Feature Extraction for Nonparametric Discriminant Analysis

Mu ZHU and Trevor J. HASTIE

In high-dimensional classification problems, one is often interested in finding a few important discriminant directions in order to reduce the dimensionality. Fisher's linear discriminant analysis (LDA) is a commonly used method. Although LDA is guaranteed to find the best directions when each class has a Gaussian density with a common covariance matrix, it can fail if the class densities are more general. Using a likelihood-basedinterpretation of Fisher's LDA criterion, we develop a general method for finding important discriminant directions without assuming the class densities belong to any particular parametric family. We also show that our method can be easily integrated with projection pursuit density estimation to produce a powerful procedure for (reduced-rank) nonparametric discriminant analysis.

Key Words: Classification; Density estimation; Dimension reduction; LDA; Projection pursuit; Reduced-rank model; SAVE.

1. INTRODUCTION

In a typical classification problem, sometimes known as pattern recognition, one has a training set $\{y_i, \mathbf{x}_i\}_{i=1}^n$, where $y_i \in \{1, 2, \ldots, K\}$ is the class label and $\mathbf{x}_i \in \mathbb{R}^d$, a vector of predictors. When d is large, it is often the case that information relevant to the separation of the classes is contained in just a few directions $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_M \in \mathbb{R}^d$, where M is much smaller than d. Our goal is to identify these important directions from the data. These directions are often known as *discriminant directions* or the important *linear features* for classification. In order to do so, one must answer a key question: what is an appropriate measure of class separation?

Fisher's linear discriminant analysis (LDA) (Fisher 1936) works with one such measure;

Mu Zhu is Assistant Professor, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, N2L 3G1, Canada, (E-mail: m3zhu@uwaterloo.ca). Trevor J. Hastie is Professor, Department of Statistics, Stanford University, Stanford, CA 94305 (E-mail: hastie@stat.stanford.edu).

Journal of Computational and Graphical Statistics, Volume 12, Number 1, Pages 101–120
DOI: 10.1198/1061860031220

the best feature is defined as

$$\underset{\alpha \in \mathbb{R}^d}{\text{argmax}} \frac{\boldsymbol{\alpha}^T \mathbf{B} \boldsymbol{\alpha}}{\boldsymbol{\alpha}^T \mathbf{W} \boldsymbol{\alpha}},$$

where $\mathbf{B}$ is the between-class covariance matrix and $\mathbf{W}$, the within-class covariance matrix. The optimal solution is the first eigenvector of $\mathbf{W}^{-1}\mathbf{B}$. In general, the matrix $\mathbf{W}^{-1}\mathbf{B}$ has $\min(K-1, d)$ nonzero eigenvalues. Hence we can identify up to this number of features—with decreasing importance. In fact, given the first $(m-1)$ discriminant directions, the mth direction is simply

$$\underset{\alpha \in \mathbb{R}^d}{\text{argmax}} \frac{\boldsymbol{\alpha}^T \mathbf{B} \boldsymbol{\alpha}}{\boldsymbol{\alpha}^T \mathbf{W} \boldsymbol{\alpha}} \quad \text{subject to} \quad \boldsymbol{\alpha}^T \mathbf{W} \boldsymbol{\alpha}_j = 0 \quad \forall \quad j < m.$$

The importance of these discriminant directions is significant. For example, classification using only the leading discriminant directions (e.g., reduced-rank LDA) can often improve classification performances on test data. Leading discriminant directions also allow us to make low-dimensional summary plots of the data. Zhu (2001) showed that discriminant directions are equivalent to the concept of *ordination axis* in correspondence analysis, which has a wide range of applications beyond pattern recognition, for example, in areas such as environmental ecology, information retrieval, and personalization.

However, LDA is not always guaranteed to find the best discriminant directions. Figure 1 shows a pathological case for LDA. The top panel shows the first two coordinates of a simulated dataset in $\mathbb{R}^{20}$ (the "multi-modality data"). Class 1 is simulated from a standard multivariate Gaussian, while classes 2 and 3 are mixtures of two symmetrically shifted standard Gaussians. In the remaining 18 coordinates, Gaussian noise is added for all three classes. The middle panel shows a two-dimensional projection produced by LDA. LDA clearly fails to recover the important features for classification. The reason why LDA fails in this case is that the class centroids coincide. This points out a fundamental restriction of LDA as a feature extraction method. This article develops a more general method for finding important discriminant directions and subspaces. We also show that our method can be easily integrated with projection pursuit density estimation to produce a powerful procedure for (reduced-rank) nonparametric discriminant analysis.

2. LITERATURE REVIEW

The pathology of LDA as illustrated above has long been recognized. Devijver and Kittler (1982, sec. 9.8) specifically pointed out that there are two types of discriminatory information: one where all such information is contained in the class centroids (the ideal case for LDA), and one where all such information is contained in the class variances (the pathological case above). A specific solution (details omitted here) is then proposed to deal with the second type, with a comment from the author that a direct feature extraction method "capable of extracting as much discriminatory information as possible regardless of its type" would require a "complex criterion" which would be "difficult to define" (Devijver and Kittler 1982, p. 339).

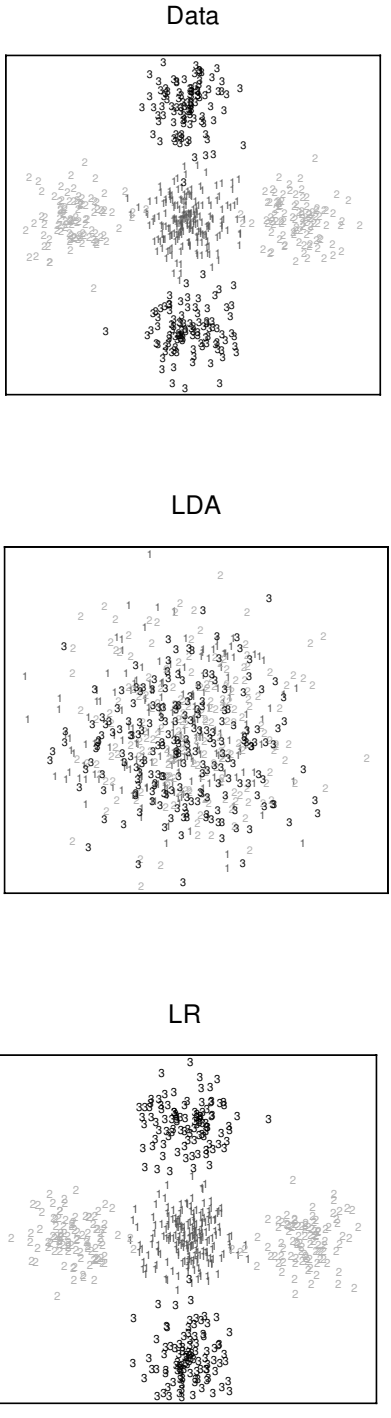

Figure 1. Top: Data are 20-dimensional. Only the first two coordinates are shown. The remaining 18 coordinates are simulated from the same *standard Gaussian distribution for all three classes. Middle: Two-dimensional discriminant subspace identified by LDA. Bottom: Two-dimensional discriminant subspace identified by recursively maximizing LR($\boldsymbol{\alpha}$), using feature removal.*

Recently, Cook and Yin (2001) proposed a new method using the *sliced average variance estimator* (SAVE), which, the authors showed, is capable of extracting both types of discriminatory information simultaneously. In Section 4, we compare our method with SAVE to gain more insight into the nature of these different methods.

