Journal of the Korean Statistical Society (2001), **30**: *2, pp 193–217*

Generalization of Fisher's linear discriminant analysis via the approach of sliced inverse regression

Chun-Houh Chen [1] and Ker-Chau Li[2]

Abstract

Despite of the rich literature in discriminant analysis, this complicated subject remains much to be explored. In this article, we study the theoretical foundation that supports Fisher's linear discriminant analysis (LDA) by setting up the classification problem under the dimension reduction framework as in Li (1991) for introducing sliced inverse regression (SIR). Through the connection between SIR and LDA, our theory helps identify sources of strength and weakness in using CRIMCOORDS(Gnanadesikan 1977) as a graphical tool for displaying group separation patterns. This connection also leads to several ways of generalizing LDA for better exploration and exploitation of nonlinear data patterns.

Keywords: Data visualization, Dimension reduction, Dynamic graphics, e.d.r. directions, Fisher's linear discriminant analysis, Nonparametric density estimation, SIR.

1. Introduction

Discriminant analysis aims at the classification of an object into one of K given classes based on information from a set of p predictor variables. Among the many available methods, the simplest and most popular approach is linear discriminant analysis (LDA). This article investigates the theoretical foundation of LDA under the dimension reduction setting of Li(1991). The connection of LDA to sliced inverse regression (SIR) is established and exploited. This leads to a variety of ways to generalize LDA so that nonlinear features in the training data can be better explored and incorporated into the discriminant rule.

A most well-known property for LDA is that LDA is a Bayes rule under a normality condition about the predictor distribution. More precisely, the condition

[1] Institute of Statistical Science, Academia Sinica, Taipei, Taiwan

[2] Department of Mathematics, UCLA, Los Angeles, CA 90024

requires that for the ith class, $i = 1, \cdots, K$, the p-dimensional predictor variable $\mathbf{x} = (x_1, \cdots, x_p)'$ follows a multi-variate normal distribution with mean μ_i and a common covariance Σ_c. Together with the prior probability π_i, $i = 1, ..., K$, about the relative occurrence frequency for each class, this assumption leads to a Bayes discriminant rule which coincides with the rule of LDA.

Another way of deriving LDA originates from the consideration about group separation when there are only two classes, $K = 2$ (Fisher 1936, 1938). The idea is to find a linear combination of the predictors, $z = a_1x_1 + \cdots, a_px_p$, that exhibits the largest difference in the group means relative to the within-group variance. The derived variate z is known as Fisher's discriminant function, or the first canonical variate. Fisher's result is further generalized by Rao(1952, Sec 9c) to the multiple class problem, $K \geq 2$. In general, after finding the first r canonical variates, the $(r + 1)$th canonical variate is the next best linear combination z that can be obtained subject to the constraint that z must be uncorrelated to all canonical variates obtained earlier. Canonical variates are also referred to as the discriminant coordinates (CRIMCOORDS) in Gnanadesikan(1977).

Empirical evidence has shown that scatterplots of the first few CRIMCOORDS can reveal interesting clustering patterns. Such graphical displays are helpful in studying the degree and nature of class separation and for detecting possible outliers. However, the nonlinear patterns often observed in such plots also point to the limitation of the normality assumption in justifying LDA. The data points within each class do not always appear elliptically distributed. Even if they do appear so, they hardly have the same orientation-violating the equal covariance assumption.

The motivation of our study stems from the concern about the theoretic foundation of LDA. To what extent, can LDA be applied effectively without the normality assumption? In what sense, can the reduction from the original p predictors to the first few CRIMCOORDS be deemed "effective"? Are there any other linear combinations more useful than the CRIMCOORDS in providing graphical information about group separation? If so, how can one find them? In this article, we address these issues by formulating the classification problems via the dimension reduction approach of Li(1991). A key notion in that article is the effective dimension reduction (*e.d.r.*) space for general regression problems.

Our paper is organized in the following way. In Section 2, we review the dimension reduction approach and bring out the connection of sliced inverse regression(SIR) with LDA. It turns out that the e.d.r. directions found by SIR are proportional to the vectors of the coefficients used in the canonical variates.

Via this connection, the theory of SIR can be applied to offer a new theoretical support for using CRIMCOORDS.

Prior information about the occurrence frequency for each class plays a crucial role in discriminant analysis. It is certainly needed in forming a Bayes rule. But how critical is it for dimension reduction? This issue is discussed in Section 3. We argue that dimension reduction can be pursued independent of the specification of a prior distribution.

LDA can be viewed as a two-stage procedure. The first stage is to find the canonical variates for reducing the predictor dimension from p to K or less; the second stage is to split the canonical space linearly into K regions for class-membership prediction via the Mahalanobis distance. While the SIR theory justifies the use of canonical variates at the first stage, the theory itself does not support the use of linear split rules at the second stage. Section 4 discusses this issue. Nonparametric classification rules can be formed using the first few canonical variates found at the first stage of LDA.

As is known, the first moment based SIR does not always work in finding the entire e.d.r. space. Such knowledge about when SIR will fail helps identify sources of potential weakness in using CRIMCOORDS. An important special case is when there are only $K = 2$ classes. There is only one CRIMCOORD available now, no matter how complex the true dimension reduction model is. This may not be enough for locating the entire e.d.r. space because the e.d.r. space can have more than one dimension. In Section 5, more general methods will be considered. There are two types of generalization. The first one follows the thoughts of Principal Hessian directions (PHD) (Li 1992a). It amounts to the comparison of the second moments of the predictors between classes. The second type of generalization explores an idea of double-slicing. Several simulation examples are provided and an application to a real data set is given.

Further discussion and some concluding remarks are given in Section 6.

2. SIR and Fisher's canonical variates

In this section, the relationship between SIR and canonical variates is established first. Then the assumptions used to guarantee the success of SIR are discussed in the context of classification. These assumptions provide more general theoretical support for the use of canonical variates than the well-known normality assumption underlying LDA.

2.1. Connection

For discussing the issue of visualization and dimension reduction in general regression problems, Li(1991) considers the model

$$Y = g(\beta_1'\mathbf{x}, \cdots, \beta_d'\mathbf{x}, \epsilon). \tag{2.1}$$

Here Y is the response variable, and g is an unknown function with $(d+1)$ arguments. The random error ϵ is independent of the p-dimensional regressor $\mathbf{x}$, but its distribution is unknown. The space spanned by the β vectors is called the *e.d.r.* space. Any vector b in the e.d.r. space is referred to as an e.d.r. vector and any linear combination $b'\mathbf{x}$ is called an e.d.r. variate. (2.1) represents the situation in which Y is related to $\mathbf{x}$ only through the e.d.r. variates. When d is smaller than p, one can reduce the regressor dimension from p to d by finding the e.d.r. directions. Plots of Y against the e.d.r. variates will be more informative than those against non-e.d.r. variates in revealing the regression structure. Cook (1994) offers an extensive discussion on the notion of e.d.r. directions.

Sliced inverse regression is a simple method for finding e.d.r. directions. We describe the population version of SIR first. Denote the covariance matrix of $\mathbf{x}$ by $\Sigma_{\mathbf{x}}$. The central idea of SIR is to reverse the roles of $\mathbf{x}$ and Y. Instead of regressing Y on $\mathbf{x}$, we may consider the inverse regression curve $E(\mathbf{x}|Y) = (E(x_1|Y), \cdots, E(x_p|Y))'$. In general, this curve is in the p dimensional space. However, Theorem 3.1 of Li(1991) shows that under (2.1) and another condition to be discussed later, the inverse regression curve indeed falls into a d dimensional subspace. This subspace is determined only by the e.d.r. directions and $\Sigma_{\mathbf{x}}$. Denote the covariance matrix of the random vector $\eta = E(\mathbf{x}|Y)$ by $\Sigma_\eta = cov(\eta) = cov(E(\mathbf{x}|Y))$. We are led to the following eigenvalue decomposition for finding e.d.r. directions:

$$\begin{aligned} \Sigma_\eta b_i &= \lambda_i \Sigma_{\mathbf{x}} b_i \\ \lambda_1 &\geq \cdots \geq \lambda_p, \end{aligned} \tag{2.2}$$

Li's theorem implies that all but the first K eigenvalues must be zero and that the eigenvectors associated with nonzero eigenvalues are the e.d.r. directions.

