Improved Linear and Integer Programming Models For Discriminate Analysis

Article By:
Fred Glover

An Excerpt from:

Creative and Innovative Approaches
To The Science of Management
Edited by Yuji Ijiri
Quarum Books. Westport, Connecticut
1993

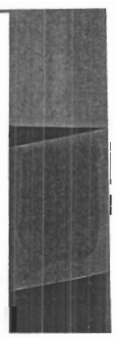
16

Improved Linear and Integer Programming Models for Discriminant Analysis

Fred Glover

There is a growing recognition that a variety of classical statistical problems can be approached advantageously using tools from the field of optimization. Reexamination of these problems and their underlying model assumptions can sometimes lead to refreshing new perspectives and alternative lines of attack. Discriminant analysis is high on the list of problems of this type and has been drawing increased attention recently because it straddles the areas of management science and artificial intelligence as well as statistics. Management science applications of discriminant analysis include decisions to make or buy, lend or invest, hire or reject (see Charnes, Cooper, and Rhodes, 1981; Kazmier, 1967; Spurr and Bonini 1976). Artificial intelligence applications involve the challenging realm of pattern recognition, including problems of signal differentiation, diagnostic classifications, code signatures, and data types (see Bobrowski, 1986; Kazmier, 1967; Tou and Gonzalez, 1974; and Watanabe, 1969).

An effort to wed statistical discrimination with optimization has come about through proposals to capture the goals of discriminant analysis in a collection of linear programming (LP) formulations (see Freed and Glover, 1981 and 1987; and Glover, Keene, and Duea, 1988). The objectives of initial forms of these models included minimizing the maximum deviation and the sum of deviations of misclassified points from a reference hyperplane, together with weighted variants of these objectives. Although the more advanced earlier variants and their recent derivatives have gone largely



unexplored (a condition that deserves to be remedied), empirical testing of the simpler variants has disclosed the "minimum sum of deviations" model to be competitive in effectiveness with the classical approach of Fisher (see Kazmier, 1967 and Markowski and Markowski, 1985). This comparative testing was carried out in contexts determined by the limited goals and assumptions of classical discriminant analysis and did not examine settings that could be advantageously exploited by the more flexible objectives of the LP discriminant approaches. Moreover, no use was made of LP postoptimization to re-weight borderline misclassified points to obtain refined solutions, one of the strategic options of the LP approaches proposed with their earliest formulations. Consequently, the effective performance of the LP discriminant analysis models under these circumstances gave encouraging evidence of their potential value in wider applications.

At the same time, however, empirical tests also disclosed that the LP formulations gave counterintuitive and even anomalous results. Follow-up examination of specially anomalous results demonstrated that these formulations are attended by certain subtleties not found in other areas to which linear programming is commonly applied (see Bajgier and Hill, 1982; Freed and Glover, 1987; and Markowski and Markowski, 1985).

Analysis has indicated that the anomalous behavior of the LP formulations stems from the implicit use of normalizations in order to avoid "null solutions" that assigned zero weight to all data elements. Several normalizations have been identified (see Freed and Glover, 1987 and Glover, Keene, and Duea, 1988) in an attempt to overcome this difficulty. The most recent of these has been demonstrated to exhibit desirable invariance properties lacking in its predecessors and has produced encouraging experimental outcomes, yielding solutions generally better than those obtained by earlier studies (see Glover, Keene, and Duea, 1988).

In spite of these advances, however, the full power of the LP models for discriminant analysis has not been achieved because the best normalization proposed to date distorts the solutions in a manner not previously anticipated. The consequences of this distortion not only inhibit the quality of "first pass" solutions obtained by the LP formulations, but also can confound the logical basis of obtaining more refined solutions by differential weighting of deviations in the objective functions and LP postoptimization.

The purpose of this chapter is to remedy these defects and to demonstrate some of the consequences for improved modeling capabilities that result. We introduce a new normalization that eliminates the previous distortions in the LP models and that has attractive properties enabling it to obtain demonstrably superior solutions.

The new normalization further allows a generalization to integer conditions and causes the integer programming problem of minimizing the

number of misclassified points to have as its continuous relaxation the LP problem of minimizing the cumulative deviations of misclassified points, permitting the latter to serve as an approximation to the former. The value of this approximation is reinforced by the demonstration that the new normalization also endows the LP formulation with an integer local optimality property, which yields a "balanced" number of misclassified points. These links between continuous and discrete solutions, and the lack of distortion that attended the most effective previous normalization, give new scope to the LP models. Finally, we show that the ability to place any desired relative emphasis on classifying particular points correctly leads to a conditionally staged application of the model, called the successive goal method, for achieving progressively more refined discrimination for both two-group and multigroup analysis.

A HYBRID LP DISCRIMINANT MODEL

We take as our starting point the hybrid LP model of Jurs (1986), which integrates features of the previous LP discriminant formulations (Freed and Glover, 1981 and 1987). Attention will initially be restricted to the two-group discriminant problem, which constitutes the main focus of our development.

We represent each data point by a row vector A_i, where membership in Group 1 or Group 2 is indicated by $i \in G_1$ or $i \in G_2$, respectively. (Different points can have the same coordinates, and efficient adaptations for this are indicated below in "Model Manipulation and Simplifications.")

To discriminate the points of the two groups, we seek a weighting vector x and a scalar b, which may be interpreted as providing a hyperplane of the form $Ax = b$, where A takes the role of representing A_i for each *i*. The goal is to assure as nearly as possible that the points of Group 1 lie on one side of the hyperplane and the points of Group 2 lie on the other, which translates into the conditions that $A_ix < b$ for $i \in G_1$ and $A_ix > b$ for $i \in G_2$.