3. GENERALIZED FEATURE EXTRACTION

3.1 THE GENERALIZED CRITERION

To better understand Fisher's criterion, let us consider it from a likelihood point of view. Suppose given $y = k$, the predictor vector $\mathbf{x}$ from class k has density function $p_k(\mathbf{x})$. Consider

$$\begin{aligned} \mathrm{H}_0 : &\quad p_k = p \quad \text{for all} \quad k = 1, 2, \ldots, K. \\ \mathrm{H}_A : &\quad p_k \neq p \quad \text{for some} \quad k = 1, 2, \ldots, K. \end{aligned}$$

In this framework, a natural candidate for measuring class differences in a fixed direction $\boldsymbol{\alpha}$ is the (marginal) generalized log-likelihood-ratio statistic:

$$\mathrm{LR}(\boldsymbol{\alpha}) = \log \frac{\max_{p_k} \prod_{k=1}^{K} \prod_{\mathbf{x}_j \in C_k} p_k^{(\alpha)}(\boldsymbol{\alpha}^T \mathbf{x}_j)}{\max_{p_k = p} \prod_{k=1}^{K} \prod_{\mathbf{x}_j \in C_k} p^{(\alpha)}(\boldsymbol{\alpha}^T \mathbf{x}_j)}, \tag{3.1}$$

where $p_k^{(\alpha)}(\cdot)$ is the marginal density along the projection defined by $\boldsymbol{\alpha}$ for class k; $p^{(\alpha)}(\cdot)$ is the corresponding marginal density under the null hypothesis that the classes share the same density function; and the notation "$x_j \in C_k$" means the jth observation belongs to class k. Then, it can be shown via straight-forward algebraic calculations (Zhu 2001) that the criterion used in LDA is a special case of $\mathrm{LR}(\boldsymbol{\alpha})$.

Result 1. *If $p_k(\mathbf{x})$ is (or is estimated by) the $N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ density function for classes $k = 1, 2, \ldots, K$, then choosing discriminant directions by maximizing LR($\boldsymbol{\alpha}$) is equivalent to Fisher's LDA.*

This simple result leads to two observations. First, it shows that when each class has a Gaussian density with a common covariance matrix (a situation where all the discriminatory information is contained in the class centroids), Fisher's linear discriminant directions are optimal. Second, it suggests a natural way to generalize Fisher's linear discriminant direction when the class densities are more general (cases where the important discriminatory information is not simply contained in the class centroids alone)—for example, when they have different covariance matrices, or, more generally, when they are non-Gaussian.

More specifically, for arbitrary class density functions, our goal is to seek $\boldsymbol{\alpha}$ that maximizes

$$\mathrm{LR}(\boldsymbol{\alpha}) = \sum_{k=1}^{K} \sum_{\mathbf{x}_j \in C_k} \log \hat{p}_k^{(\alpha)}(\boldsymbol{\alpha}^T \mathbf{x}_j) - \sum_{k=1}^{K} \sum_{\mathbf{x}_j \in C_k} \log \hat{p}^{(\alpha)}(\boldsymbol{\alpha}^T \mathbf{x}_j), \tag{3.2}$$

where $\hat{p}_k$ denotes the MLE of p_k and $\hat{p}$, the MLE of p.

Note that if $\boldsymbol{\alpha}' = c\boldsymbol{\alpha}$ for some multiple c, then LR($\boldsymbol{\alpha}'$) = LR($\boldsymbol{\alpha}$). Hence it suffices to constrain $\boldsymbol{\alpha}$ to be a unit vector, that is, $\|\boldsymbol{\alpha}\| = 1$. This means the effective search space for our maximization problem above is the unit ball in $\mathbb{R}^d$, not the entire space $\mathbb{R}^d$. This is also true for LDA except that, conventionally, LDA uses the restriction $\boldsymbol{\alpha}^T \mathbf{W} \boldsymbol{\alpha} = 1$ instead of $\|\boldsymbol{\alpha}\| = 1$, but aside from a normalizing constant, these are equivalent.

3.2 THE OPTIMIZATION PROBLEM IN DETAIL

It is important to note that the statistic LR($\boldsymbol{\alpha}$) is defined generally. We are free to restrict p_k to any desirable family of density functions. Often, one would like to work with a parametric family. Then for fixed $\boldsymbol{\alpha}$, $\hat{p}_k^{(\alpha)}$ can be evaluated quite simply by maximum likelihood. Here, however, we show that the criterion LR($\boldsymbol{\alpha}$) is general enough so that even if one chooses to work with exible but more difficult nonparametric models, it is still possible to use LR($\boldsymbol{\alpha}$) to guide the search of informative directions for classification, provided that one is willing to accept the extra computational cost.

For the optimization problem posed above, standard methods such as Newton's method are readily applicable in theory, because both the gradient and the Hessian can be estimated explicitly. To simplify the notation, we write f_k in place of $\hat{p}_k^{(\alpha)}$ and f in place of $\hat{p}^{(\alpha)}$. Then

$$g(r) \stackrel{\text{def}}{=} \frac{\partial \mathrm{LR}}{\partial \boldsymbol{\alpha}_r} = \sum_{k=1}^{K} \sum_{\mathbf{x}_j \in C_k} x_{rj} \left(\frac{f_k'(\boldsymbol{\alpha}^T \mathbf{x}_j)}{f_k(\boldsymbol{\alpha}^T \mathbf{x}_j)} - \frac{f'(\boldsymbol{\alpha}^T \mathbf{x}_j)}{f(\boldsymbol{\alpha}^T \mathbf{x}_j)} \right) \tag{3.3}$$

and, by writing $z_j = \boldsymbol{\alpha}^T \mathbf{x}_j$,

$$\begin{aligned} H(r,s) &\stackrel{\text{def}}{=} \frac{\partial^2 \mathrm{LR}}{\partial \boldsymbol{\alpha}_r \partial \boldsymbol{\alpha}_s} \\ &= \sum_{k=1}^{K} \sum_{\mathbf{x}_j \in C_k} x_{rj} \left(\frac{f_k''(z_j) f_k(z_j) - (f_k'(z_j))^2}{(f_k(z_j))^2} - \frac{f''(z_j) f(z_j) - (f'(z_j))^2}{(f(z_j))^2} \right) x_{sj}. \end{aligned} \tag{3.4}$$

Therefore, to estimate g and H, we need only estimate *univariate* marginal density functions and their first two derivatives. Various methods are available, for example, the kernel method (e.g., Silverman 1986) or the local likelihood method (e.g., Loader 1999). We do this once conditionally on each class to obtain f_k, f_k', f_k'', and once unconditionally over the entire training sample to obtain f, f', f''. Although it involves estimating the derivatives too, we shall refer to this operation simply as the "density estimation step."

In practice, one seldom uses Newton's method in its vanilla form. Some modification to the Hessian matrix is necessary, for example, when the problem is nonconvex, to ensure that one is moving in an ascent direction with each iteration. A popular variation, known as quasi-Newton, is to construct Hessian-like matrices from the gradient that are always negative-definite. More details can be found in Gill, Murray, and Wright (1981).