The sample version of SIR is easy to obtain. We simply substitute Σ_η and $\Sigma_{\mathbf{x}}$ in (2.2) by their estimates from an i.i.d. sample $(Y_i, \mathbf{x}_i), i = 1, \cdots, n$. The estimate of $\Sigma_{\mathbf{x}}$ is just the sample covariance $\hat{\Sigma}_{\mathbf{x}} = n^{-1}\sum_{i=1}^n(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$. Here $\bar{\mathbf{x}}$ denotes the sample mean. The estimate of Σ_η can be formed by first partitioning the response variable Y into H intervals, $I_h, h = 1, \cdots, H$. Within each slice, compute the mean of $\mathbf{x}$, $\hat{\mathbf{m}}_h = n_h^{-1} \sum_{Y_i \in I_h} \mathbf{x}_i$, where n_h is the number of cases in slice h. These slice means constitute a simple estimate of $E(\mathbf{x}|Y)$ and they can be combined to give a weighted covariance matrix, $\hat{\Sigma}_\eta = n^{-1} \sum_{j=1}^{H} n_j(\hat{\mathbf{m}}_j - \bar{\mathbf{x}})(\hat{\mathbf{m}}_j - \bar{\mathbf{x}})'$, for estimating Σ_η. The eigenvectors $\hat{b}_i$'s are the SIR directions and we shall call $\hat{b}_i'\mathbf{x}$ the SIR variates. Large sample properties of SIR directions and a chi-squared test about the significance of eigenvalues $\hat{\lambda}_i$'s are given in Li(1991). More recent results on SIR can be found in Hsing and Carroll (1992), Li (1992), Schott (1994), Zhu and Ng (1995), Zhu and Fang (1996), and Chen and Li(1998).

The examples and discussion in Li(1991) focused on the case where the response variable Y is continuous. But the continuity of Y is not required in (2.1). In fact, when Y is discrete and can take only K distinct values, the slicing step of SIR is automatic for $H = K$. This special circumstance fits well into our classification problem. We can regard each $(\mathbf{x}_i, Y_i)$ as one case in the training set and the response variable Y_i is just the class label for that case. The slice mean $\hat{\mathbf{m}}_j$ corresponds to the vector of the predictor's mean for the jth group. The matrix $\hat{\Sigma}_\eta$ coincides with the between group variance-covariance matrix in one-way multivariate analysis of variance (MANOVA).

To elucidate how canonical variates are related to the e.d.r. directions found by SIR, recall that the first canonical variate is derived by maximizing the ratio of the between-group variance to the within-group variance. In our notation, for a linear combination $z = \mathbf{a}'\mathbf{x}$, the group means are just $\mathbf{a}'\hat{\mathbf{m}}_j, j = 1, \cdots, K$. The between-group variance, $n^{-1} \sum n_j(\mathbf{a}'\hat{\mathbf{m}}_j - \mathbf{a}'\bar{\mathbf{x}})^2$, can be written as $\mathbf{a}'\hat{\Sigma}_\eta\mathbf{a}$. On the other hand, the within-group variance can be written as $n^{-1} \sum_{i=1}^{n}(\mathbf{a}'\mathbf{x}_i - \mathbf{a}'\hat{\mathbf{m}}_{j(i)})^2 = \mathbf{a}'\hat{\Sigma}_e\mathbf{a}$, where the class membership for the i-th case is denoted by $j(i)$ and $\hat{\Sigma}_e$ is the within-group variance-covariance matrix, $n^{-1} \sum_{j=1}^{k} n_j \hat{\Sigma}_j$. The first canonical variate is the linear combination of $\mathbf{x}$ formed by the vector $\mathbf{a}$ which solves the following maximization problem:

$$\max_{\mathbf{a}} \frac{\mathbf{a}'\hat{\Sigma}_\eta\mathbf{a}}{\mathbf{a}'\hat{\Sigma}_e\mathbf{a}}, \tag{2.3}$$

The solution of (2.3) corresponds to the largest eigenvector of the following eigenvalue decomposition:

$$\begin{aligned} &\hat{\Sigma}_\eta \mathbf{a}_i = \hat{\gamma}_i \hat{\Sigma}_e \mathbf{a}_i, \\ &\hat{\gamma}_1 \geq \hat{\gamma}_2 \geq \cdots \geq \hat{\gamma}_p \end{aligned} \tag{2.4}$$

To see the connection with SIR, we can rearrange the above eigenvalue decomposition equation by adding $\hat{\gamma}_i\hat{\Sigma}_\eta \mathbf{a}_i$ on both sides :

$$(1+\hat{\gamma}_i)\hat{\Sigma}_\eta \mathbf{a}_i = \hat{\gamma}_i(\hat{\Sigma}_\eta + \hat{\Sigma}_e)\mathbf{a}_i$$

Now we can use the identity that the sum of the between-group variance and within-group variance equals the total variance, $\hat{\Sigma}_{\mathbf{x}} = \hat{\Sigma}_\eta + \hat{\Sigma}_e$, to obtain :

$$\hat{\Sigma}_\eta \mathbf{a}_i = \frac{\hat{\gamma}_i}{1+\hat{\gamma}_i}\hat{\Sigma}_{\mathbf{x}}\mathbf{a}_i \tag{2.5}$$

Comparing this equation with the sample version of (2.2), we see that $\hat{\lambda}_i = \hat{\gamma}_i/(1+\hat{\gamma}_i)$, and $\mathbf{a}_i \propto \hat{b}_i$. We now reach the following observation.

Observation I : *The SIR variates are the same as the canonical variates except for possible differences in scaling.*

Canonical variates are often associated with LDA, which can only be theoretically justified under the normality assumption :

$$\mathbf{x}|Y=j \sim N(\mu_j, \Sigma_c). \tag{2.6}$$

If we further assume that

$$\text{the vectors } \mu_j - \mu_1,\ j=2,\cdots,K, \text{ spans a } d \text{ dimensional space,} \tag{2.7}$$

then the Bayes discriminant rule will depend on $\mathbf{x}$ only through the first d canonical variates. This is the traditional way of justifying the use of only the first few significant canonical variates in applying LDA. But (2.6) is apparently too stringent. In fact, one can even argue that if the predictors' distribution is normal, then there won't be any interesting patterns to see in the CRIMCOORDS plots. Thus to fully justify the merit of CRIMCOORDS, we need something entirely different.

By relating the canonical variates with SIR variates, Observation I brings in a very broad context for using CRIMCOORDS to reduce the dimension of the predictors. This is because SIR is developed under much weaker conditions. We shall discuss these conditions next.

2.2. Condition (2.1)

The theory of SIR is founded on two assumptions. One of them is the dimension reduction model (2.1). A general comparison of (2.1) to (2.6)-(2.7) can be made more clear by re-formulating (2.1) from the inverse regression point of view.

Put $B = (\beta_1, \cdots, \beta_d)$. (2.1) implies that the conditional density of Y given $\mathbf{x}$, $f(Y|\mathbf{x})$ depends only on $B'\mathbf{x}$; $f(Y|\mathbf{x}) = f(Y|B'\mathbf{x})$. Thus the conditional density of $\mathbf{x}$ given Y can be written as

$$\begin{aligned} f(\mathbf{x}|Y) &= \frac{f(Y|\mathbf{x})f_{\mathbf{x}}(\mathbf{x})}{f_Y(Y)} = \frac{f(Y|B'\mathbf{x})f_{\mathbf{x}}(\mathbf{x})}{f_Y(Y)} \\ &= \frac{f(Y, B'\mathbf{x})f_{\mathbf{x}}(\mathbf{x})}{f_Y(Y)f_{B'\mathbf{x}}(B'\mathbf{x})} = f(B'\mathbf{x}|Y)\frac{f_{\mathbf{x}}(\mathbf{x})}{f_{B'\mathbf{x}}(B'\mathbf{x})} \end{aligned} \tag{2.8}$$

Here all f with subscripts are marginal density functions.