Refining this goal as in Glover, Keene, and Duea (1988), we introduce external and internal deviation variables, represented by the symbols α_i and β_i, which refer to the magnitudes by which the points lie outside or inside (and hence "violate" or "satisfy") their targeted half spaces. Upon introducing objective function coefficients h_i to discourage external deviations and k_i to encourage internal deviations and defining $G = G_1 \cup G_2$, we may express the LP model as follows:

Minimize $$h_o\alpha_o + \sum_{i \in G} h_i\alpha_i - k_o\beta_o - \sum_{i \in G} k_i\beta_i \quad (1)$$

subject to

$$A_i x - \alpha_o - \alpha_i + \beta_o + \beta_i = b \qquad i \in G_1 \quad (2)$$

$$A_i x + \alpha_o + \alpha_i - \beta_o - \beta_i = b \qquad i \in G_1 \quad (3)$$

$$\alpha_o, \beta_o \geq 0 \quad (4)$$

$$\alpha_i, \beta_i \geq 0 \qquad i \in G \quad (5)$$

x, b unrestricted in sign (6)

Many variations of this model framework are possible. For example, in the "ϵ version" of the model, the variable b that constitutes the boundary term for the hyperplane can be replaced by b - ϵ for Group 1 and by b + ϵ for Group 2, where ϵ is a selected positive constant, to pursue the goal of compelling elements of Group 1 and Group 2 to lie strictly inside the half space whose boundary is demarked by *b*. (Different values of ϵ may be chosen for different points. However, under the choice of a uniform value, the ϵ version is also equivalent to a "one-sided ϵ model" that replaces *b* by b + ϵ for Group 2 only, where the ϵ value in this case is twice as large as in the "two-sided" case.)

The objective function coefficients will generally be assumed to be non-negative, although it is possible to allow the coefficients of the variables to be negative. In this latter variation the hybrid model represents a generalized form of a standard goal programming model. We also stipulate that the objective function coefficients should satisfy $h_j \geq k_j$ for $i = 0$ and $i \in G$. Otherwise, it would be possible to take any feasible solution and increase the value of α_j and β_j (for $h_j < k_j$) an indefinite amount to obtain an unbounded optimum. More complete conditions for avoiding unbounded optimality, both necessary and sufficient, are identified subsequently.

From an interpretive standpoint, the α_o variable provides a component to weight the "maximum external deviation," while the β_o variable provides a component to weight the "minimal internal deviation." This interpretation is suggestive rather than exact, however, due to the incorporation of the individual point deviation variables, α_i and β_i, in the same equations as α_o

and β_o. The effects of these variables can be segregated more fully by introducing separate constraints of the form $A_i x - \alpha_o + \beta_o \leq b$ for $i \in G_1$, and $A_i x + \alpha_o - \beta_o \geq b$ for $i \in G_2$, at the expense of enlarging the model form. By deleting the α_o and β_o variables in equations (1) through (6) or, alternatively, by deleting the α_i variables and setting the k_i coefficients to zero, the foregoing model corresponds to one of the models first proposed in Freed and Glover (1981).

THE NORMALIZATION ISSUE

To understand the potential difficulties that underlie the preceding discriminant analysis formulation, it is useful to review in greater detail the history of its development and attempted application. In the form given, the model in fact is incomplete, for it must be amended in some fashion to avoid an optimal solution that yields the null weighting $x = 0$. If the two groups can be separated by a hyperplane (or "nearly" so) and if the k_i coefficients are positive, the null weighting will be automatically ruled out, but in this case the model must be amended to assure that it is bounded for optimality. Broadly speaking, the more challenging applications of discriminant analysis arise where the two groups significantly "overlap," and in these cases a solution yielding the null weighting $x = 0$ typically will be optimal if it is not somehow rendered infeasible.

The early implementations of LP formulations for discriminant analysis undertook to avoid the null weighting by the logical expedient of setting b to a nonzero constant. It was tacitly assumed that different choices of b would serve only to scale the solution (provided at least the proper sign was chosen), and the approximation to optimality in the special case where b ideally should be 0 still would be reasonably good.

However, experimental tests of different LP model variants soon disclosed that assigning b a constant value still permitted the null weighting to occur for certain data configurations. More generally the models responded with nonequivalent, and sometimes poor, solutions to different translations of the same underlying data, where each point A_i is replaced by the point $A_i + t$ for a common vector t (see Bajgier and Hill, 1982; and Markowski and Markowski, 1985).

These unexpected outcomes prompted the observation that setting b to a constant value could be viewed as a "model normalization," and it was soon discovered that other normalizations could be identified that affected the model behavior in different ways (see Freed and Glover, 1987). Let N denote the index set for components of the x vector. Then the first two proposals for

alternative normalizations to remedy the problems of setting b to a constant can be written in the following form:

$$b + \sum_{j \in N} x_j = \text{a constant}$$

$$\sum_{j \in N} x_j = \text{a constant}$$

Of these alternatives, the latter proved in Freed and Glover (1987) to yield solutions that were equivalent for different translations of the data, a property not shared by the other normalizations. This advantage was not enough to rescue the latter normalization from defects, however. First, to use the normalization, the LP formulation had to be solved for both signs of the constant term to assure that the right sign was selected. Second, the variables had to be either directly or indirectly bounded (in a sense, yielding an auxiliary normalization) to assure bounded optimality. Third, the normalization continued to produce nonequivalent solutions for different rotations (in contrast to translations) of the problem data, where each point A_i is replaced by the point A_iR, and R is a rotation matrix.