3.3 The Density Estimation Step

In our implementation, we used C. Loader's Locfit library (Loader 1999), which is based on local likelihood theory. To focus on the main statistical ideas, we choose not to digress and go into the theory of local likelihood estimation here. Most density estimation methods, including local likelihood, have an equivalent kernel representation and, therefore, can be viewed as a special kernel estimator, at least on the conceptual level. The choice of Locfit as our density estimation module, however, is largely pragmatic. Locfit implements a rather elaborate set of evaluation structures and interpolation strategies—the density function is only evaluated at a few strategically selected points, and its values at other points are obtained by interpolation, thereby significantly reducing the amount of computation. For more details, we refer the reader to Loader (1999, sec. 12.2).

3.4 Local Maxima

The existence of local maxima adds an extra difficulty to numerical optimization. This difficulty here can be effectively mitigated by choosing relatively large bandwidths or equivalent smoothing parameters in the density estimation step. Figure 2 shows the function $\mathrm{LR}(\boldsymbol{\alpha})$ for a two-dimensional problem simulated for the purposes of illustration as $\boldsymbol{\alpha} = (\cos\theta, \sin\theta)$ goes around the unit circle, that is, for $\theta \in (0, 2\pi)$—recall that it suffices to restrict $\boldsymbol{\alpha}$ to be a unit vector. The function is estimated using nonparametric density models with different bandwidth parameters. With a large bandwidth, the local maximum essentially disappears. It is also important to point out that extremely accurate estimation of the first two derivatives of the univariate density is not necessary, as long as it can be ensured that the Newton iterates are generally moving in the correct direction. Alternatively, we are currently investigating the use of some stochastic search algorithms which are theoretically guaranteed to converge to the global maximum even for relatively rough objective functions. For smooth objective functions, of course, the convergence occurs much faster with Newton type of algorithms.

3.5 Multiple Features

In this section, we brie y address the problem of finding multiple discriminant directions. We present two of our favored strategies.

3.5.1 Orthogonalization

Suppose $\{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_{m-1}\}$ are the first $m-1$ discriminant directions, then

$$\boldsymbol{\alpha}_m = \text{argmax } \mathrm{LR}(\boldsymbol{\alpha}) \quad \text{subject to} \quad \boldsymbol{\alpha}^T \Phi \boldsymbol{\alpha}_j = 0 \quad \forall \quad j < m,$$

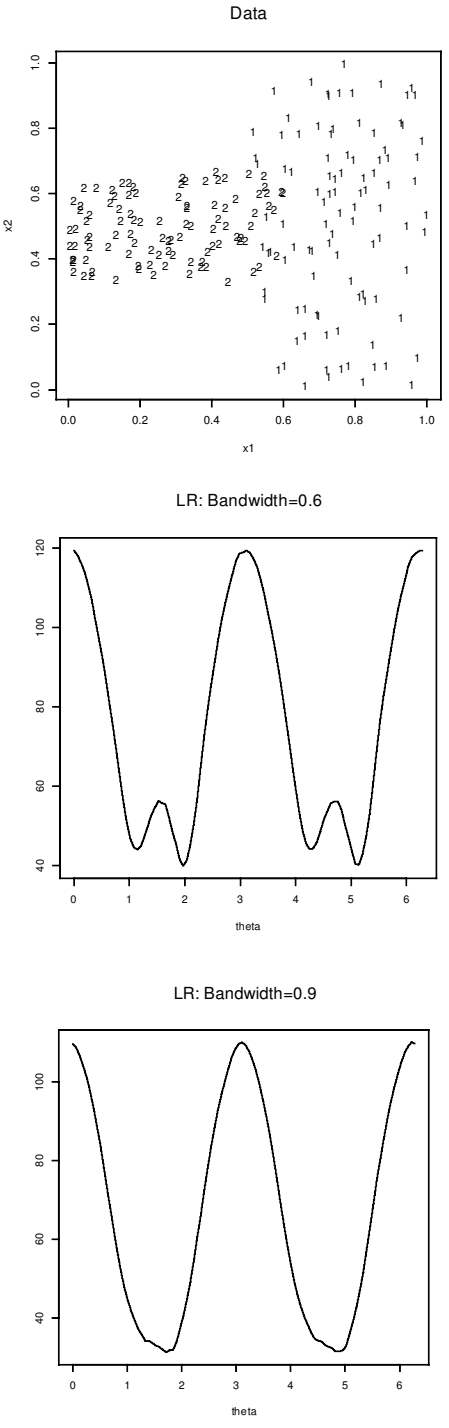

Figure 2. LR computed on simulated data as depicted in the top panel, using nonparametric density estimates with different bandwidths. A low bandwidth leads to a rough estimate with a local maximum while a large bandwidth smoothes away the local maximum.

following standard practices in classical multivariate statistics. Here "$\mathbf{x}^T\mathbf{\Phi}\mathbf{y} = 0$" means $\mathbf{x}$ and $\mathbf{y}$ are orthogonal with respect to the metric $\mathbf{\Phi}$. The main advantage of this strategy is its simplicity. The main disadvantage is the difficulty in justifying the choice of $\mathbf{\Phi}$. In classical multivariate statistics, one often assumes that the data follow a multi-dimensional Gaussian distribution, $\mathrm{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$, so there is a natural coordinate system (i.e., $\mathbf{\Phi} = \mathbf{\Sigma}$) in which it is interesting and meaningful to focus on features that are orthogonal to one another. The orthogonality $\boldsymbol{\alpha}^T\mathbf{\Sigma}\boldsymbol{\alpha}_j = 0$ implies that $\mathbf{U} = \mathbf{X}\boldsymbol{\alpha}$ and likewise $\mathbf{U}_j$ are uncorrelated. For non-Gaussian data, no such natural metric exists. In fact, for data separated into K classes, even if we assume that the data in each class are Gaussian, it is still not clear what the appropriate metric is, unless one assumes, as in LDA, that all the classes share a *common* covariance matrix. In practice, one often spheres the data prior to the analysis, that is, let $\mathbf{x}^* = \mathbf{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$, where $\boldsymbol{\mu}$ is the overall mean and $\mathbf{\Sigma}$, the total covariance matrix of the data. This is the same as using the total covariance matrix as an ad-hoc metric for orthogonalization. Another reasonable choice of $\mathbf{\Sigma}$ is the within-class covariance matrix, which, when p_k is Gaussian, is the same as LDA. In general, there is no reason why the total or the within-class covariance matrix is the appropriate metric. Before we present a more general strategy below, however, we feel it is important to add that despite it being somewhat ad-hoc, orthogonal features are still useful for some applications, such as data visualization.

3.5.2 Feature Removal

Alternatively, we can adopt Friedman's exploratory projection pursuit paradigm (Friedman 1987): Once a discriminant direction $\boldsymbol{\alpha}$ is found, we simply transform the data so that there is no class difference in the $\boldsymbol{\alpha}$ direction—in the sense that $p_k^{(\alpha)} = q$ for all k with some common q—while keeping all other directions unchanged, and search for the next direction.