For classification problems, the rightmost side in the expression (2.8) gives a useful factorization for comparing the predictor distributions in different classes. This can be summarized by the following statement:

Observation II. *For classification problems, (2.1) is equivalent to the condition that for any two classes, j and j', the ratio of their density functions of* $\mathbf{x}$ *depends only on* $B'\mathbf{x}$:

$$\frac{f(\mathbf{x}|Y=j)}{f(\mathbf{x}|Y=j')} = \frac{f(B'\mathbf{x}|Y=j)}{f(B'\mathbf{x}|Y=j')} \tag{2.9}$$

It is straightfoward to verify that (2.6) and (2.7) imply (2.9) if we take $\beta_1, \cdots, \beta_d$ to be any basis of the space spanned by the differences in μ_i's.

2.3. Condition on the predictor distribution

In addition to (2.1) (or equivalently (2.9) for classification problems), SIR requires another condition on the distribution of $\mathbf{x}$: for any $b \in R^p$,

$$\text{the conditional expectation } E(b'\mathbf{x}|\beta_1'\mathbf{x}, \cdots, \beta_d'\mathbf{x}) \text{ is linear.} \tag{2.10}$$

(2.10) is the same as the condition that for any variate $\mathbf{a}'\mathbf{x}$,

$$cov(\mathbf{a}'\mathbf{x}, B'\mathbf{x}) = 0 \text{ implies } E(\mathbf{a}'\mathbf{x}|B'\mathbf{x}) = \mathbf{a}'E\mathbf{x}, \tag{2.11}$$

(2.11) is much weaker than (2.6)-(2.7). Normality assumption is not needed here. Within group-covariances also need not be entirely the same.

One sufficient condition for (2.10) (or equivalently (2.11)) to hold is that

$$\mathbf{x} \text{ follows an elliptically-contoured distribution.} \tag{2.12}$$

But this often leads to the impression that (2.12) is equivalent to (2.10). A counter-example to this impression is indeed the normal model, (2.6) and (2.7).

As a mixture of normal distributions, the marginal distribution of $\mathbf{x}$ certainly cannot be elliptically symmetric.

The above false impression comes from a conservative view on when to apply the SIR methodology. If we want condition (2.10) to hold for *all* β vectors (including those not in the e.d.r. space), then as pointed out by Cook and Weisberg (1991), no distributions other than the elliptical ones will do.

A less conservative attitude seems more appropriate. First of all, the bias of SIR will not be significant under mild violations of (2.10). Consider the set of B with the violation of (2.10) is more than a specified amount. The result of Hall and Li (1993) assures that this set becomes smaller when the dimension of $\mathbf{x}$ gets larger. This leaves a lot of room for SIR to work well even without worrying about (2.10) prior to the analysis. After applying SIR, we can follow the analysis by a diagnostic check on this condition. On the other hand, subsampling and/or reweighting processes can also be carried out to fortify (2.11); Brillinger (1991), Cook and Natsheim (1994), Li (1991).

Remark 2.1. SIR variates are scaled to have unitary variance but canonical variates are usually scaled to have unitary *within-group* variance. Since the covariance is no longer the same for every group, we prefer the way SIR variates are scaled.

3. Prior distribution and dimension reduction

The discussion in Section 2 assumes that the training set consists of *i.i.d* observations from the same population as the target population where the test set will come from. This may not be the case in some applications. This section discusses the case that the training sample is obtained by stratified sampling. More specifically, a pre-specified number n_j of cases are drawn independently from each class j. The sampling allocation n_j/n does not always match the prior $\pi_j(= P\{Y = j\})$, the probability that a random test case from the target population falls into group j. Recall that under the 0-1 loss, the Bayes rule classifies a future observation by

$$\max_y \pi_y f_{\mathbf{x}|Y}(\mathbf{x}|y). \tag{3.1}$$

Now suppose the target population follows a dimension reduction model (2.1), or equivalently (2.9). We can translate (3.1) into

$$\max_y \pi_y f(B'\mathbf{x}|y). \tag{3.2}$$

This shows that in order to find the Bayes rule, we only have to focus on the *e.d.r.* variates.

The next question is whether SIR is still applicable for finding the e.d.r. space under stratified sampling. To answer this question, we study the population version of SIR by letting n_j tend to the infinity; while fixing $p_j = n_j/n$. We notice that SIR takes the same form as (2.2) but with a slightly different interpretation about the two covariance operators. By fixing $p_j = n_j/n$, Σ_η is still the between group variance-covariance matrix as in the one-way MANOVA with the weight for group j being p_j(instead of π_j). Similarly, $\Sigma_\mathbf{x}$ is the overall sample covariance of $\mathbf{x}$.

Theorem 3.1. *Suppose the sample is drawn by stratified sampling. Then under (2.9) and (2.11), the eigenvectors with nonzero eigenvalues in the eigenvalue decomposition(2.2) fall into the e.d.r. space.*

Proof. From (2.11), we see that for any $\mathbf{a}$ such that $\mathbf{a}'\Sigma_\mathbf{x}B = 0$, we must have $\mathbf{a}'\Sigma_\eta\mathbf{a} = 0$, or equivalently $\Sigma_\eta\mathbf{a} = 0$. This shows that the eigenspace for (2.2) associated with the zero eigenvalue must contain any such vector $\mathbf{a}$. Since all non-zero eigenvectors b_j must be orthogonal to $\mathbf{a}$, *i.e.* $\mathbf{a}'\Sigma_\mathbf{x}b_j = 0$, with respect to $\Sigma_\mathbf{x}$, they must fall into the column space of B. The theorem is now proved.

4. Nonparametric regression after SIR

Observation I, Observation II and Theorem 3.1 provide a general theoretical foundation for LDA. But this only justifies the first stage of LDA, namely using the canonical covariates to reduce the dimension. The further use of linear split rule can only be justified under normality assumption on the distributions for the e.d.r. variates are completely arbitrary. Without the normality assumption, it seems natural to apply nonparametric density estimation techniques after dimension reduction. For illustration, we shall discuss only the standard kernel estimation here. Other nonparametric procedures can similarly be applied.

Let $\mathbf{x}_{yi}, i = 1, \cdots, n_y$ be the sample drawn from class $Y = y$. The SIR directions, $\hat{b}_1, \cdots, \hat{b}_d$, converge to $b_1, \cdots, b_d$ respectively at the usual root n rate, provided that all d nonzero eigenvalues are distinct. The kernel estimate of the density function of $B'\mathbf{x}$ for class $Y = y$ takes the following form:

$$\hat{f}_{B'\mathbf{x}}(t_1, \cdots, t_d) = \frac{1}{nh^d}\sum_{i=1}^{n_y}\Pi_{j=1}^{d}\mathcal{K}(\frac{\hat{b}_j'\mathbf{x}_{yi} - t_j}{h}), \tag{4.1}$$

where the kernel $\mathcal{K}(\cdot)$ is a one-dimensional density function. The bandwidth h has to converge to 0 at an appropriate rate.

(4.1) can be compared to the "theoretical" kernel density estimate, should we be given B exactly:

$$\tilde{f}_{B'\mathbf{x}}(t_1, \cdots, t_d) = \frac{1}{nh^p} \sum_{i=1}^{n_y} \Pi_{j=1}^{k} \mathcal{K}(\frac{b_j'\mathbf{x}_{yi} - t_j}{h}). \tag{4.2}$$

The consistency of (4.2) for estimating $f_{B'\mathbf{x}}(t_1, \cdots, t_d)$ is the subject of standard kernel density estimation. This allows us to conclude that the discriminant rule obtained by substituting $f(B'\mathbf{x}|y)$ in (3.2) by the kernel estimate (4.1) is asymptotically Bayes.