The most recent attempt to settle the normalization issue occurred in Jurs (1986) with the "β normalization"

$$\beta_0 + \sum_{i \in G} \beta_i = 1$$

The need to allow for different signs of the constant term was eliminated with this normalization. More significantly, it was proved that the normalization succeeded in yielding equivalent solutions for both translations and rotations of the problem data. Experimentation further shows that the normalization provided solutions uniformly as good or better than solutions obtained with previous normalizations for the problems examined. In spite of these advances, however, this latest normalization likewise suffers undesirable limitations, which continue to distort the solutions obtained by the LP formulations.

In the following sections we illustrate the nature of the distortion inherent in the ß normalization and then show that it is compounded by a related defect that limits the generality and flexibility of the LP model when this

normalization is used. We then provide a new normalization that is free of these limitations, while exhibiting the appropriate invariance properties for transformations of data. The attributes of this normalization are explored in results that establish additional features of the LP formulations not shared by alternative approaches. Finally, we amplify the implications of these results for obtaining discrimination approaches of increased power.

LIMITATIONS TO BE OVERCOME

The limitations of the β normalization will be illustrated in an example applicable to the standard discriminant analysis context as a means of clarifying the properties that need to be exhibited by an improved normalization. Consider the simple case where each point A_i has a single coordinate, and hence the weight vector x may be treated as a scalar variable. For illustrative purposes we will use the form of the hybrid model in which α_o and β_o are deleted. In addition, for further simplicity, we suppose all the k_i coefficients are 0.

The relevant data for the example are given in Table 16.1, indicating the coordinates and the penalties for being classified in the wrong group.

Table 16.1. The Coordinates and the Penalties for Being Classified in the Wrong Group

Group 1 Points		Group 2 Points	
Coordinates	Penalties	Coordinates	Penalties
$A_1 = 0$	$h_1 = 15$	$A_4 = -1$	$h_4 = 25$
$A_2 = 1$	$h_2 = 25$	$A_5 = 0$	$h_5 = 25$
$A_3 = -2$	$h_3 = 25$	$A_6 = 2$	$h_6 = 25$
		(all $k_i = 0$)	

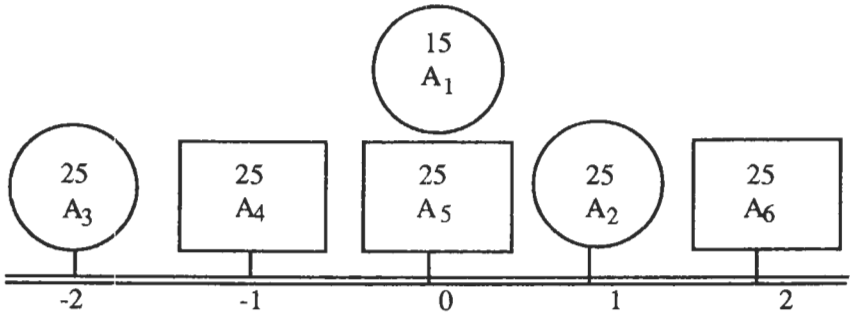
A graph of the points is shown in Figure 16.1, where Group 1 points are indicated by circles and Group 2 points are indicated by squares. The misclassification penalties are shown above each point.

The values from -2 to +2 on the line segment correspond to the values of b. It is easy to show that the best way to separate the Group 1 and Group 2

points on the line segment is to choose the value b = 0, where Group 1 points are counted as misclassified if they fall to the right of the selected value and Group 2 points are counted as misclassified if they fall to the left. Then, only A_2 and A_4 are misclassified, each with a deviation of 1 unit from the value b = 0, hence giving a total penalty cost of (1 x 25) + (1 x 25) = 50.

Without a normalization constraint, the LP model falls into the trap of finding a meaningless "optimal" solution, x = 0 and b = 0, which makes all external deviations 0 and hence also makes the total penalty cost 0. Consider the result of using the β normalization to overcome this limitation. We can choose any positive constant term for the right-hand side of this normalization, and specify the normalization to be $\Sigma\beta_i = 4$, since 4 is the sum of the internal deviations, β_i, in the case identified as best by graphical analysis. Indeed, we then obtain x = 1, b = 0 (with $\alpha_4 = \alpha_2 = 1$, $\beta_3 = \beta_6 = 2$, all other α_i and $\beta_i = 0$) as a feasible solution for the LP model, yielding a total penalty cost of 50, as before.

Figure 16.1. Group 1 and Group 2 Points with Misclassification Penalties



This solution turns out not to be optimal, however. Rather, the β normalization causes the inferior solution based on x = 1 and b = -1 to appear even better. From a graphical standpoint, the deviation variables with positive values for this solution are $\alpha_1 = 1$, $\alpha_2 = 2$, $\beta_3 = 1$, $\beta_5 = 1$, and $\beta_6 = 3$, which yield a total penalty cost of 65. However, the sum of the β_i variables equals 5, and to rescale the solution to satisfy the β normalization with a right-hand side of 4, the value of each variable must be multiplied by 4/5. The result is to multiply the total penalty cost of 65 likewise by 4/5, yielding a

penalty cost for the LP model of 42. This is better than the "best case" penalty cost of 50, causing the model to favor a less desirable solution.

This outcome is made more remarkable by noting that an earlier normalization, Σx_j = a constant (in this case chose 1 as the constant), will correctly identify the "best solution" as optimal. Yet for multidimensional problems this earlier normalization suffers from distortions not encountered by the β normalization, and empirical testing has found it generally to provide solutions that are not as good as those produced by the β normalization. Consequently, we are motivated to seek a new type of normalization that is more broadly effective and reliable.