In particular, let $z = \boldsymbol{\alpha}^T\mathbf{x}$, then for each class k, a transformation is applied to z:

$$z' = \mathbf{Q}^{-1}(F_k(z)) \stackrel{\text{def}}{=} \gamma(z), \tag{3.5}$$

where $\mathbf{Q}(\cdot)$ is the cumulative distribution function (cdf) corresponding to the common density function q and F_k, the marginal cdf of z for class k. By definition, $\gamma(\cdot)$ is a monotonic one-to-one transformation and the transformed variable, z', has density q. Let $\mathbf{A}$ be an orthogonal rotation matrix such that

$$\mathbf{z} \stackrel{\text{def}}{=} \mathbf{A}\mathbf{x} = \begin{pmatrix} \boldsymbol{\alpha}^T\mathbf{x} \\ \mathbf{A}^*\mathbf{x} \end{pmatrix}. \tag{3.6}$$

Such $\mathbf{A}$ can be constructed using the Gram–Schmidt procedure (see, e.g., Friedberg, Insel, and Spence 1989, p. 307). Then the entire transformation process can be summarized in the following diagram:

$$\begin{array}{ccc} \mathbf{x} & \stackrel{h}{\longrightarrow} & h(\mathbf{x}) \\ \downarrow \mathbf{A} & & \uparrow \mathbf{A}^{-1} \\ \mathbf{z} & \stackrel{t}{\longrightarrow} & t(\mathbf{z}) \end{array}, \tag{3.7}$$

where

$$t(z_j) = \begin{cases} \gamma(z_j) & \text{for} \quad j = 1, \\ z_j & \text{for} \quad j > 1. \end{cases} \tag{3.8}$$

Hence

$$h(\mathbf{x}) = \mathbf{A}^{-1} t(\mathbf{A}\mathbf{x}).$$

The feature removal strategy is more general because the directions are not constrained to be orthogonal, nor is there an issue of choosing the appropriate orthogonalization metric. We also have the following result, which adds some theoretical coherence to this strategy.

Result 2. *If $p_k(\mathbf{x})$ is (or is estimated by) the $N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ density function for classes $k = 1, 2, \ldots, K$, then recursively maximizing LR($\boldsymbol{\alpha}$) using exploratory projection pursuit yields the same discriminant directions as Fisher's LDA.*

The proof (Zhu 2001) of this result relies on the fact that when p_k is $\mathrm{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, the transformation γ as defined above in (3.5) is simply a location shift. The bottom panel of Figure 1 shows the "multi-modality data" projected onto the first two discriminant directions found by recursively maximizing $\mathrm{LR}(\boldsymbol{\alpha})$ using the feature removal strategy with a bandwidth parameter of 0.6 in the Locfit module. As we can clearly see, our generalized procedure successfully identifies the two discriminant directions.

3.6 EXAMPLE: 1984 CONGRESSIONAL VOTES DATA

The UCI machine-learning repository contains a dataset that describes how each member of the U.S. House of Representatives voted on 16 key issues in 1984. The dataset and any relevant information are at ftp://ftp.ics.uci.edu/pub/machine-learning-databases/voting-records/.

There are 435 Congressmen recorded in the database. Sometimes a Congressman may have been absent when the voting took place, or he may have simply voted "present"—instead of a clear "yes" or "no"—to avoid con icts of interest. These cases are represented as missing values in the database. In our analysis, each vote is characterized by two indicator variables, one indicating whether the vote is a "yes" and the other, indicating whether a clear yes-or-no vote is missing, resulting in a total of 32 predictors.

The top panel of Figure 3 shows the two-dimensional LDA subspace. Apparently, it is easy to distinguish the majority of the Democrats from the Republicans using the voting records. The bottom panel of Figure 3 shows the two-dimensional discriminant subspace found by maximizing $\mathrm{LR}(\boldsymbol{\alpha})$, using the orthogonalization strategy. Here we choose $\Phi = \mathbf{W}$ to be the within-class covariance matrix, the same as LDA. We see the same basic separation between Democrats and Republicans as above. But we also see some additional features. It seems that the voting patterns among Democrats are more polarized, or have a higher variance, than the Republicans. Moreover, within the Democratic camp, there appears to be several separate subgroups, or within-group clusters. There seem to be more "outlying" members among the Democrats, and, within the "mainstream" Democrats, there seem to be

two major subdivisions. LDA will *never* find such elaborate features. In fact, for this two-class problem, LDA can find only one meaningful discriminant direction. In this case, it is clear that there is more information in the data, and that LDA is insufficient. The additional information is obviously important. It would have been extremely interesting to examine the two subgroups within the mainstream Democrats, for example. Various interest groups could have used such information to their strategic advantage—not even politicians can undermine the power of effective data visualization!

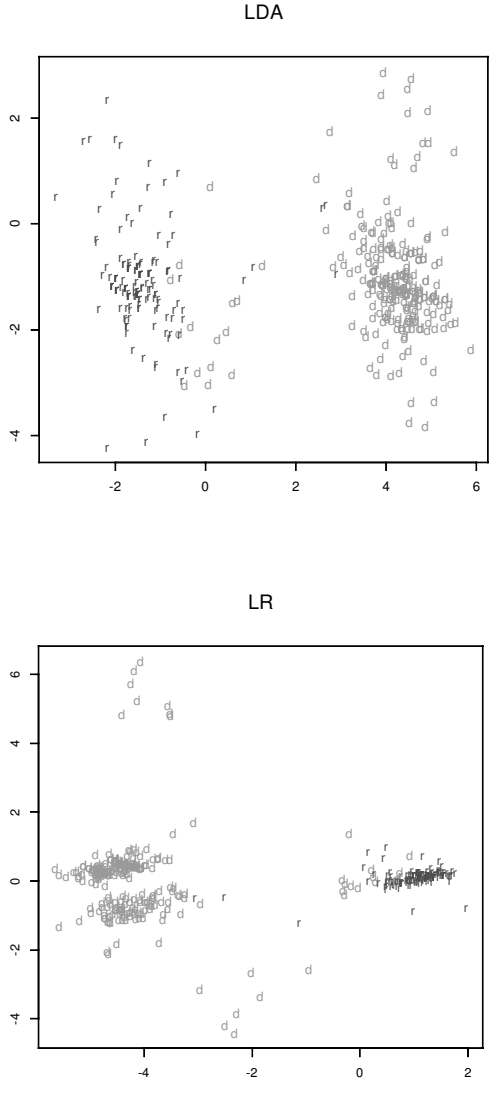

Figure 3. Congressional Votes Data. Top: Two-dimensional subspace identified by LDA—the second direction is not meaningful because the between-class covariance for a two-class problem has rank 1 . Bottom: Two-dimensional subspace identified by LR.

