}(X|C_1))\;\; \log(\widetilde{P}_{\mathbf{X}}(X|C_2))\;\; \cdots\;\; \log(\widetilde{P}_{\mathbf{X}}(X|C_N))\,].\end{aligned} \tag{1}$$

The ith component of the likelihood vector $\mathbf{Y}_X$, $Y_X^{(i)}$ is obtained as $Y_X^{(i)} = \log(\widetilde{P}_{\mathbf{X}}(X|C_i))$. We call the distributions $\widetilde{P}_{\mathbf{X}}(X|C_1), \widetilde{P}_{\mathbf{X}}(X|C_2), \ldots, \widetilde{P}_{\mathbf{X}}(X|C_N)$ the *projecting distributions*, and the N-dimensional space whose coordinates are $\log(\widetilde{P}_{\mathbf{X}}(X|C_1)), \log(\widetilde{P}_{\mathbf{X}}(X|C_2)), \ldots, \log(\widetilde{P}_{\mathbf{X}}(X|C_N))$ the *likelihood space*. $\mathbf{Y}_X$ has N components $Y_X^{(1)}, Y_X^{(2)}, \ldots, Y_X^{(N)}$, that is, as many components as the number of classes being classified. When the dimensionality of the data vector X is greater than N, the likelihood projection operation $L_N(X)$ is a dimensionality-reducing operation. When the dimensionality of X is greater than N, $L_N(X)$ is a dimensionality-increasing transformation.

2.1 Some Properties of Likelihood Projections

Likelihood vector representations have the following properties that relate to classification in likelihood spaces.

Property 1. Decision regions in the data space are compacted into contiguous regions in the likelihood space.

The projecting distributions represent a set of decision boundaries in the space of X that partition the data space into N decision regions, one for each class. Here, by the term "decision region" of a class, we mean the regions of the space that would be demarcated as belonging to that class by an optimal Bayesian classifier. Thus the decision region D_i for class C_i is the region defined by

$$X \in D_i \quad \text{if} \quad P(C_i)\widetilde{P}_{\mathbf{X}}(X|C_i) > P(C_j)\widetilde{P}_{\mathbf{X}}(X|C_j) \qquad \forall j \neq i, \tag{2}$$

where $P(C_i)$ represents the a priori probability of class C_i. The boundary regions where $P(C_i)\widetilde{P}_{\mathbf{X}}(X|C_i) = P(C_j)\widetilde{P}_{\mathbf{X}}(X|C_j)$ for some j are not attributed to any class by (2), and must be attributed to one of the competing classes based on some preset rule. The decision regions defined by (2) may consist of several disjoint regions or may be multiply connected. In the likelihood space, these (possibly disjoint or multiply connected) regions are projected into a region E_i, defined by

$$\mathbf{Y}_X \in E_i \quad \text{if} \quad Y_X^{(i)} + Z_i = Y_X^{(j)} + Z_j \qquad \forall j \neq i, \tag{3}$$

where $Z_i = \log(P(C_i))$. It is trivial to show that the region E_i is convex and therefore simply connected; from (3), we can

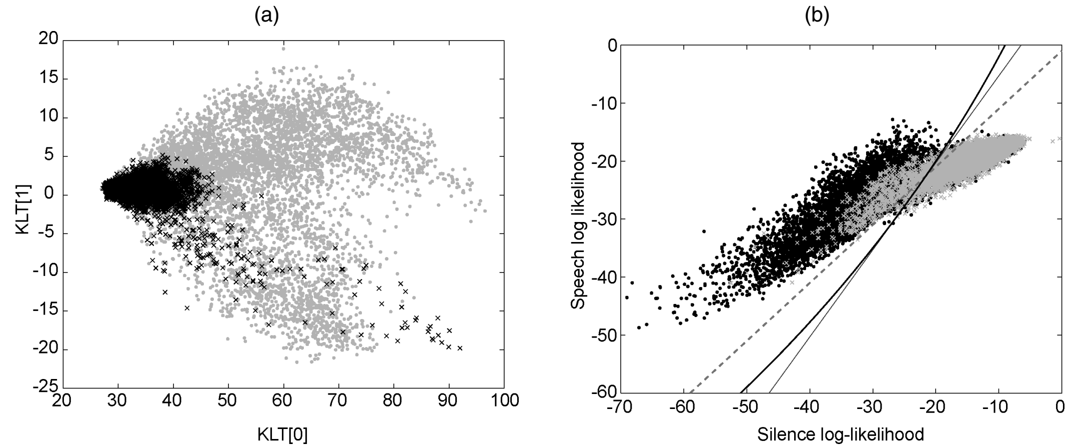

Figure 1. Scatter of Speech and Nonspeech Data in an Audio Signal. (a) The scatter in the data space; (b) the scatter in the likelihood space. The two axes represent the first and second components of vectors derived using a Karhunen Loeve transform (Jain 1976) based projection of the log spectra of 25-ms frames of the speech signal. In (a) the dark crosses represent data vectors from nonspeech regions. The gray dots represent data from speech regions. In (b) the colours are inverted for visual clarity. The projecting distributions for both classes were mixtures of 32 Gaussians, computed from speech and nonspeech training data. The dotted line in (b) represents the optimal classifier in the data space. The solid lines represent the optimal linear and quadratic discriminants in the likelihood space.

deduce that if Y_{X_1} and Y_{X_2} both lie within E_i, then, for any $0 \le \alpha \le 1$,

$$\alpha Y_{X_1}^{(i)} + (1-\alpha) Y_{X_2}^{(i)} + Z_i > \alpha Y_{X_1}^{(j)} + (1-\alpha) Y_{X_2}^{(j)} + Z_j \qquad \forall j \neq i; \tag{4}$$

that is, $\alpha Y_{X_1} + (1-\alpha) Y_{X_2}$ also lies in E_i, thereby proving that E_i is convex and therefore simply connected. Thus the likelihood projection transforms even disjoint or multiply connected decision regions in the data space to convex, simply connected regions in the likelihood space.

Figure 1 illustrates this property through an example wherein data vectors from two classes, in a recording of a parametrized speech signal, have been projected into a likelihood space using projecting distributions that were estimated from representative training data. The classes correspond to speech and nonspeech regions of the recorded signal. The figure shows the scatter of these classes in the original data space and the likelihood space. We observe that the result of the likelihood projection is to compact the classes, although the decision region for the speech class is not convex in Figure 1(a).

Property 2. The optimal classifier in the likelihood space is guaranteed to perform no worse than the optimal Bayesian classifier based on the projecting distributions.