Example 4.1 Wave recognition. This example is taken from Breiman et al. (1984, pp 49-55); see also Loh and Vanichsetakul (1988). There are three classes and 21 variables. Three triangular basic waveforms $w_1(\cdot), w_2(\cdot), w_3(\cdot)$, are involved: for $j = 1, \cdots, 21$,

$$w_1(j) = max(6 - |j - 11|, 0); \quad w_2(i) = w_1(j - 4); \quad w_3(j) = w_1(j + 4). \tag{4.3}$$

Each class is a random convex combination of two basic waveforms with noise added. Let $\mathbf{w}_i = (w_i(1), \cdots, w_i(21))', i = 1, 2, 3$, and u_1, u_2, u_3 be independent random variables uniformly distributed on $[0, 1]$. The predictor $\mathbf{x}$ is generated by

$$\begin{aligned} \mathbf{x} &= u_1\mathbf{w}_1 + (1 - u_1)\mathbf{w}_2 + \epsilon, \quad \text{for} Y = 1 \\ &= u_2\mathbf{w}_2 + (1 - u_2)\mathbf{w}_3 + \epsilon, \quad \text{for} Y = 2 \\ &= u_3\mathbf{w}_3 + (1 - u_3)\mathbf{w}_1 + \epsilon, \quad \text{for} Y = 3, \end{aligned} \tag{4.4}$$

where ϵ follows the standard normal distribution.

The two-dimensional vector space spanned by $\mathbf{w}_1 - \mathbf{w}_2, \mathbf{w}_3 - \mathbf{w}_1$ is the e.d.r. space. This can be seen by verifying (2.9).

We generate 200 cases from each group as the training sample. Then SIR is applied. Only the first two eigenvalues are nonzero, 0.651 and 0.546. After projecting the predictors along the first and the second SIR directions, kernel density estimation is applied to get the boundaries of Bayes classification rules for the uniform prior distribution $\pi_y = 1/3$ and the prior distribution $\pi_y = y/6$ respectively, Figure 4.1(a)-(b). Classification boundaries are seen to be approximately linear. This is as expected. In fact, SIR variates for the population version can be represented by mixtures of normals with means being on a equilateral triangular, Figure 4.1(c). By a geometric argument, we can show that the contours for the likelihood ratios must be straight lines.

Another interesting feature about this example is that the e.d.r. space does not depend on the distribution of u_y, $y = 1, 2, 3$. We generate another 200 cases from each group but with u_i from the density $f(u) = 3u^2$ for $u \in [0, 1]$. Apply SIR and kernel estimation again. For equal prior $\pi_y = 1/3$, the result is shown in Figure 4.1(d). Now the Bayes rules are nonlinear.

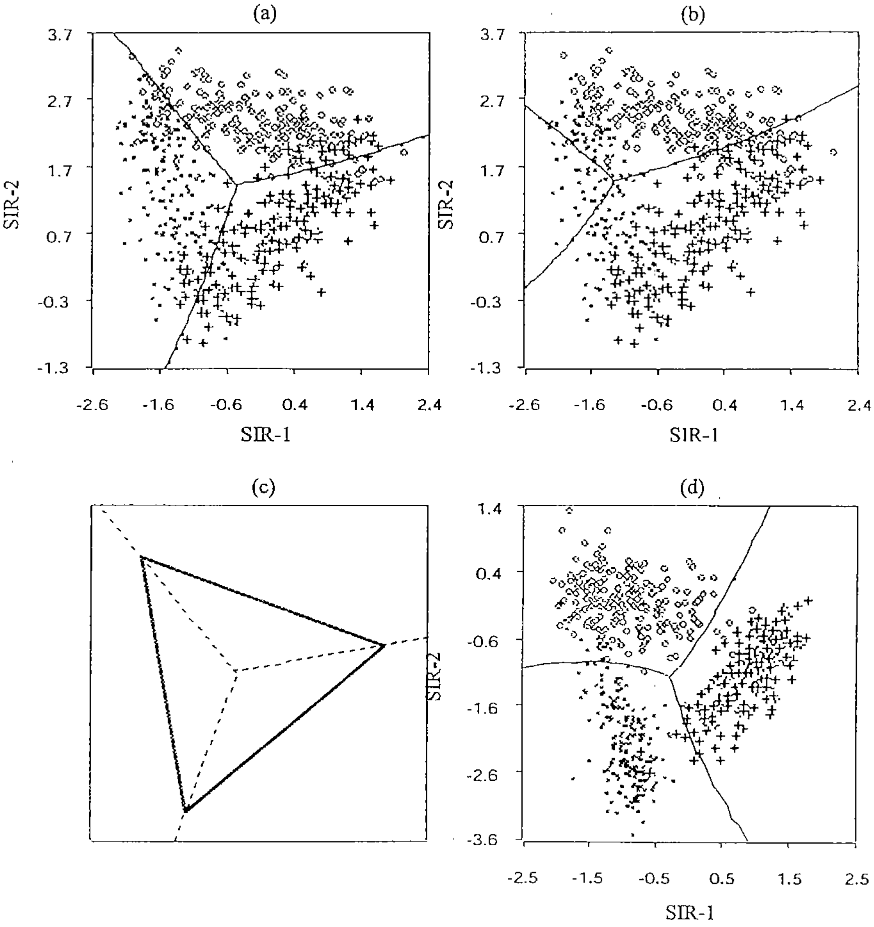

Figure 4.1: Wave Recognition Problem: (a) SIR's View with Equal Contour Boundary, $\pi_y = 1/3$; (b) SIR's View with Equal Contour Boundary, $\pi_y = y/6$; (c) SIR Variates for the Population Version; (d) SIR's View with Equal Contour Boundary, $(\pi_y = 1/3, u_i \sim f(u) = 3u, u \in [0, 1]$

5. Other SIR type methods for dimension reduction

SIR may only recover part of the e.d.r. space if the dimension of the hyperplane spanned by the group means $E(\mathbf{x}|y)$ is less than the dimension of the e.d.r. space d. When this happens, other SIR type methods can help find more e.d.r. directions that cannot be found by using CRIMCOORDS.

5.1. SIR-II

In addition to $E(\mathbf{x}|y)$, PHD (Li 1992) uses second moment of $\mathbf{x}$ for dimension reduction. In our context, it seems more natural to use SIR-II (Li 1991) which explores the variation in the group covariance matrices. Let $\Sigma_a = E[Cov(\mathbf{x}|Y)]$ be the average of the group covariance matrices. Define

$$\Sigma_{II} = E\{[Cov(\mathbf{x}|Y) - \Sigma_a]\Sigma_{\mathbf{x}}^{-1}[Cov(\mathbf{x}|Y) - \Sigma_a]\}. \tag{5.1}$$

Then the eigenvalue decomposition for SIR-II is

$$\begin{gathered}\Sigma_{II}c_i = \gamma_i \Sigma_{\mathbf{x}} c_i \\ \gamma_1 \geq \cdots \geq \gamma_p.\end{gathered}$$

The insertion of the matrix $\Sigma_{\mathbf{x}}^{-1}$ in the construction of Σ_{II} is to assure the affine invariance of the SIR-II procedure. SIR-II is similar to SAVE (Cook and Weisberg (1991)).

Compared with SIR, a condition stronger than (2.11) is required for SIR-II to find e.d.r. directions: for any variable $\mathbf{a}'\mathbf{x}$,

$$cov(\mathbf{a}'\mathbf{x}, B'\mathbf{x}) = 0, \text{ implies that } \mathbf{a}'\mathbf{x} \text{ is independent of } B'\mathbf{x}. \tag{5.2}$$

The variance-covariance matrix of $(B'\mathbf{x}, \mathbf{a}'\mathbf{x})$ for each group $Y = y$ takes a diagonal partition because $cov[(B'\mathbf{x}, \mathbf{a}'\mathbf{x})|Y = y] = 0$. The first diagonal submatrix $cov(B'\mathbf{x}|Y = y)$ depends on y, but the second one does not: $cov(\mathbf{a}'\mathbf{x}|Y = y) = cov(\mathbf{a}'\mathbf{x})$. This implies that the matrix $Cov(\mathbf{x}|Y) - \Sigma_a$ vanishes in all but the the first submatrix. The $\mathbf{a}$ must be in the eigenspace with zero eigenvalue. Now it is clear that like SIR, SIR-II can find e.d.r. directions.