4. COMPARISON WITH SAVE

The sliced average variance estimator (SAVE) was used by Cook and Yin (2001) to recover the best discriminant subspace in more general settings than the LDA, for example, when p_k is $\mathrm{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. In their approach, they first sphered the data so that the total covariance matrix is identity and then chose their discriminant directions by successively maximizing

$$\mathrm{SAVE}(\boldsymbol{\alpha}) = \boldsymbol{\alpha}^T \left(\sum_{k=1}^{K} \left(\frac{n_k}{N} \right) (\mathbf{I} - \boldsymbol{\Sigma}_k)^2 \right) \boldsymbol{\alpha} \tag{4.1}$$

over $\boldsymbol{\alpha} \in \mathbb{R}^d$ where $\|\boldsymbol{\alpha}\| = 1$. The solutions are the top eigenvectors of the kernel,

$$\sum_{k=1}^{K} \left(\frac{n_k}{N} \right) (\mathbf{I} - \boldsymbol{\Sigma}_k)^2. \tag{4.2}$$

When p_k is $\mathrm{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, it is easy to verify that, apart from a constant not depending on $\boldsymbol{\alpha}$,

$$\mathrm{LR}(\boldsymbol{\alpha}) \propto \sum_{k=1}^{K} \left(\frac{n_k}{N} \right) \left(\log \boldsymbol{\alpha}^T \mathbf{T} \boldsymbol{\alpha} - \log \boldsymbol{\alpha}^T \boldsymbol{\Sigma}_k \boldsymbol{\alpha} \right) \tag{4.3}$$

$$= \sum_{k=1}^{K} \left(\frac{n_k}{N} \right) \left(- \log \boldsymbol{\alpha}^T \boldsymbol{\Sigma}_k \boldsymbol{\alpha} \right), \tag{4.4}$$

where $\mathbf{T}$ denotes the total covariance matrix. In order to make our comparison more direct, we also sphere the data first so that $\mathbf{T} = \mathbf{I}$. This is why $\log \boldsymbol{\alpha}^T \mathbf{T} \boldsymbol{\alpha} = \log 1 = 0$ in (4.4).

4.1 A SIMPLE NUMERICAL EXAMPLE

We first construct a simple numerical example (see also Hastie and Zhu 2001) to compare the two methods. For simplicity, we stay in $\mathbb{R}^2$ and let there be only $K = 2$ classes of equal prior probability ($n_1 = n_2 = N/2$). To simplify matters further, we let $\boldsymbol{\Sigma}_k$'s have common eigenvectors. Without loss of generality, simply assume that the $\boldsymbol{\Sigma}_k$'s are diagonal. We create two competing directions, one in which the two classes differ only by the mean (first-order difference), and one in which the two classes differ only by the variance (second-order difference). In particular, the parameters are

$$\boldsymbol{\mu}_1 = (-\sqrt{0.5}, 0)^T, \quad \boldsymbol{\mu}_2 = (\sqrt{0.5}, 0)^T$$

and

$$\mathbf{S}_1 = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.3 \end{pmatrix}, \quad \mathbf{S}_2 = \begin{pmatrix} 0.5 & 0 \\ 0 & 1.7 \end{pmatrix}.$$

One can easily verify that the overall mean is $(0, 0)^T$ and the total covariance matrix is the identity. In other words, the data are properly sphered. A simulated dataset for this example is

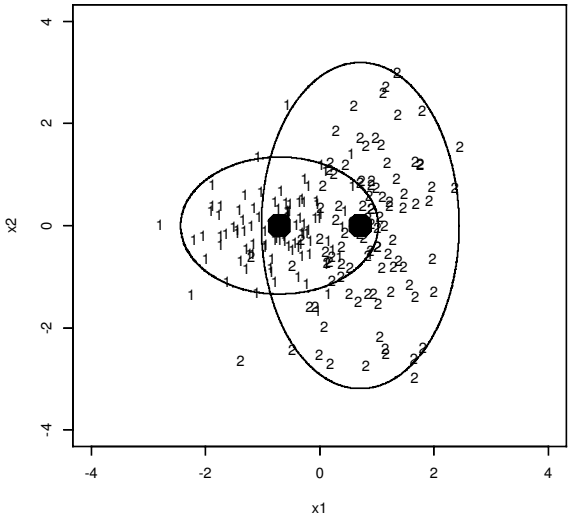

Figure 4. Simulated Data. Two Gaussians. The black dots are the centroids. There is a mean shift between the two classes in x_1 but no difference in the marginal variances. The means in x_2 are identical for the two classes while the marginal variances differ.

shown in Figure 4. In this case, the two eigenvectors of (4.2)—the only directions considered by SAVE—are simply the coordinate axes, $\mathbf{x}_1$ and $\mathbf{x}_2$. Table 1 shows the values of LR($\boldsymbol{\alpha}$) and SAVE($\boldsymbol{\alpha}$) evaluated at $\mathbf{x}_1$ and $\mathbf{x}_2$. We can see that SAVE will conclude that $\mathbf{x}_2$ is more important for discrimination while our criterion LR will conclude that $\mathbf{x}_1$ is the more important direction.

From Figure 4, it is hard to tell exactly which direction is actually more important for discrimination. However, in this simple case, we know the optimal Bayes decision rule if we were only to use one of the directions. If the decision were to be based on $\mathbf{x}_1$ alone, then the optimal Bayes rule would be:

$$\hat{y}_1 = \begin{cases} 1 & \text{if} \quad x_1 \leq 0, \\ 2 & \text{if} \quad x_1 > 0. \end{cases} \tag{4.5}$$

If the decision were to be based on $\mathbf{x}_2$ alone, then the optimal Bayes rule would be:

$$\hat{y}_2 = \begin{cases} 1 & \text{if} \quad |x_2| \leq 0.795, \\ 2 & \text{if} \quad |x_2| > 0.795. \end{cases} \tag{4.6}$$

Table 1. Evaluation of LR($\boldsymbol{\alpha}$) and SAVE($\boldsymbol{\alpha}$) Over the Two Eigenvectors of the SAVE Kernel, $\mathbf{x}_1 = (1,0)^T$ and $\mathbf{x}_2 = (0,1)^T$

$\boldsymbol{\alpha}$	*LR($\boldsymbol{\alpha}$)*	*SAVE($\boldsymbol{\alpha}$)*
$\mathbf{x}_1$	0.69	0.25
$\mathbf{x}_2$	0.34	0.49

Table 2. Misclassication Errors for the Two Different One-Dimensional Bayes Rules

Classication method	*Misclassication error*
$\hat{y}_1$	15.9%
$\hat{y}_2$	30.2%

The misclassification errors of these two rules can be easily calculated and are displayed in Table 2. Clearly, $\mathbf{x}_1$ is a much better discriminant direction than $\mathbf{x}_2$, yet if allowed to pick only one direction, SAVE would pick $\mathbf{x}_2$ while our generalized log-likelihood-ratio criterion would pick $\mathbf{x}_1$.

4.2 AN EXPERIMENT

We now carry this numerical example a bit farther by conducting a simple experiment. In this experiment, everything is kept the same as above except we now let $\boldsymbol{\mu}_1 = (-\sqrt{b}, 0)^T$ and $\boldsymbol{\mu}_2 = (\sqrt{b}, 0)^T$, where b is a free parameter.

It is easy to understand that as b changes, the Bayes misclassification errors of $\hat{y}_1$ and $\hat{y}_2$ also change. More specifically, when $b = 0$, it is clear that the direction $\mathbf{x}_1$ has no discriminating power; as b moves from 0 to 1, $\mathbf{x}_1$ gradually becomes the better discriminant direction than $\mathbf{x}_2$. The goal of this experiment is to answer the following questions: At what value of b will this reversal take place, and can the criteria LR($\boldsymbol{\alpha}$) and SAVE($\boldsymbol{\alpha}$) detect this reversal correctly?