This follows as a consequence of Property 1. In the data space, the optimal minimum-error Bayesian classifier is given by the rule (Duda et al. 2000)

$$X \in C_i: \qquad i = \arg\max_j \{P_{\mathbf{X}}(X|C_j)P(C_j)\}; \tag{5}$$

that is, X is classified as belonging to the class C_i, such that i indexes the class with the largest value for $P_{\mathbf{X}}(X|C_i)P(C_i)$. A classifier that uses the set of estimated distributions approximates this as

$$X \in C_i: \qquad i = \arg\max_j \{\widetilde{P}_{\mathbf{X}}(X|C_j)P(C_j)\}, \tag{6}$$

which can be equivalently stated in terms of log-likelihoods as

$$X \in C_i: \qquad i = \arg\max_j \bigl\{\log(\widetilde{P}_{\mathbf{X}}(X|C_j)) + \log(P(C_j))\bigr\}. \tag{7}$$

Equation (7) can be restated as a sequence of pairwise comparisons between classes. Classification between any two classes C_i and C_j is performed as

$$X \in \begin{cases} C_i & \text{if } \log(\widetilde{P}_{\mathbf{X}}(X|C_i)) - \log(\widetilde{P}_{\mathbf{X}}(X|C_j)) > T_{ij} \\ C_j & \text{otherwise,} \end{cases} \tag{8}$$

where T_{ij} is $\log(P(C_j)) - \log(P(C_i))$. Classification between N classes requires $N-1$ pairwise classifications of the kind defined by (8). The pairwise comparisons represented by (8) can be easily translated into the likelihood space. To do this, we define a vector $\mathbf{A}_{ij}$ as $\mathbf{A}_{ij} = [0\ 0\ 1\ 0\ \cdots\ -1\ 0\ \cdots]$, where the 1 occurs in the ith position and the -1 is in the jth position. Equation (8) can now be redefined in the likelihood space as

$$X \in \begin{cases} C_i & \text{if } \mathbf{A}_{ij}^T \mathbf{Y}_X > T_{ij} \\ C_j & \text{otherwise,} \end{cases} \tag{9}$$

where $\mathbf{Y}_X$ represents the likelihood projection of X. Equation (9) is a simple linear discriminant with a slope of unity. In the likelihood space, as in the data space, classification between N classes requires $N-1$ classifications of the kind defined by (9). It is thus possible to define a classifier in the likelihood space that performs identically to a Bayesian classifier based on the projecting distributions in the space of X. It follows that the performance of the optimal classifier in the likelihood space cannot be worse than that obtainable with the projecting distributions in the data space. It also follows that if the projecting distributions are the true distributions of the classes $P_{\mathbf{X}}(X|C_j)$, then the optimal classification performance in the likelihood space is identical to the optimal classification performance in the data space.

2.2 Classification in Likelihood Spaces

As a consequence of Property 2 in Section 2.1, the performance of the optimal classifier in the likelihood space is *lower* bounded by the classification accuracy obtainable with the optimal Bayesian classifier based on the projecting distributions in the data space. Therefore, it may actually be possible to estimate classifiers in the likelihood space that perform better than the optimal Bayesian classifier estimated from the projecting distributions. This constitutes the subject of discussion in this section.

In the data space the true distributions of the data may be extremely complex, and the distributions modeling the classes could result in complicated, possibly even multiple, disjoint estimated decision boundaries. Likelihood projections map the regions demarcated by these boundaries onto a single contiguous region in the likelihood space. A Bayesian classifier between any two classes in the data space maps onto a linear discriminant of slope 1.0 in the likelihood space. When projecting densities are continuous at the decision boundaries in the data space, data points that are misclassified in the data space, but lie adjacent to the decision boundaries, get mapped onto the region adjoining this linear discriminant in the likelihood space, regardless of the spatial complexity of the boundaries in the data space.

The geometrical simplicity of having misclassified regions adjoin the convex region representing any class in the likelihood space makes it possible to easily determine a different functional form for the discriminant that reduces the average classification error, compared with the linear discriminant of slope 1.0. Even simple classifiers, such as linear, quadratic, or logistic discriminants, that are effective only on contiguous classes, can be used. This is illustrated in Figure 1(b), where the dotted line represents the optimal Bayesian classifier estimated in the original data space. The slope of the line is 1.0. The $\mathbf{Y}$ intercept of the line was estimated using held-out test data. The thin solid line represents the optimal linear discriminant in the likelihood space, also estimated using the same held-out data. This discriminant results in 4.5% lower classification error relative to the dotted line. The solid curve represents a quadratic discriminant function, also estimated on the same held-out data, that results in even lower error than the thin solid line.

The determination of a new linear discriminant can be interpreted as corresponding to the determination of linear or nonlinear transformations of class distributions in the data space to achieve better approximation of optimal classification boundaries. For instance, a linear discriminant of slope 1.0 with a $\mathbf{Y}$ intercept other than that of the original linear discriminant corresponds to scaling of class distributions in the data space. A linear discriminant of slope other than 1.0 in the likelihood space corresponds to exponentiating the class densities by some power in the data space. The transformations of the densities result in a different set of decision boundaries than those obtained from the original class-conditional densities. The discriminants in the likelihood space can be construed to map onto these modified decision boundaries in the data space. Figure 2 illustrates this with an example. In this example, 120-dimensional log-spectral vectors, derived from a

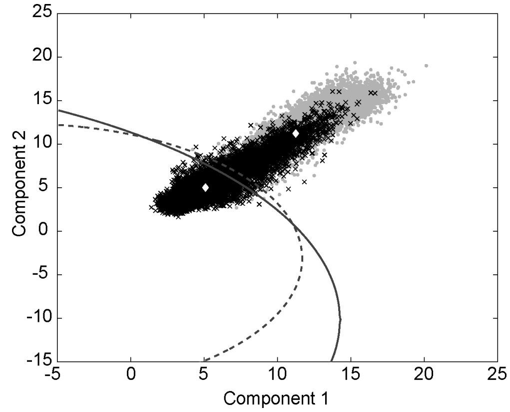

Figure 2. Illustration of Classification Boundaries Obtained From Original Class Distributions and From the Transformed Class Distributions Represented by Linear Discriminants of Nonunit Slope in Likelihood Space. The gray and black regions represent the scatter of data from two classes. The white spots in the centers of these classes represent the location of their means. The dotted curve represents the decision boundary obtained by modeling both classes as Gaussians. The solid curve represents the mapping of the optimal linear classifier in the likelihood space defined by the Gaussian class densities back into the data space.

speech signal as explained later in Section 3, have been projected into two dimensions. The probability density of each of the classes was modeled by a single Gaussian density. The dotted curve shows the classification boundary obtained from these Gaussian densities. The solid curve shows the decision boundary obtained by mapping the optimal linear discriminant separating the two classes in the corresponding likelihood space back into the data space. The reverse mapping of the linear discriminant is simple in this case. Let C_1 and C_2 represent the two classes, let $\widetilde{P}_{\mathbf{X}}(X|C_1)$ and $\widetilde{P}_{\mathbf{X}}(X|C_2)$ be their estimated Gaussian densities, and let $\mathbf{Y}_X$ represent the likelihood vector derived by projecting a vector X using these densities. We have

$$\mathbf{Y}_X = \left(Y_X^{(1)}, Y_X^{(2)}\right) = \left(\log(\widetilde{P}_{\mathbf{X}}(X|C_1)), \log(\widetilde{P}_{\mathbf{X}}(X|C_2))\right). \quad (10)$$

The optimal linear discriminant in the likelihood space can be represented as

$$AY_X^{(1)} + B = Y_X^{(2)}. \quad (11)$$

This can be represented in terms of the projecting densities as

$$\widetilde{P}_{\mathbf{X}}(X|C_1)^A e^B = \widetilde{P}_{\mathbf{X}}(X|C_2). \quad (12)$$

The new decision boundary is thus the locus of all vectors X that satisfy (12).