Theorem 5.1. *Under dimension reduction framework, (2.10) (or (2.1)), if (5.2) holds, then there are at most d nonzero eigenvalues in (5.1) and the corresponding eigenvectors are in the e.d.r. space.*

However, under the weaker condition (2.11), we can only conclude that some of the eigenvectors with nonzero eigenvalues might be in the e.d.r. space. If none of them are in the e.d.r. space, then the e.d.r. space must be contained in the eigenspace with zero eigenvalue.

Example 5.1 Spherical Distribution Problem

This was considered in Loh and Vanichsetakul (1988). There are two classes and ten variables with the following distributions:

Group $Y = 1$: (1) $x_1, \cdots, x_d$ are distributed uniformly within a d-dimensional spherical slab centered at the origin, with inner radius r_1 and outer radius r_2; (2) $x_{d+1}, \cdots, x_{10}$ are independent $N(0,1)$.

Group $Y = 2$: $(x_1, \cdots, x_{10})$ is a 10-dimensional multivariate normal centered at the origin, with identity covariance matrix.

The last $10 - d$ variables are just noise. Because of the perfect symmetric pattern, SIR fails to find the *e.d.r.* directions, but SIRII does a good job. For $d = 2$, $r_1 = 3.5, r_2 = 4.0$ and $n_1 = n_2 = 200$ cases, best view of x_2 against x_1 shows that the first class almost completely surrounds the second; Figure 5.1(a). Figure 5.1(b) is the SIRII view of the first two directions being found, which also illustrates the equal-contour line of the two densities which can be used as the boundary classifier for classification of future observations. The eigenvalues for this example are (0.613, 0.526, 0.063, 0.031, $\cdots$).

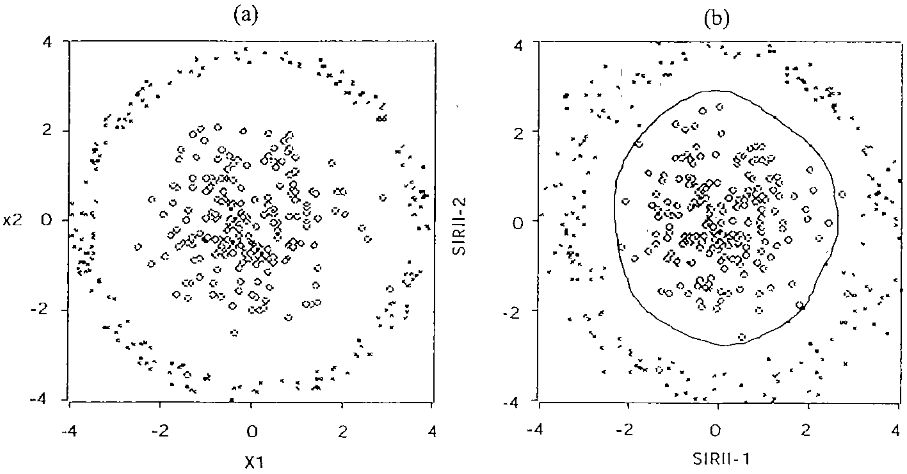

Figure 5.1: Spherical Distribution Problem: (a) Best View (b) SIRII's View with Equal-Contour Boundary

5.2. Double slicing

We begin with the binary case K=2. SIR-I can only find at most one e.d.r. direction. The rest of them can rely on the 2nd moment based methods to recover. But as the following example shows, this may not be enough.

Example 5.2 Tai Chi. Consider Figure 5.2a, the well-known Tai Chi figure in the Asian culture. The regions are in black and in white called Ying and Yang respectively. The concepts of Yin and Yang and the Five Agents provide the intellectual framework for much of ancient Chinese scientific development especially in fields like biology and medicine (Ebrey 1993).

The basic structure of Tai Chi is formed by drawing one large circle, two medium half circles and two small circles. The two small Yin and Yang circles located at the centers of the Yang and Yin half circles which are tangent to each other and are also to the large circle.

We set up the model as follows:

(1). Let x_1 and x_2 be the coordinates of a random point within the large circle. We then assign the class label $Y = 1$ if the point falls in the Yin region and assign $Y = 2$, otherwise.

(2). $x_3, \cdots, x_p$ are independent $N(0,\, 1)$.

The classification problem is to predict Y from $(x_1, \cdots, x_p)$.

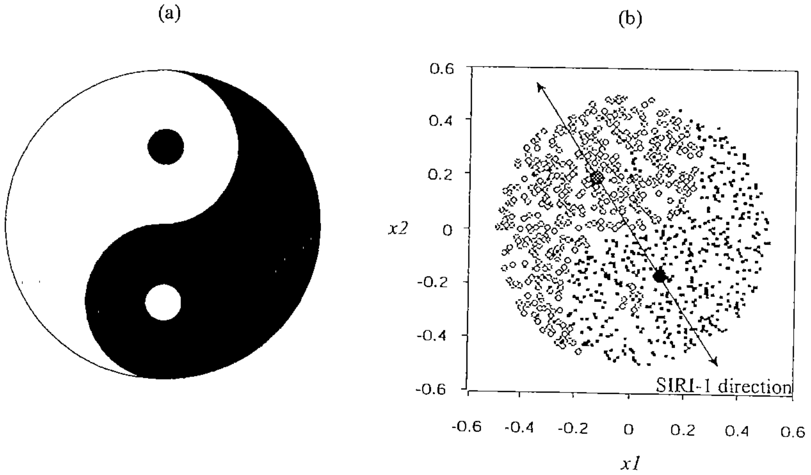

Figure 5.2: Tai Chi Example: (a) The Tai Chi Model with Yin and Yang Classes (b) Simulation of Tai Chi Model with 1000 Observations and the SIRI direction.

For this model, SIR-I can only find a single e.d.r. direction. This is the direction which passes through the mass centers of Yin and Yang regions (Figure 5.2b). However SIR-II can not identify any e.d.r. direction. This is because the Yin and the Yang regions are anti-symmetry to each other, implying that covariance matrix of (x_1, x_2) for Y=1 is the same as that for Y=2. One simple way to find the second direction necessary for completing the e.d.r. space is to slice the joint space of the direction identified by SIR-I and the class label Y.

In general, suppose that an e.d.r. direction b_0 is already obtained. We can take $\Sigma_\eta = cov(E(\mathrm{x}|b_0'\mathrm{x}, Y))$ and conduct the eigenvalue decomposition (2.1). Under the same condition as SIR, the eigenvectors with nonzero eigenvalues can be shown to fall into the e.d.r. space.

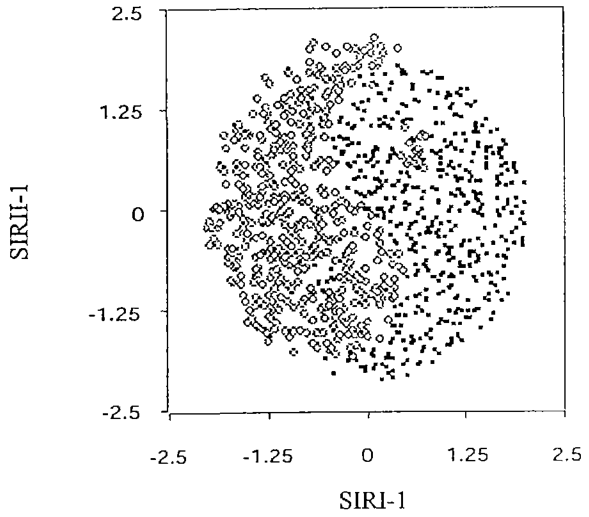

Figure 5.3: SIR's View for the Tai Chi problem with Double Slicing.Figure 5.4 Twist Problem: (a) Best View with SIRI direction (b) SIR's View with Double Slicing.