An answer is provided in Figure 5. The top panel shows that the Bayes misclassification error of $\hat{y}_1$ drops from 50% to 0% as b moves from 0 to 1. The misclassification error of $\hat{y}_2$ stays constant since changes in b do not affect $\hat{y}_2$. The cross-over took place at around $b = 0.2$: when $b < 0.2$, $\hat{y}_2$ has smaller misclassification error, indicating that $\mathbf{x}_2$ should be the better discriminant direction than $\mathbf{x}_1$; when $b > 0.2$, the reverse is true. The middle panel shows the function LR($\boldsymbol{\alpha}$) evaluated at $\mathbf{x}_1$ and $\mathbf{x}_2$ as b changes. We can see LR starts to pick $\mathbf{x}_1$ over $\mathbf{x}_2$ as the better discriminant direction when b is over 0.3. The bottom panel shows the function SAVE($\boldsymbol{\alpha}$) evaluated at $\mathbf{x}_1$ and $\mathbf{x}_2$ as b changes. We see SAVE does not pick $\mathbf{x}_1$ over $\mathbf{x}_2$ as the better direction until b is over 0.7.

The conclusion from this simple experiment is that SAVE seems to over-emphasize second-order differences among the classes. The problem that SAVE over-emphasizes second-order information manifest itself quite regularly, as we illustrate with a real example below.

4.3 EXAMPLE: PEN DIGIT DATA

The Pen Digit database from the UCI machine-learning repository contains 10,992 samples of handwritten digits (0, 1, 2, . . . , 9) collected from 44 different writers. Each digit is stored as a 16-dimensional vector. The 16 attributes are derived using standard temporal and spatial resampling techniques in order to create vectors of the same length for every

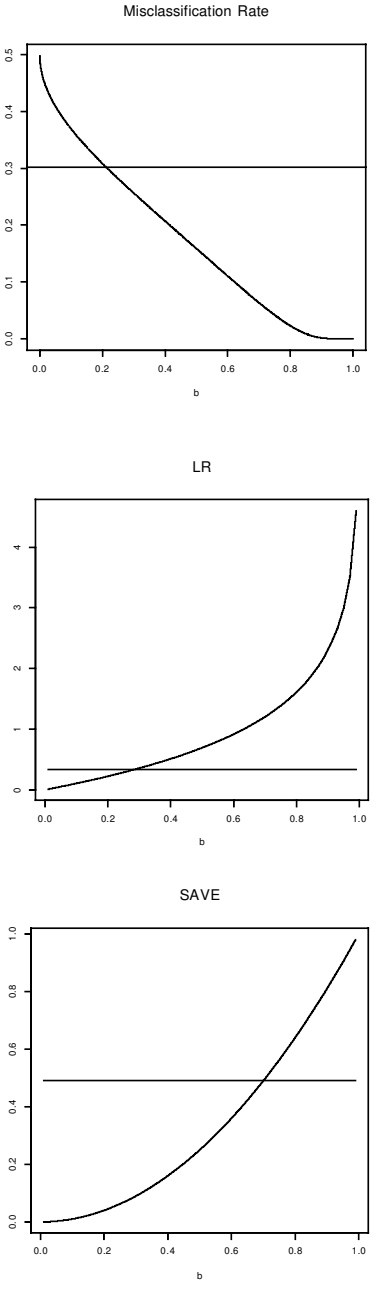

Figure 5. Top: Bayes misclassification errors of $\hat{y}_1$ and $\hat{y}_2$ as b changes, the horizontal line being that of $\hat{y}_2$. Middle: LR($\boldsymbol{\alpha}$) evaluated at $\mathbf{x}_1$ and $\mathbf{x}_2$ as b changes, the horizontal line being the one at $\mathbf{x}_2$. Bottom: SAVE($\boldsymbol{\alpha}$) evaluated at $\mathbf{x}_1$ and $\mathbf{x}_2$ as b changes, the horizontal line being the one at $\mathbf{x}_2$.

character. More details are available at ftp://ftp.ics.uci.edu/pub/machine-learning-databases /pendigits/. For our purposes, it suffices to state that this is a 10-class problem in a 16-dimensional space. The dataset is divided into a learning set (7,494 cases) and a test set (3,498 cases).

For better graphical display, we only select the 6's, 9's, and the 0's, three easily confused digits, as an illustration. For this three-class subproblem, there are 2,219 cases in the training set and 1,035 cases in the test set. We apply LDA, SAVE, and LR to the three-class subproblem. In each case, we search for the two leading discriminant directions. We then project the test data onto these two directions. In order to avoid having to justify the bandwidth we pick, the pictures presented here are produced with a parametric model, using the $N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ density as the model for p_k—note that this is close to the assumptions in SAVE but still more general than the assumptions in LDA. Figure 6 shows the results. This is a real problem, so we do not know the true discriminant subspace. However, from the graphs, it is clear that LDA separates the classes reasonably well. SAVE, on the other hand, picks a subspace in which the 9's have a much larger variance than the 0's and the 6's, while LR picks essentially the same subspace as LDA.

4.4 LR VERSUS SAVE

The experiments and examples above show the criterion LR is more robust than SAVE: while it is exible enough to pick up high-order differences among the classes, LR does not over-emphasize high-order information when the first-order difference is dominant.

5. NONPARAMETRIC DISCRIMINANT ANALYSIS: AN APPLICATION

Using projection pursuit density estimation (Friedman, Stuetzle, and Schroeder 1984), we can easily integrate our feature extraction technique into a non-parametric density model; this allows us to consider (reduced-rank) non-parametric discriminant analysis, a problem that has remained difficult due to the "curse of dimensionality." In particular, let

$$p_k(\mathbf{x}) = p_0(\mathbf{x}) \prod_{m=1}^{M} f_{mk}(\boldsymbol{\alpha}_m^T \mathbf{x}). \tag{5.1}$$

That is, one starts with a *common* initial guess for all the classes (often a Gaussian distribution), and augments each density function with a series of univariate ridge modifications to distinguish the classes, thereby bypassing the "curse of dimensionality." Model (5.1) can also be recursively written as

$$p_{m,k}(\mathbf{x}) = p_{m-1,k}(\mathbf{x}) f_{mk}(\boldsymbol{\alpha}_m^T \mathbf{x}).$$

For fixed direction $\boldsymbol{\alpha}_m$, the optimal ridge function f_{mk} for class k is given by (see Friedman et al. 1984):