More generally, however, such simple interpretations are not possible. For instance, a quadratic discriminant of the form

$$\left(Y_X^{(1)}\right)^2 + D\left(Y_X^{(2)}\right)^2 + EY_X^{(1)}Y_X^{(2)} + F = 0 \quad (13)$$

maps onto the following discriminant in data space:

$$\widetilde{P}_{\mathbf{X}}(X|C_1)^{\log(\widetilde{P}_{\mathbf{X}}(X|C_1)) + E\log(\widetilde{P}_{\mathbf{X}}(X|C_2))} \times \widetilde{P}_{\mathbf{X}}(X|C_2)^{D\log(\widetilde{P}_{\mathbf{X}}(X|C_2))} e^F = 1. \quad (14)$$

Clearly, this cannot be obtained by any simple transformation of the individual class distributions, due to the presence of the cross-term $Y_X^{(1)} Y_X^{(2)}$. Other, more complex discriminants in likelihood space are mapped onto even more complex functions of class distributions in the data space.

2.3 Dependence of Classifiers in Likelihood Spaces on Projecting Distributions

The reduced classification error in the likelihood space is a consequence of compensation for errors in modeling class distributions in the data space. In the context of classification, distribution modeling errors can result from two causes. First, the analytical model chosen to represent the distribution of a dataset may be inappropriate for the data. Second, the parameters of the model for any class are usually estimated such that the resulting distribution best represents the distribution of the training data for that class, without reference to the distributions of other classes. Figure 3(a) illustrates the problems that can result from this using a synthetic example. In the example, the data are one-dimensional. Two classes with Rayleigh distributions have been erroneously modeled as Gaussian. The dotted curves in Figure 3(a) show the true probability densities of the two classes, and the solid curves show the estimated Gaussian densities. The first and second moments of the Gaussians shown in the figure are identical to those of the true (Rayleigh) distribution of the data; that is, they represent the maximum likelihood Gaussian estimates that would be obtained with unlimited training data from the two classes. The optimal decision boundary, B_{true}, is the value of the abscissa at the point where the true densities cross over. This is indicated by the vertical dotted line. The estimated decision boundary, $B_{\text{estimated}}$, occurs at the abscissa where the Gaussian estimates of the densities cross over and is indicated by the vertical solid line. The gray-shaded region represents data that will be misclassified due to the difference between B_{true} and $B_{\text{estimated}}$. This error is the direct result of erroneous modeling of Rayleigh distributions as Gaussian.

Figure 3(b) shows the two-dimensional likelihood projection of data from the two classes. We note that the curve represents a one-dimensional manifold in the two-dimensional likelihood space. This is expected because the projection is a deterministic dimensionality increasing transform (Conlon 1993). The estimated Bayesian classifier in the data space is represented by the solid line of slope 1.0. The star on the curve represents the optimal decision threshold, B_{true} in the data space. Thus the optimal classifier in the likelihood space can be any line or curve that passes through the point marked by the star, for example, the linear discriminant represented by the dotted diagonal line.

As explained in Section 2.2, classification with a linear determinant other than the solid line in the figure is equivalent to classification with a transformed version of the class distributions in the data space. For example, the optimal discriminant represented by the dotted line in Figure 3(b) is equivalent to classification with the scaled Gaussians shown in Figure 3(c). As a result of the scaling, the Gaussians now cross over at the optimal classification boundary.

The optimal classification boundary may also be obtained by modeling the classes with a different set of Gaussians in the first place, by discriminatively training them to optimize classification. Several methods for such discriminative training of class distributions have been proposed in the literature (e.g., Normandin, Cardin, and De Mori, 1994). Figure 3(c) also shows an example of such discriminative Gaussian estimates for the Rayleigh class distributions of Figure 3(a). They too cross over at the optimal classification boundary. The principle of classification in likelihood spaces remains valid, however. Even when class distributions are discriminatively trained, the performance of the optimal classifier in the likelihood space derived from these distributions is only lower bounded by that of the Bayesian classifier based on the distributions in the data space. Also, regardless of the manner in which class distributions are trained, the form of the classification boundaries in the data space is constrained by the model chosen for the distributions. For instance, if class distributions are modeled as Gaussian, then the resultant Bayesian classifier is constrained to be a quadratic discriminant. On the other hand, the data-space discriminants corresponding to a discriminant in like-

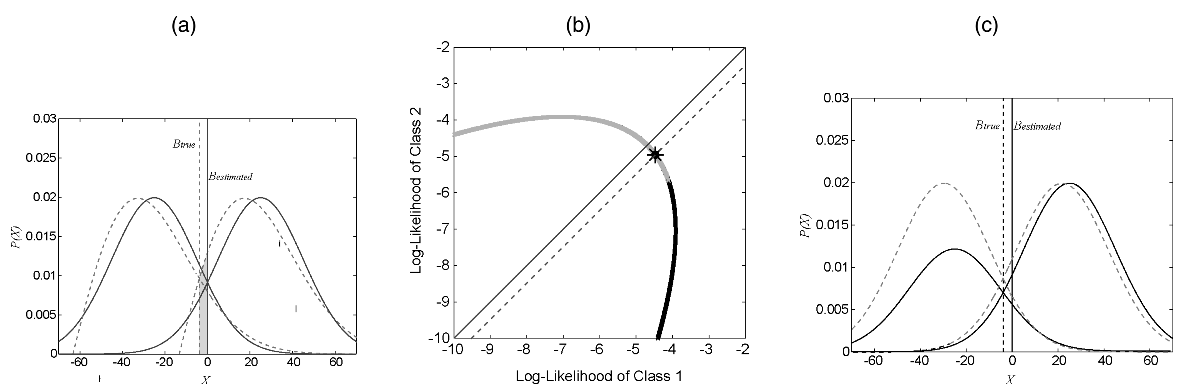

Figure 3. Synthetic Two-Class Example Illustrating Why It May Be Possible to Obtain Improved Classification Performance in Likelihood Spaces. (a) The true densities of the classes are Rayleigh (shown by the dotted curves) but are inaccurately modeled as Gaussians (solid curves). The gray region between the true decision boundary B_{true} and the estimated decision boundary $B_{estimated}$ represents data that will be misclassified. (b) The scatter of likelihood space representations of the data from the two classes. The gray and black portions of the figure represent data from the two classes. The solid line represents $B_{estimated}$, and the star represents the true optimal decision threshold B_{true}. The dotted line passing through the star represents an optimal linear discriminant. (c) The dark solid curves represent the scaled versions of the Gaussians in (a) that are implicit in the optimal (dotted) linear discriminant in (b). They intersect at the optimal classification boundary. The lighter dotted curves represent examples of discriminatively estimated Gaussian distributions for the classes. They too intersect at the optimal classification boundary.

lihood space can be significantly more complex than those obtainable with the Bayesian classifier in data space. For example, when class distributions are Gaussian, even a simple quadratic discriminant in the likelihood space with no cross terms corresponds to a fourth-order polynomial discriminant in the data space. It is therefore plausible that a superior classifier may be obtained in the likelihood space even when class distributions are discriminatively trained.

It must be clear from the discussion thus far that when classifiers in the likelihood space are simple linear or quadratic discriminants, improved classification in the likelihood space is largely a consequence of compensating for classification errors in regions adjoining the classification boundaries in the data space. Such discriminants cannot be expected to compensate for classification errors that occur for other reasons. Such errors, for example, can occur when the distributions modeling the classes in the original space miss entire regions of the optimal decision regions (given by the true class distributions) altogether.