Example 5.2 (continued) For p = 6, we simulated 1000 i.i.d. cases of (x, Y) using the model specified. The result is in Figure 5.3 and Table 5.1. The double-sliced SIR-I and SIR-II have recovered the complete e.d.r space for the Tai Chi structure.

Table 5.1: Eigenvalues and eigenvectors of SIRI (a) and SIRII (b) for the Tai Chi problem with Double Slicing.

(a) SIRI

eigenvalues	(.987 .027 .019 .013 .006 .001)
1st eigenvector	(-2.299 3.269 -0.005 -0.011 -0.021 -0.034)
2nd eigenvector	(3.088 2.260 0.039 0.123 0.024 0.312)

(b) SIRII

eigenvalues	(.225 .133 .109 .081 .073 .001)
1st eigenvector	(-3.360 -2.287 -0.009 0.038 0.033 0.060)
2nd eigenvector	(0.200 -0.004 0.386 0.745 0.520 0.102)

Example 5.3 The Twist Problem. This example was originally introduced as a clustering problem by Koontz and Fukunaga (1972), see also Koontz et. al. (1975), and Fukunaga (1990). There are two C-shaped trigonometric curves with random normal noise tangled with each other in a two-dimensional space (Figure 5.6a). There are two classes, one for each curve:

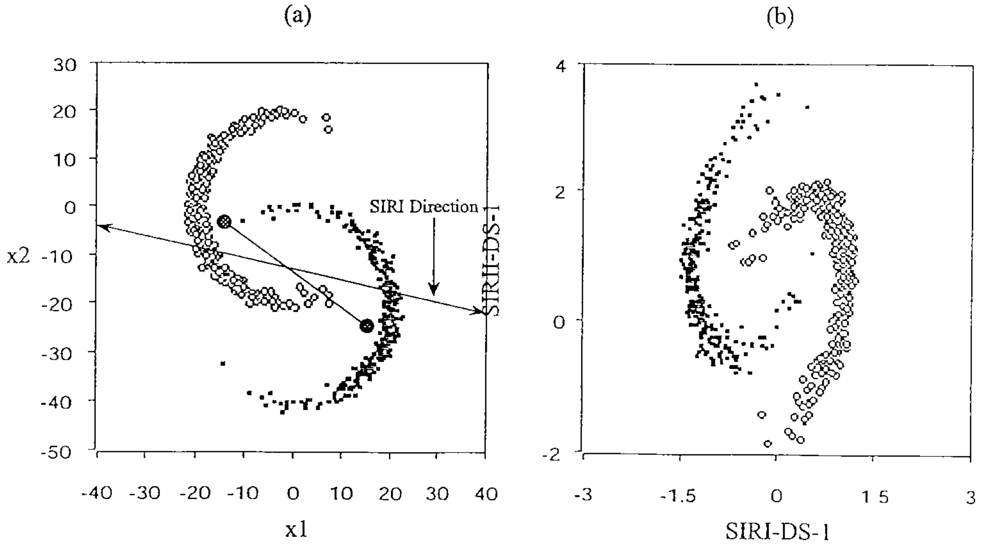

Figure 5.4: Twist Problem: (a) Best View with SIRI direction (b) SIR's View with Double Slicing.

Group $Y = 1$:

(1). $x_1 = 20\cos\theta + z_1$,
$x_2 = 20\sin\theta + z_2$,
where, z_1, z_2 are independent $N(0,1)$ and θ is $N(\pi, (0.25\pi)^2)$.

(2). $x_3, \cdots, x_p$ are independent $N(0,1)$.

Group $Y = 2$:

(1). $x_1 = 20\cos\theta + z_1$,
$x_2 = 20\sin\theta - 20 + z_2$,
where, z_1, z_2 are independent $N(0,1)$ and θ is $N(0, (0.25\pi)^2)$.

(2). $x_3, \cdots, x_p$ are independent $N(0,1)$.

For $p = 10$, we generate 300 cases from each class. Since the first two variables are correlated through the structure of θ, SIR-I direction will not pass through the mass centers of these two curves, see Figure 5.4a. In this case, again SIR-II fails to identify the second e.d.r. direction necessary for obtaining the complete e.d.r. space. But with double-slicing, we can find the entire e.d.r. space. The results are shown in Figure 5.6b and Table 5.3.

Table 5.2: Eigenvalues and eigenvectors of SIRI (a) and SIRII (b) for the Twist problem with Double Slicing.

(a) SIRI

eigenvalues	(.973 .129 .041 .018 .008 .001 .000 .000 .000 .000)
1st eigenvector	(-0.861 0.200 0.034 -0.003 0.049 0.007 0.037 0.014 -0.038 0.023)
2nd eigenvector	(0.158 0.089 0.630 0.115 0.379 0.113 0.480 0.055 -0.365 0.283)

(b) SIRII

eigenvalues	(.694 .130 .118 .095 .089 .079 .071 .049 .041 .004)
1st eigenvector	(-0.926 -1.219 -0.108 0.107 0.010 -0.013 -0.161 -0.001 -0.055 0.024)
2nd eigenvector	(-0.051 -0.211 0.386 0.310 -0.300 0.209 0.630 0.170 -0.089 0.416)

Example 5.4 Sonar data. This data set can be found in Gorman and Sejnowski (1988). Sonar signals bounced off a metal cylinder (class 1) or off a roughly cylinder rock (class 2) are recorded in 60 channels. The training set consists of 111 cases from class 1 and 97 cases from class 2. Thus the raw data consist of $p = 60$ predictors with $n_1 = 111, n_2 = 97$. The direct application of LDA or any generalization to 60 predictors with a sample of only 208 training

cases is questionable. This is because of the instability in estimating the covariance matrices (see Appendix C). To reduce dimension, feature extraction is often considered in the engineering literature.

We view the signal recording for each case as one curve $s(f_i), i = 1, \cdots, 60$, where f_i's denote the given channel frequencies. One way of feature extraction can be proceeded as follows. First, we find a small number of basis functions so that each curve can be represented well as a linear combination of the basis functions. According to a scheme which we describe in detail in Appendix A, four basis functions denoted by $\phi_1(f_i), \cdots, \phi_4(f_i)$ are selected. Suppose each curve is fitted by least squares:

$$s(f_i) = \alpha + \beta_1\phi_1(f_i) + \cdots + \beta_4\phi_4(f_i) + \epsilon_i, i = 1, \cdots, 60.$$

Then we can extract 5 feature variables, $x_1 = \alpha, x_2 = \beta_1, \cdots, x_5 = \beta_4$,from the original data. Since some curves are fitted better than others, we would like to include a sixth feature variable x_6 which is defined by $\log(r^2/(1-r^2))$, where r^2 is the R-squared value from the least squares fit.

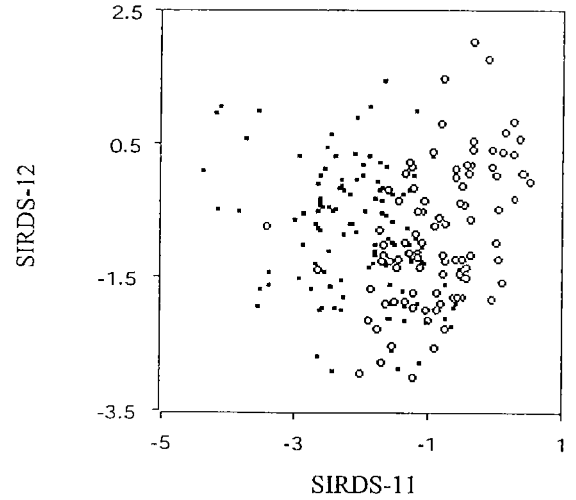

Figure 5.5: Plot of SIRDS-12 against SIRDS-11 for the Sonar Data with Six Extracted Feature Variables.