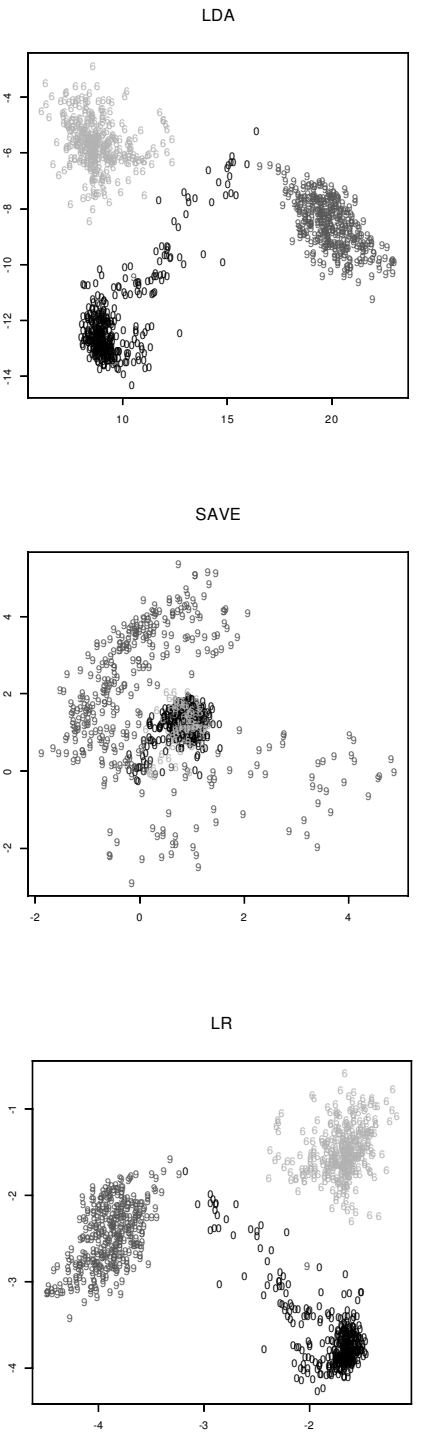

Figure 6. Pen Digit Data. Shown here are test data projected onto the leading two discriminant directions found by LDA, SAVE, and LR, respectively.

$$f_{mk}(\boldsymbol{\alpha}_m^T \mathbf{x}) = \frac{p_k^{(\alpha)}(\boldsymbol{\alpha}_m^T \mathbf{x})}{p_{m-1}^{(\alpha)}(\boldsymbol{\alpha}_m^T \mathbf{x})}.$$

So the important question here is how to choose the ridge directions $\boldsymbol{\alpha}_m$ at each iteration.

The idea of using model (5.1) for discriminant analysis was also proposed in Polzehl (1995), where $\boldsymbol{\alpha}_m$ was chosen to minimize the misclassification error. But misclassification error is, in itself, not a continuous function in $\boldsymbol{\alpha}$, so an additional smoothing step has to be introduced. Instead, we choose $\boldsymbol{\alpha}_m$ to maximize our general measure of class separation:

$$\mathrm{LR}(\boldsymbol{\alpha}) = \log \frac{\prod_{k=1}^K \prod_{\mathbf{x}_j \in C_k} p_{m,k}(\mathbf{x}_j)}{\prod_{k=1}^K \prod_{\mathbf{x}_j \in C_k} p_m(\mathbf{x}_j)} = \log \frac{\prod_{k=1}^K \prod_{\mathbf{x}_j \in C_k} p_{m-1,k}(\mathbf{x}_j) f_{mk}(\boldsymbol{\alpha}^T \mathbf{x}_j)}{\prod_{k=1}^K \prod_{\mathbf{x}_j \in C_k} p_{m-1}(\mathbf{x}_j) f_m(\boldsymbol{\alpha}^T \mathbf{x}_j)}.$$

The definition of $\mathrm{LR}(\boldsymbol{\alpha})$ here is slightly different from (3.1); the difference merely re ects the special projection pursuit density models being used for each class.

Here are some important advantages of this model, which, we feel, were not emphasized enough by Polzehl (1995): First, we have a exible, nonparametric model for the density functions in each class, while still avoiding the "curse of dimensionality." Second, forcing the ridge directions to be the *same* for all classes allows the complexity of the model to be easily *regularized* by selecting only a few directions where there are significant differences between classes. Therefore, although this is a density-based classification method, we actually do *not* waste any effort in directions which may contain interesting features in the density function itself but do not contribute to class differentiation. Third, because the initial density is *common* for all the classes, the decision boundary between classes k and K simply depends on the ratio of the augmenting ridge functions alone. In particular, the decision boundary is characterized by, assuming equal priors,

$$\prod_{m=1}^M \frac{f_{mk}(\boldsymbol{\alpha}_m^T \mathbf{x})}{f_{mK}(\boldsymbol{\alpha}_m^T \mathbf{x})} \stackrel{\text{def}}{=} \prod_{m=1}^M g_{mk}(\boldsymbol{\alpha}_m^T \mathbf{x}) = 1.$$

This implies that there is a generalized projection pursuit additive model (Roosen and Hastie 1991) for the posterior log-odds:

$$\log \frac{p(y=k|\mathbf{x})}{p(y=K|\mathbf{x})} = \log \frac{\pi_k}{\pi_K} + \log \frac{p_k(\mathbf{x})}{p_K(\mathbf{x})} \stackrel{\text{def}}{=} \beta_k + \sum_{m=1}^M \log(g_{mk}(\boldsymbol{\alpha}_m^T \mathbf{x})),$$

where π_k denotes the prior probability of class k.

5.1 ILLUSTRATION: EXCLUSIVE OR

As an illustration, we examine a difficult classification problem that involves an "exclusive or" relation in the predictors. Let

$$\boldsymbol{\mu}_A^1 = \begin{pmatrix} 1.5 \\ 1.5 \end{pmatrix}, \quad \boldsymbol{\mu}_A^2 = \begin{pmatrix} -1.5 \\ -1.5 \end{pmatrix},$$

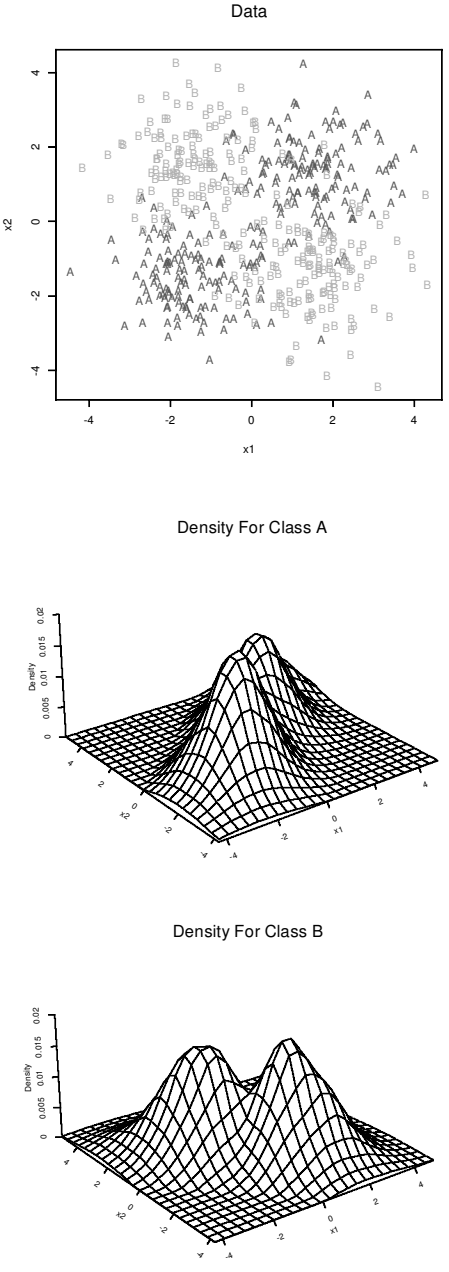

Figure 7. Top: A simulated instance of the "exclusive or" situation, with $n_A = n_B = 250$. Middle and Bottom: Estimated density functions for the two classes, after two ridge modifications.