Classifiers that are more complex than simple linear or quadratic discriminants may also be defined in the likelihood space. For instance, one may define distribution-based classifiers within the likelihood space. Such classifiers may result in better classification than linear or quadratic discriminants. In general, however, as the decision boundaries in the data space approach the optimal boundaries, the gains to be expected from classifying in likelihood spaces quickly diminish. Also, in this situation the decision boundaries in the data space onto which the optimal discriminant in the likelihood space maps approach the decision boundaries given by the class densities themselves.

It must be recognized that we are guaranteed only that the best classifier in the likelihood space performs at least as well as the best Bayesian classifier in the data space based on the projecting distributions. There is no guarantee that it performs at least as well as the best classifier of any kind in the data space. In fact, there is no assurance that the best possible classifier in the likelihood space can perform comparably to the best possible classifier in the data space, unless the likelihood projection is invertible.

2.4 Localization of Data Vectors by Their Likelihood Projections

The likelihood projection would be invertible if it could be guaranteed that no more than a single data vector projects onto any likelihood vector. But likelihood projections are generally not invertible, as shown in Figure 4, and the likelihood projection of a data vector cannot be guaranteed to uniquely identify the data vector. Nevertheless, we do note that as the number of class distributions in the likelihood projection increases, the likelihood projection of a vector increasingly *localizes* it in the data space. Consider a likelihood vector $\mathbf{Y}_X$ with components $Y_X^{(1)}, Y_X^{(2)}, \ldots, Y_X^{(N)}$, that has been obtained by the projection of a vector X. Let U_X^i represent the region in the data space such that

$$\exp\bigl(Y_X^{(i)}\bigr) \le \widetilde{P}_{\mathbf{X}}(X : X \in U_X^i | C_i) \le \exp\bigl(Y_X^{(i)}\bigr) + \varepsilon, \tag{15}$$

where ε is an infinitesimally small number. By this definition, U_X^i is the set of all data vectors that have a class-conditional probability for C_i lying in the interval $[\exp(Y_X^{(i)}), \exp(Y_X^{(i)}) + \varepsilon]$. The size of U_X^i is the volume of the data space that lies within it. Knowledge of $Y_X^{(i)}$ localizes X to lie in U_X^i. Further, knowledge of $Y_X^{(i)}$ and $Y_X^{(j)}$ localizes X to the region $U_X^i \cap U_X^j$. Thus, knowing the first j components of the likelihood vector localizes X to lie in the region V_X^j defined by

$$V_X^j = \bigcap_{i=1}^{j} U_X^i. \tag{16}$$

It is easy to see that

$$V_X^1 \supseteq V_X^2 \supseteq \cdots \supseteq V_X^N; \tag{17}$$

that is, V_X^j is a decreasing series. Knowledge of the likelihood vector $\mathbf{Y}_X$ is equivalent to knowing that X lies within V_X^N; that is, $\mathbf{Y}_X$ contains the *positional* information that X lies in V_X^N. We note that V_X^N is guaranteed not to be larger than the smallest U_X^i, whereas it can be much smaller. We also note that V_X^N may be empty for many likelihood vectors and is nonempty only if the likelihood vector has been generated from any data vector. Conversely, for any likelihood vector $\mathbf{Y}_X$ that has been

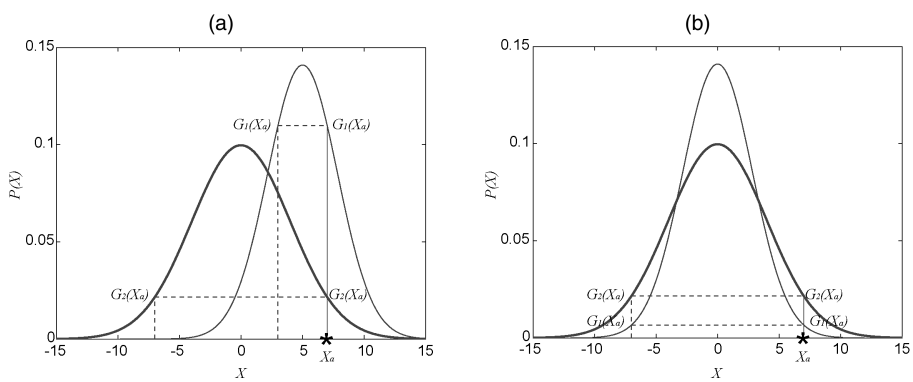

Figure 4. Illustration of the Invertibility of Likelihoods. The two Gaussians transform the data point X_a into the pair of density values $G_1(X_a)$ and $G_2(X_a)$. In (a), the two Gaussians have different means. The two vertical dotted lines show the other values of X that result in density values $G_1(X_a)$ and $G_2(X_a)$. It is clear that there is only one point that results in both $G_1(X_a)$ and $G_2(X_a)$. Thus $G_1(X_a)$ and $G_2(X_a)$ uniquely identify X_a. In (b), the means of the two Gaussians are identical. In this case there is a second value of X that has density values $G_1(X_a)$ and $G_2(X_a)$, and the mapping from X values to density values cannot be inverted.

generated thorough the projection of a data vector X, V_X^N cannot be empty and must contain at least one data point, namely X itself.

3. EXPERIMENTS

In the discussion so far, we have only discussed the *existence* of classifiers in the likelihood space that can classify no worse than a Bayesian classifier in the data space. But the mere existence of such classifiers provides no assurance that they can in fact be estimated, or that the actual classification performance of the classifiers estimated in likelihood space will always be superior to that of the Bayesian classifier. Estimation of classifiers is always difficult, and the final performance of any estimated classifier is governed by many factors, including the estimation procedure used, size of the training data, and so on. We hypothesize that since decision regions of the Bayesian classifier are mapped onto convex regions of the likelihood space, it would be simpler to estimate better classifiers in the likelihood space. This hypothesis must be experimentally substantiated, and we do so in this section with experiments on the Brodatz texture database (Brodatz 1966) and the TIMIT speech database (Zue, Seneff, and Glass 1990).

3.1 Classification of Visual Textures

Visual textures are images characterized by some degree of homogeneity that typically contain repeated structures, often with some random variation. Thus images of the surface of water, fabrics, cloudy skies, and even wallpaper are all considered textures. In 1966 a photographer named Phil Brodatz published a set of 112 textures, including pictures of walls, matted surfaces, and so on, in a book titled *Textures: A Photographic Album for Artists and Designers*. The "Brodatz texture database" has been derived by extracting subimages from 8-bit 512×512 pixel digitization of these images (e.g., Picard, Kabir, and Liu 1993). Nine nonoverlapping 128×128 pixel subimages have been extracted from each of the textures, resulting in a set of 1,008 images. Figure 5 shows a few examples of Brodatz's textures.

We evaluated classification in likelihood spaces on this database. For our experiments, the 9 subimages for each texture were separated into a training set of 8 images and 1 test image, resulting in an overall training set of 996 images and a test set of 112 images. The partitioning into train and test sets was done in 9 different ways in a jackknife procedure, effectively increasing the test set size to 1,008 images. The aim of all experiments was to identify the textures from which the test images were drawn.

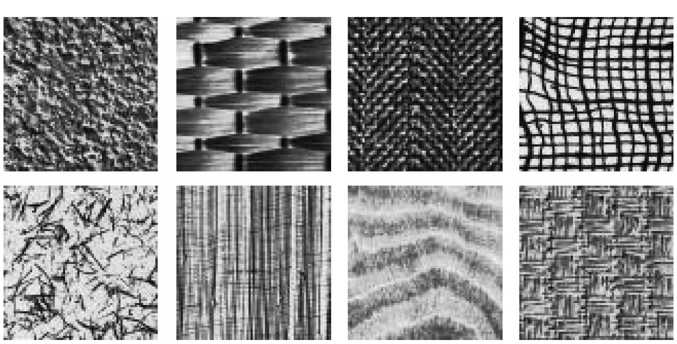

Figure 5. Examples of Brodatz Textures.