As a step toward identifying additional properties this new normalization ideally should have, and pitfalls it should avoid, we next examine the behavior of the normalization in the context of an integer programming formulation.

The Integer Programming Case

An integer programming (IP) discriminant model for minimizing the *number* of misclassified points can be formulated as a simple variant of the LP model. We write the IP model as follows.

$$\text{Minimize} \quad \sum_{i \in G} z_i \tag{7}$$

subject to

$$A_i x - U z_i + \beta_i = b \qquad i \in G_1 \tag{8}$$

$$A_i x + U z_i - \beta_i = b \qquad i \in G_2 \tag{9}$$

$$\beta_i \geq 0 \qquad i \in G \tag{10}$$

$$z_i = 0 \text{ or } 1 \qquad i \in G \tag{11}$$

$$x, b \text{ unrestricted in sign} \tag{12}$$

The constant U is assumed to be chosen large enough that the inequality of $A_i x \leq b + U z_i$ will be redundant for $i \in G_1$ and the inequality of $A_i x \geq b$ -

Uz_i will be redundant for $i \in G_2$, when z_i is set equal to 1. The β_i variables may be interpreted as slack and surplus variables for these inequalities. More generally, the constant U can be replaced by different constants U_i for each i in G. Likewise a constant can be added to the right-hand side of equation (8) and subtracted from the right-hand side of equation (9) to seek a minimizing solution for strict group separation.

To apply the β normalization to the IP formulation, we need to know how large the U should be. From equations (8) and (9), we note the normalization can be expressed as

$$\sum_{i \in G_1} (b - A_i x + Uz_i) + \sum_{i \in G_2} (A_i x - b + Uz_i) = 1$$

Hence, in particular this yields:

$$U = [1 - \sum_{i \in G_1} (b - A_i x) - \sum_{i \in G_2} (A_i x - b)] / \sum_{i \in G} z_i$$

Thus, we see that the value of U depends intimately on the optimal values of the problem variables, which cannot be known in advance. If the IP formulation is applied to points A_i of the numerical example of Table 16.1 and Figure 16.1, the value of U must uniquely be selected to be 1/4 to permit the optimal integer solution to be found. This serious deficiency of the β normalization from an integer programming model standpoint identifies a further type of limitation an improved normalization should seek to overcome.

THE NEW NORMALIZATION

The normalization we propose is

$$(-n_2 \sum_{i \in G_1} A_i + n_1 \sum_{i \in G_2} A_i)x = 1 \qquad (N)$$

where n_1 and n_2 are respectively, the number of elements in G_1 and G_2, and the right-hand side of 1 is an arbitrary scaling choice for a positive constant.

(An alternative scaling that tends to yield x_j values closer to an average absolute value of 1 is to choose this constant to be $2n_1n_2$.) An equivalent form of this normalization occurs by adding n_2 times each equation of (1) and subtracting n_1 times each equation of (3) to yield the constraint

$$2n_1n_2\,(\beta_o\alpha_o) + n_2 \sum_{i\in G_1} (\beta_i - \alpha_i) + n_1 \sum_{i\in G_2} (\beta_i - \alpha_i) = 1 \qquad (N^*)$$

Expressing the normalization in the form (N) has certain advantages for analysis, while expressing it in the form (N*) is convenient for incorporation into the LP formulation [since the coefficients of the variables do not require calculation as in (N)]. It may be noted that the weights h_i and k_i in the objective function should not be chosen in proportion to the coefficients of corresponding variables in (N*), or else the normalization effectively constrains the objective function to equal a constant, and the minimization goal becomes superfluous. [If the k_i coefficients are proportional to corresponding coefficients of (N*), then a similar effect occurs in the case where it is possible to completely separate Group 1 and Group 2 points—where all α_i become 0.]

To understand the properties of the normalization given by (N) and (N*), let d_i denote the *net internal deviation* of the point A_i from the hyperplane generated by the discriminant model; that is

$$d_i = b - A_ix \quad \text{for } i \in G_1$$

$$d_i = A_ix - b \quad \text{for } i \in G_2.$$

Hence, d_i is positive (or zero) if A_i lies within its targeted half space and negative otherwise. [The "ϵ version" of the model for seeking strict separation replaces the quantity b by $b - \epsilon$, for a positive constant ϵ, in the definition of d_i. This results in increasing the constant term of the normalization (N) by the quantity $\epsilon(n_1 + n_2)$, while leaving the constant term of the normalization (N*) unchanged.]

Note that if Group 1 and Group 2 have the same number of points and are "separable" to any meaningful extent by a hyperplane, then the internal deviations should sum to a larger value than the external deviations, and, hence, the sum of all the d_i values should be positive. More broadly, if

Group 1 and Group 2 have a different number of points, then upon weighting the d_i values to give equal representations to the groups relative to their sizes (i.e., multiplying each d_i in Group 1 by n_2 and each d_i in Group 2 by n_1), a meaningful separation should yield a positive value for this weighted sum. We embody this observation in the following definition.

A hyperplane creates a *meaningful separation* of Group 1 and Group 2 if

$$n_2 \sum_{i \in G_1} d_i + n_1 \sum_{i \in G_2} d_i > 0.$$

On the basis of this definition we may at once state the following result.

Theorem 1

The normalization (N) is equivalent (under scaling) to requiring a meaningful separation and eliminates the null weighting $x = 0$ as a feasible solution.