Table 3. The Misclassication Error for 15,000 New Observations. The Bayes error for this problem is 12.5%.

Number of ridge functions (M)	*Test error*
1	23.5%
2	12.9%
3	14.2%
4	14.9%

$$\boldsymbol{\mu}_B^1 = \begin{pmatrix} -1.5 \\ 1.5 \end{pmatrix}, \quad \boldsymbol{\mu}_B^2 = \begin{pmatrix} 1.5 \\ -1.5 \end{pmatrix},$$

and consider two mixture classes, A and B, with

$$\begin{aligned} p_A(\mathbf{x}) &= \frac{1}{2}\phi(\mathbf{x}; \boldsymbol{\mu}_A^1, \mathbf{I}) + \frac{1}{2}\phi(\mathbf{x}; \boldsymbol{\mu}_A^2, \mathbf{I}) \\ p_B(\mathbf{x}) &= \frac{1}{2}\phi(\mathbf{x}; \boldsymbol{\mu}_B^1, \mathbf{I}) + \frac{1}{2}\phi(\mathbf{x}; \boldsymbol{\mu}_B^2, \mathbf{I}), \end{aligned}$$

where $\phi(\mathbf{x}; \boldsymbol{\mu}, \mathbf{I})$ denotes the $\mathrm{N}(\boldsymbol{\mu}, \mathbf{I})$ density function.

The top panel of Figure 7 shows one simulated instance of this situation. This is a particularly difficult problem, because when considered alone, neither $\mathbf{x}_1$ nor $\mathbf{x}_2$ contains any information for classification; they must be considered together. The Bayes decision rule for this case is an "exclusive or" condition:

$$\hat{y} = \begin{cases} \text{B} & \text{if (not } (x_1 > 0) \text{ and } (x_2 > 0)) \text{ or (not } (x_2 > 0) \text{ and } (x_1 > 0)), \\ \text{A} & \text{otherwise.} \end{cases}$$

The symmetry makes it easy to compute that the Bayes error for this problem is equal to $2\Phi(1.5)(1 - \Phi(1.5)) \approx 12.5\%$. Using model (5.1), we start with $p_A = p_B = p_0$, where p_0 is Gaussian, and modify each by a series of ridge functions. The ridge directions are chosen to be directions in which there are significant differences between the two classes as measured by $\mathrm{LR}(\boldsymbol{\alpha})$. The middle and bottom panels of Figure 7 show the resulting density surfaces after two ridge modifications. The ridge functions also determine the classification rule. The misclassification results on test data are summarized in Table 3. We can see that when we use two ridge modifications for each density function (the optimal number for this problem), we can achieve an error rate very close to the Bayes error, despite the difficulty of the problem. The fact that the test error increases for $M > 2$ is a sign of over-fitting caused by the additional and unnecessary ridge modifications. In general, one must rely on cross-validation (or similar methods) to select the optimal M.

6. SUMMARY

We have proposed a general, automatic, and adaptive feature extraction method for classification and pattern recognition, of which LDA is a special case. It has the ability to pick up rather complicated features (such as within-group clusters) that are important

for distinguishing the classes. Unlike its competitors such as SAVE, it also has the ability to capture the right amount of trade-off between low-order and high-order features. We also showed that our method can be incorporated into formal probabilistic models to allow (low-rank) nonparametric discriminant analysis. The equivalence between discriminant directions and the concept of ordination axes in correspondence analysis (Zhu 2001) also makes this method applicable to a wide variety of applications in areas such as environmental ecology and e-commerce.

ACKNOWLEDGMENTS

We thank Professors Robert Tibshirani and Jerome Friedman for their interesting discussions and helpful comments during this work. We are also indebted to an associate editor for comments and suggestions which greatly improved our manuscript. Both authors were partially supported by grant DMS-9803645 from the National Science Foundation, and grant ROI-CA-72028-01 from the National Institutes of Health. In addition, Mu Zhu was also partially supported by the Natural Sciences and Engineering Research Council of Canada.

[Received February 2001. Revised August 2002.]

REFERENCES

Cook, R. D., and Yin, X. (2001), "Dimension Reduction and Visualization in Discriminant Analysis" (with discussion), *Australian and New Zealand Journal of Statistics*, 43, 147–199.

Devijver, P. A., and Kittler, J. (1982), *Pattern Recognition: A Statistical Approach*, Englewood Cliffs, NJ: Prentice-Hall International.

Fisher, R. A. (1936), "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, 7.

Friedberg, S. H., Insel, A. J., and Spence, L. E. (1989), *Linear Algebra*, Englewood Cliffs, NJ: Prentice Hall.

Friedman, J. H. (1987), "Exploratory Projection Pursuit," *Journal of the American Statistical Association*, 82, 249–266.

Friedman, J. H., Stuetzle, W., and Schroeder, A. (1984), "Projection Pursuit Density Estimation," *Journal of the American Statistical Association*, 79, 599–608.

Gill, P. E., Murray, W., and Wright, M. H. (1981), *Practical Optimization*, New York: Academic Press.

Hastie, T. J., and Zhu, M. (2001), Discussion of "Dimension Reduction and Visualization in Discriminant Analysis," by Cook and Yin, *Australian and New Zealand Journal of Statistics*, 43, 179–185.

Loader, C. (1999), *Local Regression and Likelihood*, New York: Springer-Verlag.

Polzehl, J. (1995), "Projection Pursuit Discriminant Analysis," *Computational Statistics and Data Analysis*, 20, 141–157.

Roosen, C., and Hastie, T. J. (1991), "Logistic Response Projection Pursuit," Technical Report BLO11214-930806-09TM, AT&T Bell Laboratories.

Silverman, B. (1986), *Density Estimation for Statistics and Data Analysis*, New York: Chapman and Hall.

Zhu, M. (2001), "Feature Extraction and Dimension Reduction with Applications to Classification and the Analysis of Co-occurrence Data," Ph.D. dissertation, Stanford University.

This article has been cited by:

1. Santiago Velilla . 2008. A Method for Dimension Reduction in Quadratic Classification ProblemsA Method for Dimension Reduction in Quadratic Classification Problems. *Journal of Computational and Graphical Statistics* **17**:3, 572-589. [Abstract] [PDF] [PDF Plus] [Supplementary material]
2. Iain Pardoe , Xiangrong Yin , R. Dennis Cook . 2007. Graphical Tools for Quadratic Discriminant AnalysisGraphical Tools for Quadratic Discriminant Analysis. *Technometrics* **49**:2, 172-183. [Abstract] [PDF] [PDF Plus]
3. Adolfo Hernández , Santiago Velilla . 2005. Dimension Reduction in Nonparametric Kernel Discriminant AnalysisDimension Reduction in Nonparametric Kernel Discriminant Analysis. *Journal of Computational and Graphical Statistics* **14**:4, 847-866. [Abstract] [PDF] [PDF Plus]