For the experiments, each 128×128 pixel image was parameterized into a set of 4,096 64-dimensional vectors as follows. The image was segmented into squares of 8×8 pixels, where adjacent squares overlapped by 6 pixels. The edges of the image were padded with zero-valued pixels such that every pixel in the image was included in exactly 16 squares. A 64-component discrete cosine transform (DCT) was computed for each square (Vasconcelos and Carneiro 2002). The 64-dimensional DCT vectors for any image were assumed to be independent and identically distributed. The distributions of the DCT vectors for the textures were assumed to be mixtures of Gaussians with diagonal covariance matrices. The number of Gaussian components in the mixtures represented a parameter that controlled the degree to which the estimated density fit the data. Mixtures with 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1,024, and 4,096 Gaussians were trained from the 32,768 vectors derived from the 996 training images for each texture. In each experiment, the distributions for all textures had the same number of Gaussian components. The a priori probabilities of all textures were assumed to be identical.

Classification in the data space was performed using the joint log-likelihood of all 4,096 feature vectors obtained from the test image. Joint log-likelihoods of the classes were also used to project test images into likelihood space. Although the number of classes (112) is greater than the number of components in the feature vector (64), the projection into likelihood space nevertheless constituted a dimensionality-reducing transform, because the entire set of 4,096 64-dimensional vectors for each image was projected onto a single 112-dimensional likelihood vector. For classification in likelihood space, linear discriminants were trained to classify between each pair of classes using a least squares procedure (Duda et al. 2000). Because there were 8 training images from each texture, only 16 likelihood vectors were available to train any linear discriminant. A total of 6,216 linear discriminants were trained. Classification was performed using the voting mechanism based on exhaustive pairwise classification suggested by Friedman (1996), where pairwise classification was performed between all pairs of classes. The class that was selected most frequently by the pairwise classifiers was chosen to be the output of the multi-class classifier.

Figure 6 shows the combined results from the 9 jackknife experiments. The dotted line shows the classification accuracies obtained in the data space as a function of the number of Gaussian components in the class distributions, and the solid line shows classification accuracies obtained in the corresponding likelihood space. In almost all cases, the classification accuracy obtained in the likelihood space is higher than that in the data space. In the data space, the best classification result is obtained with mixtures of 128 Gaussians. In the likelihood space, the best classification accuracy is obtained when the projecting densities are mixtures of 64 Gaussians. For mixtures of more than 16 Gaussians and fewer than 512 Gaussians, however, the difference between the performance obtained in the data and likelihood spaces is statistically insignificant as measured using McNemar's test (Siegel 1956). On the other hand, the differences between the two at the extremes of the curves in the figure are significant to the .05 level or better.

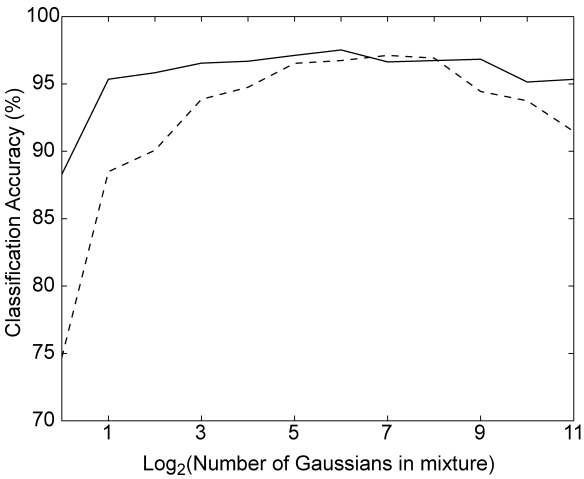

Figure 6. Classification Accuracy on Brodatz Textures in the Space of the Original 64-Dimensional DCT Vectors and in the Likelihood Space. The X-axis represents the log of the number of Gaussians in the mixture Gaussian distributions used to model class distributions in the data space. The dotted line represents classification accuracy obtained by a Bayesian classifier in the data space, and the solid line represents classification accuracy in the corresponding likelihood space.

These results are as expected from our discussion in Section 2. When the projecting class distributions are suboptimal for classification, large gains are obtained by classifying in likelihood space. The gains diminish as the projecting distributions become more optimal. In some cases, classification in likelihood space is in fact less accurate than that in the data space. This is not unexpected either, because only 16 training examples were available to estimate each pairwise classifier in the likelihood space, and hence the estimated classifiers did not generalize to the test data better than the Bayesian classifier in data space. In general, however, the classification performance in likelihood space is observed to be much more robust to variations in the class distributions than the data-space Bayesian classifier based on those distributions.

Although Figure 6 demonstrates only the robustness of classification in likelihood space to the number of Gaussians in the mixture Gaussian class distributions, it was also found to be robust to variations in the estimates of the distributions themselves. In our experiments, the expectation maximization (EM) algorithm used to train the mixture Gaussians was observed to be rather sensitive to the initial settings for the parameters, especially for mixtures with 512 or more components. To estimate the distributions reliably, we estimated each mixture Gaussian density several times, by restarting the EM algorithm with different initial values. The results in Figure 6 were obtained with distributions that resulted in the highest likelihood for the training data.

Table 1 gives classification accuracies obtained with two different sets of mixture Gaussian densities on one of the nine train/test partitions of the Brodatz textures. Because the test set here consisted of only 112 images, the table reports the actual number of images correctly classified, rather than the percentage of accuracy. The mixture densities in the first set, labelled "Gaussian mixture 1" in the table, were poorly trained and resulted in poor classification in the data space. The second set of densities, labelled "Gaussian mixture 2," were well trained and resulted in significantly better classification than the first set. In both cases, better classification was achieved in the likelihood space. More importantly, the classification performance in likelihood space was almost identical for both sets of projecting distributions, the difference being statistically insignificant. Classification in likelihood space thus appears to compensate for the poor generalizability of the distributions in Gaussian mixture 1.

Table 1. Number of Brodatz Textures Correctly Classified Using Mixture Gaussian Densities With 512, 1,024, and 2,048 Mixture Components

Number of Gaussians in mixture		*512*	*1,024*	*2,048*
Gaussian mixture 1	Baseline classification	91	70	54
	Classification in likelihood space	104	106	102
Gaussian mixture 2	Baseline classification	101	100	98
	Classification in likelihood space	103	103	103

NOTE: The test set has 112 texture images in all. The first two rows, labelled as Gaussian mixture 1, show classification results obtained with poorly trained mixture Gaussian densities that do not generalize well to the test data. The third and fourth rows, labelled as Gaussian mixture 2, show classification results obtained with well-trained densities that generalize well to the test data.