Proof. First, (N) reduces to an inconsistent equation when $x = 0$ and hence renders the null solution infeasible. To see that (N) is equivalent to requiring a meaningful separation, expand the inequality that defines a meaningful separation by substituting the appropriate values for d_i, according to membership of i in G_1 or G_2, thereby obtaining

$$n_2 \sum_{i \in G_1} (b - A_i x) + n_1 \sum_{i \in G_2} (A_i x - b) > 0$$

Algebraic manipulation and reduction permit this inequity to be reexpressed in the form

$$-n_2 \sum_{i \in G_1} A_i x + n_1 \sum_{i \in G_2} A_i x > 0$$

whose left-hand side corresponds to the left-hand side of (N). Given any feasible solution to the LP formulation that satisfies this inequality, upon dividing the values of all variables in the solution by the positive left-hand-

side quantity, the result is again feasible for the LP problem and satisfies (N). Hence, allowing for scaling, the solutions are equivalent. Similarly, any feasible solution that satisfies (N) automatically satisfies the definition of a meaningful separation. This completes the proof.

Corollary

A meaningful separation exists if and only if there exists a hyperplane such that $n_2 \sum_{i \in G_1} d_i + n_1 \sum_{i \in G_2} d_i \neq 0$.

It also exists if and only if there exists some component A_{ij} of each point A_i, $i \in G$, such that $n_2 \sum_{i \in G_1} A_{ij} \neq n_1 \sum_{i \in G_2} A_{ij}$

Proof. The corollary is a direct consequence of Theorem 1 and the form of (N).

It may be noted by the proof of Theorem 1 that upon choosing non-negative scalars u_i such that $\sum_{i \in G_1} u_i = \sum_{i \in G_2} u_i > 0$, a normalization of the form

$$(-\sum_{i \in G_1} u_i A_i + \sum_{i \in G_2} u_i A_i)x = 1$$

will correspondingly eliminate the null solution and be consistent with a biased meaningful separation defined by the inequality

$$\sum_{i \in G_1} u_i d_i + \sum_{i \in G_2} u_i d_i > 0$$

Specifically, if there is reason to ensure that a weighted sum of internal deviations should exceed a correspondingly weighted sum of external deviations (as where particular points command more importance, and hence larger weights, than others), then such a biased normalization can be employed. We will not undertake to pursue the issue of these biased normalizations further, but simply note that our subsequent results can be readily adapted to treat them as well.

Useful additional insights into the nature of (N) and its consequences for the hybrid LP discriminant formulation are provided by examining the linear

programming dual of equations (1) through (6) with (N) attached. To create this dual, it is convenient first to rewrite the constraint equation (2) by multiplying it through by -1. Then, associating a variable v_i with the equations (2) and (3) for each $i \in G$ and a variable v_o with (N), we obtain the following result.

Dual Model Formulation

Maximize v_o

subject to

$$A_o v_o - \sum_{i \in G_1} A_i v_i + \sum_{i \in G_2} A_i v_i = 0$$

$$h_o \geq \sum_{i \in G} v_i \geq k_o$$

$$h_i \geq v_i \geq k_i \qquad i \in G$$

$$\sum_{i \in G_1} v_i - \sum_{i \in G_2} v_i = 0$$

where

$$A_o = -n_2 \sum_{i \in G_1} A_i + n_1 \sum_{i \in G_2} A_i$$

Our interest is analyzing this dual is to determine circumstances that provide a feasible dual solution and hence that assure that the LP discriminant formulation is bounded for optimality.

Necessary conditions for bounded optimality of the formulation (1) through (6) are immediately evident from the dual formulation, as are necessary conditions in order for certain variables of the LP discriminant

formulation to be nonzero at optimality. The following is established by reference to the quality theory of linear programming.

Necessary Conditions for Bounded Optimality

$h_i \geq k_i \qquad i \in G \text{ and } i = 0,$

$\sum_{i \in G} k_i \leq h_0$

Necessary conditions for Variables to be Nonzero

For β_0: $h_0 < \sum_{i \in G} h_i$

For β_0: $k_0 > \sum_{i \in G} k_i$

To avoid trivial solution values for dual variables, it is appropriate to stipulate $h_i > k_i$ for $i \in G$. In general, interpretation of the inequalities for bounded optimality in the context of the LP discriminant formulation suggests they reasonably may be required to be strict. It may be noted that $h_i > k_i$ implies that at most one of α_i and β_i will be positive, an outcome that also holds when $h_i = k_i$ in the case of extreme point solutions. (This is not true for the β normalization.)

We seek to go beyond the foregoing observations, however, by providing sufficient as well as necessary conditions for bounded optimality.

Theorem 2

The LP discriminant model in equations (1) - (6) with the normalization (N) is bounded for optimality whenever

$$\text{Min } (h_0/2, n_1h_1 : i \in G_1, n_2h_i : i \in G_2)$$

is at least as large as

$$\text{Max } (k_0/2,\ n_1k_1{:}i\in G_1,\ n_2k_i{:}i\in G_2)$$

Proof. Replace (N) by (N*) in the primal formulation, whereon the dual problem becomes

Maximize v_0

subject to

$$-\sum_{i\in G_1} A_iv_i \quad +\sum_{i\in G_2} A_iv_i = 0$$

$$h_0 \geq -2n_1n_2v_0 + \sum_{i\in G} v_i \geq k_0$$

$$h_i \geq -n_2v_0 + v_i \geq k_i \qquad i\in G_1$$

$$h_i \geq -n_1v_0 + v_i \geq k_i \qquad i\in G_2$$

Here v_0 is the same variable as in the preceding dual formulation, but the v_i variables, $i \in G$, are different. In this new dual formulation, we set $v_i = 0$ for all $i \in G$. The resulting partial solution satisfies the first problem constraint and leaves the remaining inequalities in the form of bounds on v_0. Expressing these as bounds on $-n_1n_2v_0$ in each case, and then comparing terms, yields the inequalities stated in the theorem. This completes the proof.

