Chapter 14

Generalizing Fisher's linear discriminant analysis via the SIR approach

This chapter is a minor modification of Chen and Li(1998).

Despite of the rich literature in discriminant analysis, this complicated subject remains much to be explored. In this chapter, we study the theoretical foundation that supports Fisher's linear discriminant analysis (LDA) by setting up the classification problem under the dimension reduction model (1.1) of chapter 1. Through the connection between SIR and LDA, our theory helps identify sources of strength and weakness in using CRIMCO-ORDS(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.

14.1 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).

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 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 basic normality 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 commonly-used 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.

This chapter 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 **a** used in the canonical variates. Via this connection, the theory of SIR is 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 more effective than LDA 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. 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 to help find more e.d.r. directions that cannot be found by SIR. 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 mentioned. 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.

14.2 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.

14.2.1 Connection.

Recall the dimension reduction model (1.1) from Chapter 1. For ease of presentation, let's rewrite it here:

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

Here Y is the response variable, and g is an unknown function with $(d+1)$ arguments. Notice that we have changed the notation a little bit : d is used to replace K as the dimension of the e.d.r. space. This change is because K is reserved for the number of classes.

Recall the population version of SIR from Chapter 2 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 to 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.

The examples and discussion in earlier chapters focuse 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. 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) is 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 :

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

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 support for using 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 to consider a different situation where CRIMCOORDS can serve as an effective way of conveying the importance informance in the predictors.

By relating the cannocial 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 can be justified under much weaker conditions. We shall discuss these conditions next.

14.2.2 Condition (2.1).

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 $B = (\beta_1, \cdots, \beta_k)$. (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.

14.2.3 Condition on the predictor distribution.

Recall that in addition to (2.1) (or equivalently (2.9) for classification problems), SIR requires another condition on the distribution of $\mathbf{x}$: **(L.D.C)**: 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.

As we have known before, 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. As we have seen in Chapter 8, this predictor distribution condition is not too restrictive.

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.

14.3 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 $p_j = n_j/n$ and 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* $\mathbf{a}$ *with respect to* $\Sigma_{\mathbf{x}}$*, they must fall into the column space of* B*. The theorem is now proved.*

14.4 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 is only 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 vector space parallel to the three-dimensional hyperplane spanned by $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3$ is the e.d.r. space. This can be seen by verifying (2.9).

Now generate 200 cases from each group. Then SIR is applied. The eigenvalues are (0.651, 0.546, 0, $\cdots$). Kernel estimation is applied. Figures 4.1(a)-(b) show the Bayes rules with $\pi_y = 1/3$ and $\pi_y = y/6$ respectively. 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.

14.5 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.

14.5.1 SIR-II.

In our context, SIR-II 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

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

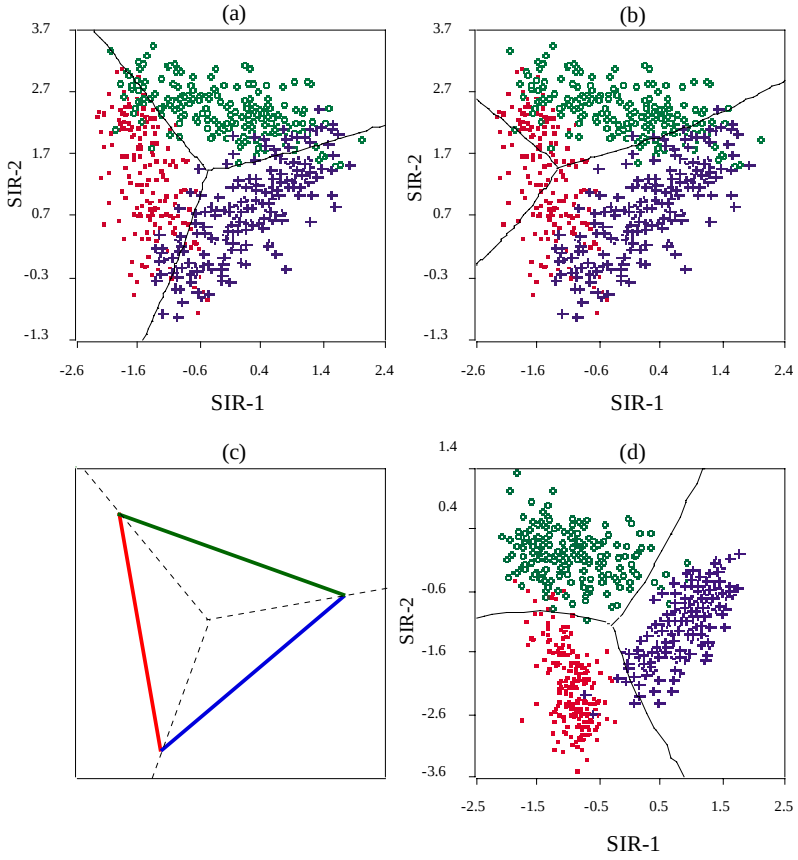

Figure 4.1: Wave Recognition Problem:
(a). SIR's View with Equal Contour Boundary, $\pi_y = \frac{1}{3}$;
(b). SIR's View with Equal Contour Boundary, $\pi_y = \frac{y}{6}$;
(c). SIR Variates for the Population Version;
(d). SIR's View with Equal Contour Boundary, ($\pi_y = \frac{1}{3}, u_i - f(u) =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

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.

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

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

Thus the covariance of $(B'\mathbf{x}, \mathbf{a}'\mathbf{x})$ for each group $Y = y$ takes a diagonal partition: $cov[(B'\mathbf{x}, \mathbf{a}'\mathbf{x})|Y = y] = 0$. The first diagonal $cov(B'\mathbf{x}|Y = y)$ depends on y, but the second one does not: $cov(\mathbf{a}'\mathbf{x}|Y = y) = cov(\mathbf{a}'\mathbf{x})$. This implies that $\Sigma_{II}\mathbf{a} = 0$. 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.