3.2 Classification of Speech Sounds

We conducted experiments using the TIMIT speech database (Zue et al. 1990) provided by the Linguistic Data Consortium (LDC). TIMIT is a standard database used by speech researchers for the development of signal processing and classification algorithms. The TIMIT corpus consists of 5.38 hours of individually recorded spoken utterances, of which 3.94 hours have been designated as training data and 1.44 hours as test data. In this corpus, the sounds in American English have been categorized into 61 phonemes (or sound units) by linguistic experts. Phoneme boundaries have been manually marked and provided with signals. The classes considered in our experiments were obtained by grouping the 61 phonemes into 10 sets, as listed in Table 2. Note that although the names given to the sets are coincident with those provided with the TIMIT corpus, the composition is not the one specified in the corpus. The names here are simply indicative of broad phonetic characteristics of the elements of the sets.

For our experiments, each speech signal in the TIMIT corpus was first transformed into a sequence of feature vectors. For this, the signal was divided into segments, or *frames*, of 20 ms, where adjacent frames overlapped by 10 ms. Thus each second of speech yielded 100 frames. From each frame, a 40-dimensional Mel-frequency log-spectral vector was derived (Davis and Mermelstein 1980). Each vector was further

Table 2. Listing of Phoneme Groupings to Generate Classes

Set name	*Phoneme composition*
Affricates	/jh/ /ch/
Back	/ih/ /eh/ /ae/ /aa/ /ah/ /ao/
Closures	/kcl/ /tcl/ /pcl/ /gcl/ /pcl/ /bcl/ /pau/ /epi/ /h#/
Diphthongs	/iy/ /ey/ /ay/ /aw/ /oy/
Fricatives	/s/ /sh/ /z/ /zh/ /f/ /th/ /v/ /dh/
Nasals	/m/ /n/ /ng/ /em/ /en/ /eng/ /nx/
Round	/ow/ /uw/ /ux/ /uh/
Schwa	/ix/ /ax/ /axr/ /ax-h/ /er/
Semivowels	/l/ /r/ /w/ /y/ /hh/ /hv/ /el/
Stops	/b/ /d/ /g/ /p/ /t/ /k/ /dx/ /q/

NOTE: Each entity enclosed in "/ · /" represents a phoneme.

augmented by a 40-dimensional difference vector, computed as the difference of the log-spectral vectors of the succeeding and preceding frame, and a 40-dimensional double difference vector, computed as the difference between the difference vectors of the succeeding and preceding frame. The final vector representing any frame of speech was thus 120-dimensional. Note that Mel-frequency log-spectral representations derived in this manner, or their linear transformations, have been empirically determined to be highly effective for classifying speech (Davis and Mermelstein 1980). There were 142,910 phonetic segments composed of approximately 1.42 million vectors available for training the 10 classes and 51,681 phonetic segments composed of approximately .5 million vectors in the test set.

The goal of the experiments was to classify each phonetic *segment* in the test data into one of the 10 classes (and not merely to classify individual frames). The joint evidence of all the frames in a segment was used to classify it. For the purpose of this experiment, log-spectral vectors within any segment were assumed to be independent and identically distributed. The probability distribution of the log-spectral vectors belonging to each sound class was modeled by a mixture of Gaussians. Mixture Gaussian distributions are widely used to model the distributions of Mel-frequency log-spectra and their linear derivatives for the purpose of classifying speech sounds (Huang, Acero, and Hon 2001). Mixtures with 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1,024, 2,048, 4,096, and 8,192 Gaussian components were computed for each of the classes, using the EM algorithm. All Gaussians were assumed to have diagonal covariance matrices.

Classification in the data space was performed using the joint log-likelihood of all frames within a segment. The normalized joint log-likelihoods of the classes were also used to project speech segments into likelihood space. The normalization was performed by dividing the joint log-likelihood of the frames in a segment by the number of frames in the segment. This was necessary, because different segments have different numbers of frames. Each segment was thus represented by a single vector in likelihood space.

3.2.1 Discriminant-Based Classifiers in Likelihood Space. To perform discriminant-based classification, linear discriminants were trained to classify between each pair of classes using a least squares procedure (Duda et al. 2000). A total of 45 linear discriminants were trained. Classification was performed using the voting mechanism based on exhaustive pairwise classification as suggested by Friedman (1996).

Figure 7(a) shows classification error rates obtained on 120-dimensional log-spectral feature vectors in the data and likelihood spaces. Classification in likelihood space is observed to be superior to classification in data space in all cases. The difference between the two is particularly large when the number of Gaussians in the projecting class distributions is either very small or very large. The best performance is obtained with mixtures of 1,024 Gaussians. Even here, classification in likelihood space is significantly superior to classification in data space.

Although the best performance is obtained with 1,024 component Gaussian mixture class densities, the fact that classification performance is better in the likelihood space, even with simple linear discriminants, indicates that the estimated Gaussian mixtures do not optimally model class densities in the 120-dimensional space. We therefore projected the 120-dimensional vectors down into a 9-dimensional subspace using linear discriminant analysis (Duda et al. 2000). Linear discriminant analysis identifies subspaces within which the classes are most separated, and the lower dimensionality of the space makes it simpler to estimate class distributions. Gaussian mixture class distributions were trained for the 9-dimensional vectors and used both for Bayesian classification and for projection of segments into likelihood space.

Figure 7(b) shows classification performance on the 9-dimensional vectors. We observe that classification performance on the 9-dimensional data is superior to that on the 120-dimensional data when the projecting Gaussian mixtures have a small number of Gaussian components. The best performance is obtained with mixtures of 128 Gaussians. Again, classification in likelihood space is consistently superior to classification in the data space. Surprisingly, as the number of

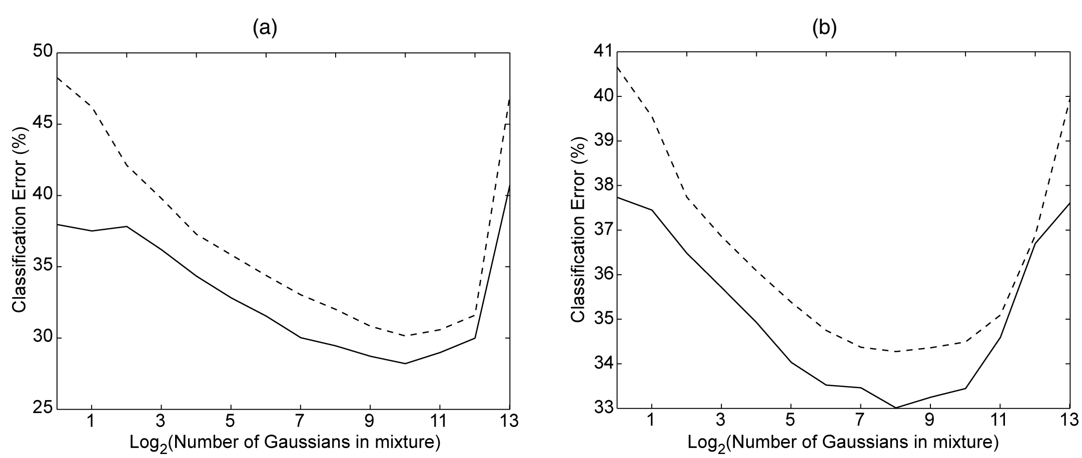

Figure 7. Classification Error on the TIMIT Data. (a) Classification error in the space of 120-dimensional log-spectral vectors, and the likelihood spaces derived from it. (b) Classification error in the space of 9-dimensional projections of log-spectral vectors, and the likelihood spaces derived from it. In both panels the X-axis represents the log of the number of Gaussians in the mixture Gaussian distributions used to model class distributions in the data space. The dotted lines represent classification error rates obtained by a Bayesian classifier in the data space, and the solid lines represent classification error rates in the corresponding likelihood space.