The sufficiency conditions of Theorem 2 are generally more restrictive than required to assure bounded optimality. When the theorem is applied to the model variant where α_0 and β_0 is deleted the corresponding term involving h_0 or k_0 is deleted from its statement. Where both α_0 or β_0 are deleted, and the two groups have the same number of elements, the conditions of the theorem simplify to $\text{Min}(h_i{:}i\in G) \geq \text{Max}(k_i{:}i\in G)$.

Theorem 2 has an additional attractive feature. Suppose that h_i and k_i values initially have been chosen subject only to the condition that all h_i (including h_0) are positive. If the inequality of Theorem 2 is not satisfied, let R be the ratio of the Max term to the Min term of the theorem. Then upon replacing each h_i by Rh_i in the objective (1), the condition of the theorem is satisfied. This modification of the coefficients of objective (1) leaves the relative magnitudes of the h_i coefficients, and also of the k_i coefficients,

of objective (1) to reflect any desired *relative emphasis* on the correct classification of particular points, and bounded optimality can be assured by a simple adjustment of the objective function coefficient that preserves this relative emphasis.

Our next goal is to show that the normalization (N) is stable across rotations and translations of problem data. For this, we employ a useful result from Glover, Keene, and Duea (1988). Consider the following pair of related problems:

I. Minimize g(y)

subject to

$$A_i x + \beta_i(y) = b \qquad i \in G \qquad (13)$$
$$y \in Y$$

II. Minimize g(y)

subject to

$$(A_i R + t)x + \beta_i(y) = b \quad i \in G \qquad (14)$$
$$y \in Y$$

The terms g(y) and $\beta_i(y)$ for $i \in G$ in these problems represent arbitrary functions of y. To connect these problems to the LP and IP discriminant formulations, the vectors A_i and the variables x and b may be construed the same as indicated previously. The vector of variables y may accordingly include all remaining variables of the LP and IP formulations, while the condition $y \in Y$ may summarize non-negativity and integer requirements.

It is important to note that $y \in Y$ can also incorporate the normalization constraint (N), using the observation in the proof of Theorem 2 that reexpresses this constraint in terms of the α and β variables [which similarly leads to an expression for (N) in terms of the z and β variables for the IP problem]. The objective function g(y) and the constraints function $\beta_i(y)$ of I and II may likewise encompass the associated linear functions of the LP and IP discriminant models as a special case.

The constraints that differentiate the two problems are constraints (13) and (14). The latter constraint set achieves the effect of transforming each point A_i by means of a rotation matrix R and translating the point by means of

point A_i by means of a rotation matrix R and translating the point by means of a row vector t. More generally, we assume that R is nonsingular and disregard the stipulation that the transpose of a rotation matrix is also its inverse. Then we may state the following result (stability theorem—see Glover, Keene, and Duea, 1988).

Stability Theorem

The optimum objective function values for Problems I and II are the same. Moreover, if Y' and Y" represent the optimal solution sets Y for Problems I and II, respectively, then Y' = Y".

Proof. We show more particularly that if the solution (y', x', b') is optimal for I, then $(y', R^{-1}x', b' + tR^{-1}x')$ is optimal for II, and if (y", x", b") is optimal for II, then (y", Rx", b" - tx") is optimal for I. By substituting and rearranging terms, it is clear that the solutions claimed to be optimal for problems I and II, given the assumed optimality of (y', x', b') and (y", x", b"), must respectively be feasible for these two problems. By feasibility for II we obtain $g(y') \geq g(y")$, and by feasibility for I we obtain $g(y") \geq g(y')$. Consequently $g(y') = g(y")$ and the stated conclusions are established.

By our observations linking the LP and IP discriminant formulations to Problems I and II, the Stability Theorem gives the desired result.

Theorem 3

The optimum objective function values and optimal values for the α and β deviation variables in the LP discriminant formulation and for the z and β variables in the IP formulation, are unchanged for all rotations and translations of the problem data.

Proof. This theorem is a direct consequence of the preceding observations.

It may be noted that the general form of Problems I and II also makes the foregoing results applicable to the case where strict group separation is sought by replacing b with $b - \epsilon$ in the constraints applicable to Group 1 and with $b + \epsilon$ in the constraints applicable to Group 2.

We conclude this section by observing that the defect illustrated in Table 16.1 and Figure 16.1 for the β normalization is overcome by (N). In particular, the distortion of the solution caused by the β normalization in this example occurred because a shift of b (from its "best value" of 0) caused the

to satisfy the β normalization. As a result, it was impossible to hold x constant to find the optimal b value, given x, since moving b forced x to change as well. The normalization (N) is free of this defect for the important reason that it is entirely possible to hold x constant and change b without any effect on the normalization constraint. Thus, the normalization (N) gives the same objective function values as the graphical analysis of the example of Table 16.1 and Figure 16.1, and identifies the same solution as optimal.

MODEL MANIPULATION AND SIMPLIFICATIONS

Our primary goal will be to identify how the model (1) - (6) can be manipulated to achieve an "equal representation" of the points in Group 1 and Group 2. This hinges on another more basic observation, which makes it possible to reduce the size of the model in the case where some points may have the same coordinates as others [to avoid including a separate constraint equation (and corresponding α_i and β_i variables) for each duplicate point].