SIR is then applied to this set of six feature variables. The first direction found by SIR and the class label Y are used as the directions for running double-sliced SIR (SIR_{ds}). Two directions are found by SIRDS, which we denote as

SIRDS-11 and SIRDS-12 (Figure 5.5). The eigenvalues and first two eigenvectors for SIRDS are displayed in Table 5.3. We observe that the correlation coefficients between x_6 and SIRDS-11 with SIRDS-12 are -0.07 and 0.6826 respectively. Thus although SIR (or equivalently LDA) does not use x_6, the information contained in x_6 is used in SIRDS.

Table 5.3: The first two eigenvectors and eigenvalues of SIRDS for six base functions, Sonar data.

first vector	(-0.51 -21.61 0.00 -1.28 -0.35 -1.05)
second vector	(1.64 -3.07 -3.52 -3.60 -3.78 -3.60)
eigenvalues	(0.91 0.19 0.10 0.08 0.05 0.03)

After reducing to the two SIRDS variates, a k-nearest-neighbor classifier is applied. For k=1,3, ...,15, the resubstitution error rates are (24.52%, 22.60%, 23.56%, 22.12%, 19.23%, 20.19%, 23.56%, 24.04%) respectively with a minimum resubstitution error-rate of 19.23% at k=9.

6. Conclusion

LDA is a popular method for classification. This article re-investigates its theoretical property from the dimension reduction point of view. The canonical variates are found to be the same as the SIR variates except for the scaling. We examine in detail the assumption underlying SIR and apply them to the classification problems. This helps justify the use of CRIMCOORDS for informative graphical display of separation patterns between different classes. However the theory of SIR does not justify the use of linear rules. We illustrate that nonparametric density estimation following dimension reduction can be more informative then LDA. As in known, SIR may not be able to find "all" e.d.r. directions. We investigate two types of generalizations for finding more e.d.r directions. One of them is the second moment based method. This method explores differences between group covariance. Compared to SIR, one drawback of this method is the uncertainty introduced by covariance estimation. Another method, double-slicing, is less sensitive to covariance estimation. These methods extend the power of LDA and can be used to reveal more complicated data pattern.

Appendices: More on Sonar data.

A. Description of the procedure in choosing the basis functions. The first two basis functions $\phi_1(f_i)$, $\phi_2(f_i)$ are taken as the average of all curves $s(f_i)$ from the first class and the second class, respectively (Figure A.1). Each curve is

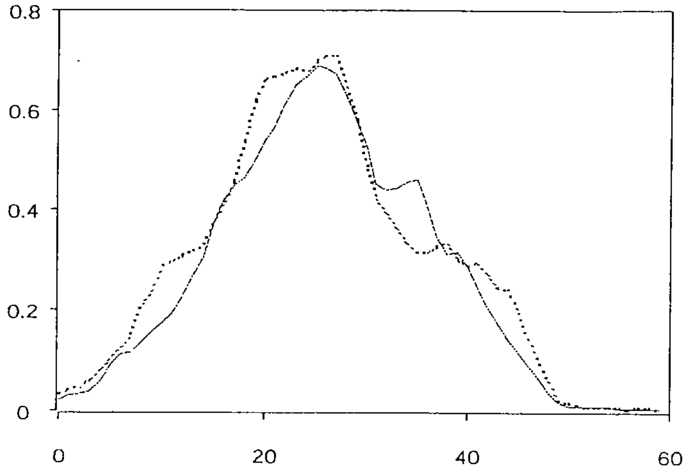

Figure A.1: First two basis functions, $\phi_1(f_i)$, and $\phi_2(f_i), i = 1, \cdots, 60$. The solid (red) and dashed (blue) curves represent the mean signals of the two groups of metal ($y = 1$) and rock ($y = 2$).

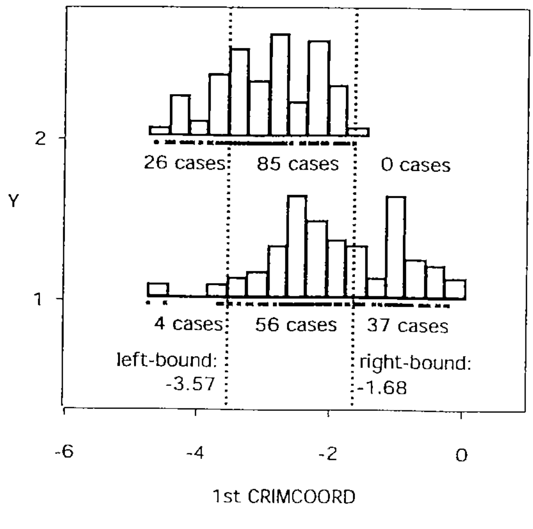

Figure A.2: Fisher's linear discriminant analysis for the three new predictors with 141ambiguous signals.

tentatively fitted with these two basis functions. Then an LDA is conducted on the three predictors, α, β_1, β_2. Figure A.2 of first CRIMCOORD shows a good portions of cases in the middle part cannot be distinguished well. This portion is extracted out, which has 141 cases. Our third and fourth basis functions $\phi_3(f_i)$ and $\phi_4(f_i)$ are just the average of all curves in this portion that came from class 1 and class 2 respectively (Figure A.3).

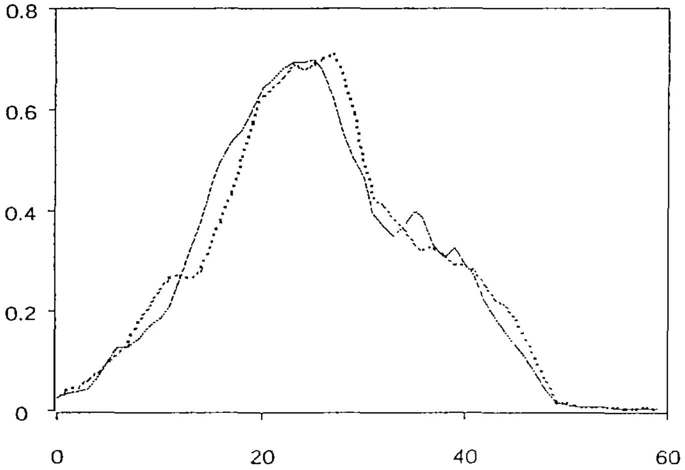

Figure A.3: Additional Basis Functions, $\phi_3(f_i)$, and $\phi_4(f_i), i = 1, \cdots, 60$. The solid (red) and dashed (blue) curves represent the mean signals of the two groups of metal ($y = 1$) and rock ($y = 2$) for the 141 ambiguous cases.

B. Stability. To see how stable the proposed classification procedure by SIRDS is, we proceed with the following simulations.

Each time we split all 208 cases into a training set and a test set with probabilities equal to 0.75 and 0.25 respectively. From a training set we first identify the 4 basis functions using the same procedure as described in Appendix A. Then we go through the same curve fitting step again and find two SIRDS directions. Signals in the test set are then projected to the obtained SIRDS directions. 1000 simulations with k=1,3, ... ,15 are carried out, the result, Figure B.1.c. The lowest average error rate of 22.68% for test set is reached at k=15 with an average standard deviation of 0.056. For comparison, the same simulation data is also used to carry out the LDA analysis and k-NN analysis for the original 60 variables. The average error for test set for LDA is 26.63% (Figure B.1.a), much higher than the k-NN results for the 6 base functions. The k-NN analysis for the original 60 variables are also consistently worse than that of the 6 base functions one, Figure B.1.b.