Gaussians in the class distributions increases, classification in the 120-dimensional space is superior to classification in the 9-dimensional space. The best segment level classification performance is obtained for the 120-dimensional features, with mixtures of 1,024 Gaussians. Although we do not speculate on the reason for these results, we point out that the lowest overall classification error (28.2%) is obtained with *likelihood projections* of the 120-dimensional features.

3.2.2 Distribution-Based Classifiers in Likelihood Space. A major distinction between distribution-based and discriminant-based classifiers lies in the fact that whereas class distributions in distribution-based classifiers can be trained independently of one another, discriminant-based classifiers are discriminatively trained; that is, they are trained to explicitly optimize some measure of the expected classification error and thus must consider all classes. Thus, whereas the class distributions for the classifiers in our experiments were not discriminatively trained, the classifiers in the likelihood space were discriminatively trained, and thereby optimized for classification.

A question that arises naturally is whether the observed improved classification in likelihood space is simply a consequence of the discriminative training of the classifiers in likelihood space, or whether the projection into likelihood space makes it simpler to estimate good classifiers. To investigate this, we evaluated the performance of distribution-based classifiers in the likelihood space. The experiments were conducted on likelihood projections of the 120-dimensional log-spectral feature vectors. In a preliminary diagonalization step, the likelihood vectors were rotated by multiplication with the matrix of eigenvectors of the overall covariance of the training set. Mixture Gaussian class distributions were trained from the (rotated) likelihood vectors for each of the classes. In every experiment, all class distributions had an identical number of Gaussians. Likelihood-space distributions were not discriminatively trained. As a result, there was no discriminatively trained component in the classifier.

The results of the experiment are given in Table 3. The first row presents classification errors obtained in the likelihood space, when the projecting class densities were single Gaussians. The second row presents classification errors when the projecting densities were the ones that resulted in the best classification in the data space, that is, mixtures of 1,024 Gaussians. The first two columns of both rows of the table present the classification errors obtained in the data space and with linear discriminants in the likelihood space. Subsequent columns provide classification errors obtained with distribution-based classifiers in the likelihood space.

We observe that classification with distribution-based classifiers in the likelihood space did in fact improve significantly on classification in the data space itself. In fact, when projecting class distributions (in the data space) were single Gaussians, the best distribution-based classifier was observed to outperform the discriminant-based classifier. Even when the projecting class distributions were mixtures of 1,024 Gaussians, the best distribution-based classifier in the likelihood space performed significantly better than classification in data space, although the discriminant-based classifier was better still. This suggests that there is inherent merit to the mapping performed by likelihood projections themselves that enables us to improve on the classification performance obtained in the data space. In a follow-up experiment, it was determined that classification in a second likelihood space, obtained using the class densities in the likelihood space as a projecting distributions, did not result in additional improvements; that is, there is no advantage to recursively projecting data into newer likelihood spaces.

4. DISCUSSION AND CONCLUSIONS

As is evident from the experiments reported in Section 3, classification is likelihood space is very robust to errors in the modeling and estimation of class distributions in the data space. Variations of classification performance with changes in class distributions are much smaller in the likelihood space than in the data space. The advantages to be derived from this fact are clear. It may often be simpler to estimate a relatively crude set of class distributions and perform the final classification in the likelihood space than to search for the optimal set of class distributions. In many situations, the computational requirements of the classifier are important. The combined computational requirement of a likelihood projection using simple models for class distributions, followed by a simple classifier in likelihood space, may be significantly lower than that of a more complicated classifier in data space, while providing the same performance.

For the most part, in our experiments we have restricted the explored classifiers to linear discriminants, because our goal was only to demonstrate that better classification is possible in likelihood spaces, not to obtain the best classifier for the data considered. One advantage of linear discriminants is that the optimal Bayesian classifier in the data space is also a linear discriminant in the likelihood space. Thus any search for an optimal linear discriminant in the likelihood space will also consider this classifier. This ensures that the classifier in the likelihood space does not perform worse than the one in the data space, at least on the training data. However, better classification performance may be possible through the use of other discriminant functions, such as quadratic discriminants (Gupta, Riley, and White 1986) or logistic regressors (Darlington 1990).

Table 3. Percent Classification Errors Obtained With Distribution-Based Classifiers in Likelihood Spaces

	Baseline	*Discriminant*	*1 Gaussian*	*2 Gaussians*	*4 Gaussians*	*8 Gaussians*	*16 Gaussians*	*32 Gaussians*	*64 Gaussians*
1 Gaussian	48.2	38.0	41.3	40.8	39.1	37.7	36.9	35.7	35.5
1,024 Gaussians	30.1	28.2	33.7	31.2	30.0	29.1	29.1	28.8	29.1

NOTE: The two rows show classification errors obtained when projecting class distributions are single Gaussians, and mixtures of 1,024 Gaussians. The first column shows the baseline Bayesian classification error in the data space. The second column shows the percent error obtained with a discriminant-based classifier in the likelihood space. The remaining columns show errors obtained with distribution-based classifiers in the likelihood space. The numbers in the heading rows of the columns indicate the number of Gaussian components in the mixture Gaussian likelihood-space class distributions.

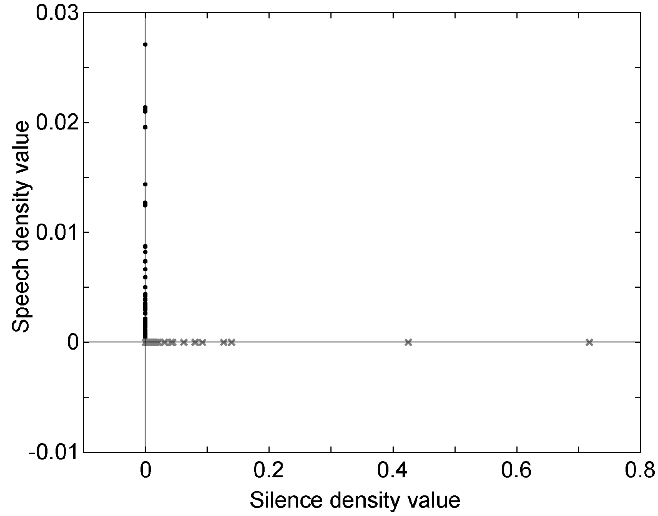

Figure 8. Scatter of Density Values of Data Shown in Figure 1, Measured Using the Densities of Two Classes. This must be compared with the scatter of log-likelihood values shown in Figure 1.

Also, discriminant-based multiclass classification has been performed by the combination of binary classifiers using the voting mechanism of Friedman (1996). Several other methods have been proposed, including the use of cyclic redundancy codes (Dietterich and Bakiri 1995) or pairwise coupling (Hastie and Tibshirani 1998), which might result in better performance.

In this article we have considered only log-likelihoods as projections. However, much of the discussion herein would also apply if we were to use the logarithm of estimated a posteriori class probabilities as projections. This is because likelihoods and a posteriori class probabilities are related—the former are just a scaled version of the latter. As mentioned in Section 1, a posteriori probability-based projections have been used earlier in speech recognition systems and have been found to result in greatly improved recognition performance, as compared with recognition using the data vectors (Hermansky et al. 2000).