Specifically, let S denote a collection of points all in G_1 or all in G_2, such that $A_p = A_q$ for each p, q in S. If S is a subset of G_1, then the equations of (2) corresponding to $i \in S$ can be replaced by a single representative equation $A_r x - \alpha_0 - \alpha_r + \beta_0 + \beta_r = b$, where A_r is the common vector A_i for all $i \in S$. If S is a subset of G_2, the equations of (3) corresponding to $i \in S$ can similarly be replaced by the representative equation $A_r x + \alpha_0 + \alpha_r - \beta_0 - \beta_r = b$. In each case, assuming that the h_i and k_i values are chosen in accordance with the stipulations of the preceding section and that the normalization (N) is employed, it follows that an optimal solution before the replacement occurs must yield the same values of α_i and β_i for each $i \in S$, and, hence, we are at liberty to interpret the values received by α_r and β_r as representing the common values.

To assure that the optimal solutions before and after replacement are the same under this interpretation, it suffices to let h_r and k_r, respectively, equal the sums of the h_i and k_i coefficients for $i \in S$. (It is reasonable in the original model to give these coefficients the same two values—say, h* and k*, for all $i \in S$, in which case $h_r = h^*|S|$ and $k_r = k^* S|$.) The necessary and sufficient conditions for bounded optimality identified in the previous section will hold after the replacement if they held before the replacement.

The manner in which this model simplification can be used to achieve an "equal representation" of Group 1 and Group 2 is as follows. If the two groups are of different sizes, we make n_2 copies of each point in G_1 and n_1 copies of each point in G_2, so that the two groups effectively are given the

same number of elements. The resulting representation does not enlarge the model formulation, since by the foregoing observation we may replace each h_i and k_i by n_2h_i and n_2k_i for $i \in G_1$, and by n_1h_i and n_1k_i for $i \in G_2$, without requiring the creation of additional variables or constraints in order to handle the implicitly generated copies of the original points.

By analogy with the case where all h_i (and all k_i) begin with the same value for the two groups, we may generally regard the objective function coefficients to be unbiased with respect to the sizes of the sample groups G_1 and G_2, if after the indicated adjustment, $\sum_{i \in G_1} h_i = \sum_{i \in G_2} h_i$ and $\sum_{i \in G_1} k_i = \sum_{i \in G_2} k_i$.

In the integer programming case, if the numbers of points in the two groups are not the same, then instead of minimizing the number of misclassified points, it may be more reasonable to minimize a weighted sum that gives Group 1 and Group 2 equal representation in the foregoing sense. Using the approach indicated for the LP case, we may minimize the number of misclassifications for this adjusted problem by replacing the IP objective (7) with

$$\text{Minimize} \quad n_2 \sum_{i \in G_1} z_i + n_1 \sum_{i \in G_2} z_i$$

On the basis of the preceding observations, we now examine connections between the LP and IP formulations.

LINKS BETWEEN THE LP AND IP DISCRIMINATION MODELS

Our first result link in the IP and LP formulations is to show that the LP formulations using (N) enjoy a special property that causes an optimal solution to "balance" the number of misclassified points across the two groups, whenever the objective weights each point equally.

We focus attention on the *Min Sum LP Model,* where α_o and β_o are deleted and the objective function takes the following simple form:

$$\text{Minimize:} \quad \sum_{i \in G} \alpha_i$$

It is to be emphasized that our results also apply to problems more general than the Min Sum model by making use of the constructions of the previous section.

We define the number of misclassified points to be balanced between Group 1 and Group 2 if the number of points with $\alpha_i > 0$ in each group does not exceed the number of points with $\alpha_i \geq 0$ in the other group. In the absence of points with $\alpha_i = 0$, this condition implies the number of misclassified points in each group will be the same.

Theorem 4

An optimal solution to the Min Sum LP model using normalization (N) yields a balanced number of misclassified points.

Proof. Since at most one of α_i and β_i will be positive for each i, it follows that $\alpha_i > 0$ for $i \in G_1$ only if $A_i x > b$, and that $\alpha_i > 0$ for $i \in G_2$ only if $A_i x < b$. Let n_k^+ and n_k^0, respectively denote the number of positive and zero α_i for $i \in G_k$ and k=1, 2. If b is increased by a small positive value ϵ, then all $\alpha_i > 0$ for $i \in G_1$ are decreased by ϵ, and all $\alpha_i \geq 0$ for $i \in G_2$ are increased by ϵ, yielding a net increase in $\Sigma\alpha_i$ of $(n_2^+ + n_2^0)\epsilon - n_1^+ \epsilon$. Since this value must be non-negative, we conclude that $n_2^+ + n_2^0 \geq n_1^+$. Consideration of decreasing b by ϵ similarly yields $n_1^+ + n_1^0 \geq n_2^+$. This completes the proof.

One consequence of Theorem 4 for linking the LP and IP formulations is that a balanced number of misclassified points must be "close" to a minimum number of misclassified points, when x is held constant and b is allowed to vary. This is expressed more precisely in the following result.

Corollary

Starting from an optimal LP solution, the greatest reduction in the number of misclassified points that can be obtained by holding *x* constant and varying *b* cannot exceed Min (n_1^+, n_2^+).

Proof. By the reasoning of the proof of Theorem 4, if b is increased, the largest possible reduction in the number of misclassified points is $n_1^+ -$

n_2^0, which is at most n_2^+ (as well as at most as n_1^+). The corresponding conclusion for decreasing b yields the result of the corollary.