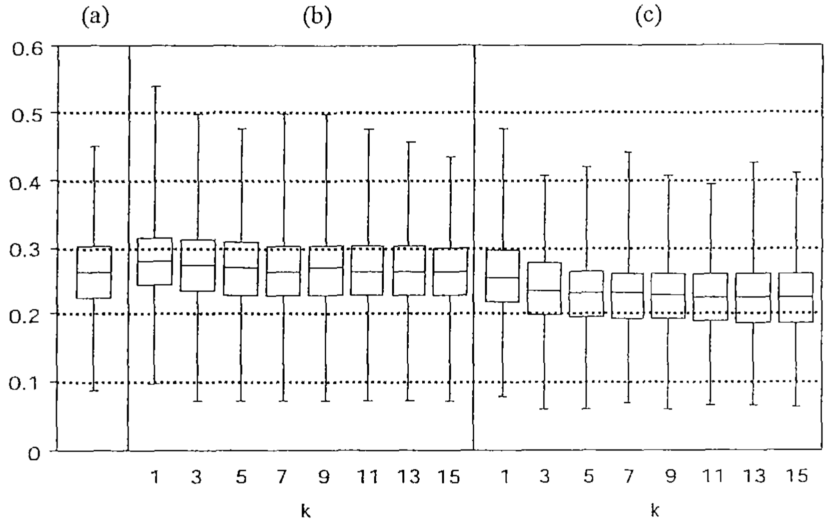

Figure B.1 Box-Plots for Simulation Result with 1000 runs: (a) LDA with 60 original variables; (b) k-nearest-neighbor classifier with 60 original variables; (c) k-nearest-neighbor classifier with 6 basis functions.

C. LDA with 60 predictors. The resubstitution error rate of Fisher's linear discriminant analysis with the original sonar data is 9.615%, which corresponds to 20 misclassified signals (12 metal signals and 8 rock signals each), Figure C.1. We suspect that this 9.615% resubstitution error rate is too low to be true. Since there are only 208 subjects in total with 60 variables, the estimation of the covariance matrix will be unstable, which may create a problem of overfitting. We carry out the following leave-one-out procedure to verify this suspicion. Each time we use one of the 208 signals as a test signal and use the other 207 signal to find the Fisher's linear discriminant function for predicting the class label of that selected test signal. Among all 208 runs, 51 signals are misclassified which corresponds to a leave-one-out error rate of 24.52%. This leave-one-out error rate of 24.52% is much higher than the resubstitution error rate of 9.615% by the single run linear discriminant analysis for the full 208 cases.

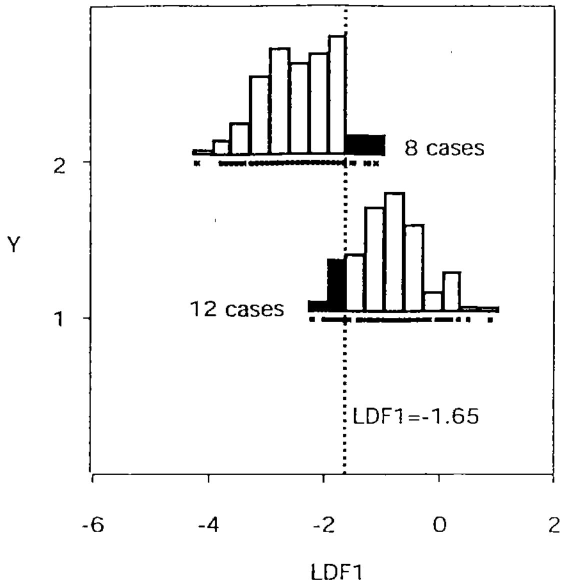

Figure C.1 Fisher's linear discriminant analysis for the full sonar data set with 20 misclassified signals. The upper and lower histograms represent the distributions of observations projected onto the Fisher's linear discriminant function (LDF) of the original 60 variables from the metal (y=1) and rock (y=2) groups respectively. The dashed line is the cutting point for prediction from LDF. The black bars represent the cases that are misclassified by LDF.

References

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

Brillinger, D. R. (1991), Comments on "Sliced Inverse Regression for Dimension Reduction", by K. C. Li, *J. Amer. Stat. Assoc.*, **86**, 333-333.

Chen, C. H., and Li, K. C. (1998), "A three-way subclassification approach to multiple-class discriminant analysis", *Academia Sinica, Ser.*

Cook, R. D. (1994), "On the Interpretation of Regression Plots," *J. Amer. Stat. Assoc.*, **89**, 177-189.

Cook, R. D., and Nachtsheim, J. C. (1994), "Reweighting to Achieve Elliptically Contoured Covariates in Regression," *J. Amer. Stat. Assoc.*, **89**, 592-599.

Cook, R. D., and Weisberg, S. (1991), Comments on "Sliced Inverse Regression for Dimension Reduction", by K. C. Li. *J. Amer. Stat. Assoc.*, **86**, 328-333.

Ebrey, P. (1993), *Chinese Civilization : A Sourcebook*, (2nd ed.), New York: Free Press, 77-79.

Fisher, R. A. (1936), "The Use of Multiple Measurements in Taxonomic Problems," *Ann. Eugen.*, **7**, 179-188.

Fisher, R. A. (1938), "The Statistical Utilization of Multiple Measurements," *Ann. Eugen.*, **8**, 376-386.

Friedman, J. H. (1977), "A Recursive Partitioning Decision Rule for Nonparametric Classification," *IEEE Transactions on Computers*, **26**, 404-408.

Fukunaga, Keinosuke (1990), *Introduction to Statistical Pattern Recognition.*

Gorman, R. P. and Sejnowski, T. J. (1988), "Analysis of hidden units in a layered network trained to classify sonar targets." *Neural Networks*, **1**, 75-89.

Gnanadesikan, R. (1977), *Methods for statistical data analysis of multivariate observations*, New York: John Wiley & Sons.

Hall, P., and Li, K. C. (1993), "On Almost Linearity of Low Dimensional Projection from High Dimensional Data," *Ann. Stat.*, **21**, 867-889.

Hsing, T and Carroll, R. J. (1992), "An Asymptotic Theory for Sliced Inverse Regression," *Ann. Stat.*, **20**, 1040-1061.

Koontz, W. and Fukunaga, K. (1972), "A Nonparametric Valley-Seeking Technique for Cluster Analysis," *IEEE Transactions on Computers*, **21**, 171-178.

Koontz, W., Narendra, P. and Fukunaga, K. (1975), "A Graphic-Theoretic Approach to Nonparametric Cluster Analysis," *IEEE Transactions on Computers*, **25**, 936-944.

Li, K. C. (1991), "Sliced Inverse Regression for Dimension Reduction," (with discussion), *J. Amer. Stat. Assoc.*, **86**, 316-342.

Li, K. C. (1992a), "Uncertainty Analysis for Mathematical Models with SIR", in *Probability and Statistics*, eds. Z. P. Jiang, S. H. Yan, P. Cheng, and R. Wu, Singapore: World Scientific, pp. 138-162.

Li, K. C. (1992b), "On Principal Hessian Directions for Data Visualization and Dimension Reduction: Another Application of Stein's Lemma," *J. Amer. Stat. Assoc.*, **87**, 1025-1039.

Loh, W. Y. and Vanichsetakul, N. (1988), "Tree-Structured Classification via Generalized Discriminant Analysis," (with discussion), *J. Amer. Stat. Assoc.*, **83**, 715-728.

Rao, C. R. (1952), *Advanced Statistical Methods in Biometric Research*, New York: John Wiley & Sons.

Schott, J. R. (1994), "Determining the Dimensionality in Sliced Inverse Regression," *J. Amer. Stat. Assoc.*, **89**, 141-148.

Van Ness, J. W. and Simpson, C. (1976), "On the Effects of Dimension in Discriminant Analysis," *Technometrics*, **18**, 175-187.

Zhu, L. X., and Fang, K. T. (1996), "Asymptotics for Kernel Estimate of Sliced Inverse Regression," *Ann. Stat.* **24**, 1053-1068.

Zhu, L. X., and Ng. (1995), "Asymptotics of Sliced Inverse Regression," *Statistica Sinica*, **5**, 727-736.