The logarithm that we have used in the likelihood projections is an important component of these projections. Most data points have very low likelihoods for at least one of the classes. Consequently, any density-based projection that does not incorporate the logarithm projects most of the data points into regions that are very close to one of the axes, making it difficult to obtain simple discriminants for the data. The logarithm function tends to expand this region out, simplifying the problem. Figure 8 illustrates this pictorially. Other functions with similar properties could have also been used instead of the logarithm.

Although distribution-based classifiers in the likelihood space are effective, they may be difficult to estimate when the number of classes in the likelihood projection, and thereby the dimensionality of the likelihood space, is greater than the dimensionality of the data space. In such situations, data vectors are projected onto manifolds of the same dimensionality as the data space within the likelihood space (Conlon 1993). Figure 3 shows such an example, where one-dimensional data are projected onto a one-dimensional manifold in two-dimensional space. In such situations, the likelihood space is largely empty. This makes using continuous densities difficult, because they would also attempt to account for data in the empty regions of the space. To avoid this problem, it may be advantageous to unwrap the manifold into a lower-dimensional Euclidean space using such methods as charting (Brand 2002), before classification. This hypothesis remains to be evaluated.

Finally, we note from the TIMIT experiments in Section 3.2.2 that for segment-level classification, distribution-based classifiers in the likelihood space derived from the 120-dimensional log-spectral vectors are far more effective than distribution-based classifiers in the 9-dimensional space derived by linear discriminant analysis of the 120-dimensional space. Both linear discriminant analysis and likelihood projections project the 120-dimensional data into a lower-dimensional space; however, the likelihood projection has the added advantage of gathering class data from potentially disconnected regions into convex regions. It is not clear whether the superior performance obtained with likelihood projections is due entirely to this reason, or whether this result will hold up on other data. Further experiments are needed to resolve this question.

ACKNOWLEDGMENT

The authors thank Dr. Paris Smaragdis for discussions and help with the classification experiments. Rita Singh was sponsored by Space and Naval Warfare Systems Center grant N66001-99-1-8905. The content of the information in this publication does not necessarily reflect the position or the policy of the U.S. Government, and no official endorsement should be inferred.

[Received July 2002. Revised June 2003.]

REFERENCES

Brand, M. (2002), "Charting a Manifold," *Proceedings on Neural Information Processing Systems*, Vancouver, BC, Canada.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984), *Classification and Regression Trees*, Belmont, CA: Wadsworth.

Brodatz, P. (1966), *Textures: A Photographic Album for Artists and Designers*, New York: Dover.

Brown, H., and Prescott, R. (2000), *Applied Mixed Models in Medicine*, London: Wiley.

Burges, C. J. C. (1998), "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, 2, 1–43.

Conlon, L. (1993), *Differentiable Manifolds: A First Course*, Cambridge, MA: Birkhäuser.

Cortes, C., and Vapnik, V. (1995), "Support Vector Networks," *Machine Learning*, 20, 273–297.

Darlington, R. B. (1990), *Regression and Linear Models*, New York: McGraw-Hill.

Dawid, A. P. (1976), "Properties of Diagnostic Data Distributions," *Biometrics*, 32, 647–658.

Davis, S. B., and Mermelstein, P. (1980), "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28, 357–366.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.

Dietterich, T., and Bakiri, G. (1995), "Solving Multiclass Learning Problems via Error-Correcting Output Codes," *Journal of Artificial Intelligence Research*, 2, 263–286.

Duda, R. O., Hart, P. E., and Stork, D. G. (2000), *Pattern Classification* (2nd ed.), New York: Wiley.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1999), *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge, U.K.: Cambridge University Press.

Friedman, J. H. (1996), "Another Approach to Polychotomous Classification," technical report, Stanford University, Dept. of Statistics.

Gupta P. L., Riley J. T., and White T. J. (1986), "Misclassification Probabilities for Quadratic Discrimination," *SIAM Journal of Scientific and Statistical Computing*, 7, 1400–1417.

Hastie, T., and Tibshirani, R. (1998), "Classification by Pairwise Coupling," *The Annals of Statistics*, 26, 451–471.

Hermansky, H., Ellis, D. P. W., and Sharma, S. (2000), "Tandem Connectionist Feature Extraction for Conventional HMM Systems," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP2000*, Istanbul, Turkey, pp. 1635–1638.

Highleyman, W. H. (1962), "Linear Decision Functions With Applications to Pattern Recognition," *Proceedings of the IRE*, 50, 1501–1514.

Huang, X. D., Acero, A., and Hon, H. (2001), *Spoken Language Processing*, Englewood Cliffs, NJ: Prentice-Hall.

Jain, A. K. (1976), "A Fast Karhunen–Loeve Transform for a Class of Random Processes," *IEEE Transactions on Communications*, 24, 1023–1029.

Lee, Y., Lin, Y., and Wahba, G. (2001), "Multicategory Support Vector Machines," Technical Report TR-1043, University of Wisconsin, Dept. of Statistics.

Mantengna, R. N., and Stanley, H. E. (2000), *An Introduction to Econophysics: Correlations and Complexity in Finance*, Cambridge, U.K.: Cambridge University Press.

McLachlan, G. J. (1992), *Discriminant Analysis and Statistical Pattern Recognition*, New York: Wiley.

McLachlan, G. J., and Peel, D. (2000), *Finite Mixture Models*, New York: Wiley.

Normandin, Y., Cardin, R., and De Mori, R. (1994), "High-Performance Connected Digit Recognition Using Maximum Mutual Information Estimation," *IEEE Transactions on Speech and Audio Processing*, 26, 299–311.

Parzen, E. (1962), "On the Estimation of a Probability Density Function and Mode," *Annals of Mathematical Statistics*, 33, 1065–1076.

Picard, R. W., Kabir, T., and Liu, F. (1993), "Real-Time Recognition With the Entire Brodatz Texture Database," *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, New York, NY, pp. 638–639.

Schölkopf, B., Mika, S., Burges, C., Knirsch, P., Müller, L.-R., Ratsch, G., and Smola, A. J. (1999), "Input Space vs. Feature Space in Kernel-Based Methods," *IEEE Transactions on Neural Networks*, 10, 1000–1017.

Schölkopf, B., Sung, K. K., Burges, C. C., Girosi, F., Niyogi, P., Poggio, T., and Vapnik, V. (1997), "Comparing Support Vector Machines With Gaussian Kernels to Radial Basis Function Classifiers," *IEEE Transactions On Signal Processing*, 45, 2758–2765.

Siegel, S. (1956), *Nonparametric Statistics for the Behavioral Sciences*, New York: McGraw-Hill.

Tresp, V. (2001), "Mixtures of Gaussian Processes," in *Advances in Neural Information and Processsing Systems 13*, (eds. T. K. Leen, T. G. Dietteriech, and V. Tresp,) Cambridge, MA: MIT Press.

Vapnik, V. (1998), *Statistical Learning Theory*, New York: Wiley.

Vasconcelos, N., and Carneiro, G. (2002), "What Is the Role of Independence for Visual Recognition?" in *Proceedings of the European Conference on Computer Vision*, Copenhagen, Denmark.

Weston, J., and Watkins, K. (1998), "Multiclass Support Vector Machines," Technical Report CSD-TR-98-04, University of London.

Wilks, D. S. (1995), *Statistical Methods in the Atmospheric Sciences: An Introduction*, London: Academic Press.

Zue, V., Seneff, S., and Glass, J. (1990), "Speech Database Development at MIT: TIMIT and Beyond," *Speech Communication*, 9, 351–356.