The foregoing observations also make it possible to identify a value b that minimizes the number of misclassified points subject to holding x at its optimal LP value. That is, instead of relying on the worst case bound of the corollary, we may apply a method that identifies precisely the amount of reduction in misclassified points that is possible by shifting *b* and that further identifies the value of b that achieves this reduction. The method is as follows.

Method to Optimize b, Given x

0. Begin with $\theta^* = n_1^+ + n_2^+$, and perform the following steps for each Group k, k = 1, 2, such that $n_k^+ > 0$. Upon termination, θ^* will be the minimum number of misclassified points.
1. Restrict attention to those $i \in G_k$ such that $\alpha_i > 0$, and arrange these α_i in ascending order, reindexing for simplicity so that $\alpha_1 \leq \alpha_2 \leq \ldots \leq \alpha_r$, where $r = n_k^+$.
2. Examine the α_i, i = 1, 2, ..., r in sequence, considering in turn that $\alpha_h < \alpha_{h+1}$ (or such that h = r).
   a. Define $b_h = b = \alpha_h$ for k = 1 and $b_h = b - \alpha_h$ for k = 2. If b is given the value b_h, the number of misclassified points is $\theta_h = n_1 + n_2 - n_k^0 + \Delta_h - h$, where Δ_h is the number of points A_i, for $i \in G_k$, such that $0 < \beta_i < \alpha_h$.
   b. If $\theta_h < \theta^*$, let $\theta^* = \theta_h$. Stop for the current value of k if $\theta_h - n_k^+ + h > \theta^*$; otherwise, return to Step 2a for the next value of $h \leq r$.

The stopping criterion of step 2b is based on the observation that the number of positive α_i not yet examined is $n_k^+ - h$, and hence this number represents an upper limit on the possible decrease in θ_h. The justification of the method derives from the logic underlying the proof of Theorem 4, from which it follows that an optimal b value is identified as the one that yields θ^*.

It is natural to ask whether a connection between the LP and IP formulations can be established that gives an indication of the quality of the LP solution for the IP problem in a more global sense, in contrast to the local sense of holding x constant, while b varies. The following answers this affirmatively.

Theorem 5

Assume that Group 1 and Group 2 have a meaningful separation and that the normalization (N) is employed. Then there exists a finite positive U* such that for all $U \geq U^*$, the Min Sum LP formulation is a valid continuous relaxation of the IP formulation (7) - (12) under the scaling $\alpha_i = Uz_i$, $i \in G$.

Proof. Starting with the IP formulation and substituting a_i/U for z_i causes equations (8) and (9) to become the same as equations (2) and (3), while equation (10) becomes $\alpha_i = 0$ or U, which relaxes to $0 \leq \alpha_i \leq U$, for $i \in G$. The objective (7) then becomes Minimize $(1/U) \sum_{i \in G} \alpha_i$, and, hence, the Min Sum LP formulation with $\alpha_i \leq U$, $i \in G$, has the same set of optimal solutions as the IP formulation upon relaxing $z_i = 0$ or 1 to $0 \leq z_i \leq 1$, $i \in G$. The key is therefore to demonstrate the existence of U* such that the IP formulation achieves its intended purpose of minimizing the number of positive α_i, while simultaneously assuring the relaxation is valid for all $U \geq U^*$.

Replace the objective of the Min Sum formulation by Minimize $\sum_{i \in G} f(\alpha_i)$, where $f(\alpha_i) = 1$ if $\alpha_i > 0$ and $f(\alpha_i) = 0$ otherwise. This problem has an optimal solution x*, b*, α_i^*, and β_i^*, $i \in G$, under normalization (N) with all α_i^* finite, and we can require that at least one of α_i^* and β_i^* is 0 for all i. (To see this, minimize $\sum_{i \in S} \alpha_i$ for every subset S of G, holding $\alpha_i = 0$ for $i \in S$. A smallest cardinality subset that has a feasible solution yields the indicated finite solution.) Then for $U^* = \text{Max}(\alpha_i^* : i \in G)$ and any $U \geq U^*$, the solution determined by setting $x = x^*$, $b = b^*$, and $\alpha_i = U$ if and only if $\alpha_i^* > 0$, is optimal for the Min Sum problem subject to the added condition $\alpha_i = 0$ or U. This follows from the fact that increasing α_i^* and β_i^* by the same amount, $U - \alpha_i^*$, for all i such that $\alpha_i^* > 0$, yields a solution that continues

to satisfy all problem constraints and the normalization (N) without changing x* and b*, giving an objective function value of $U \sum_{i \in G} f(\alpha_i^*)$. There cannot be a better solution to the Min Sum problem subject to $\alpha_i = 0$ or U, using (N), since by reversing the preceding derivation we would thereby obtain a solution better than the one assumed optimal for the problem of minimizing $\sum_{i \in G} f(\alpha_i)$. The proof is completed by allowing U* to increase, if necessary, so that the constraint $\alpha_i \leq U^*$, $i \in G$, is redundant for the Min Sum LP formulation.

We note that the crucial aspect of the preceding theorem was to establish the ability to choose any $U \geq U^*$, something not possible with the β normalization. In the case of the (N) normalization, we can additionally replace U by positive values $U_i \geq \alpha_i^*$ for each $i \in G$, provided the objective for the LP problem is correspondingly replaced by Minimize $\sum_{i \in G} \alpha_i/U_i$. (If this is done, the relaxation theoretically may be tightened by adding the constraints $\alpha_i \leq U_i$ for $i \in G$, yielding progressively better relaxations as U_i is chosen closer to α_i^*. However, approximate knowledge of α_i^* values, and of relative differences between them, is not typically possible.)

The proof of Theorem 5 in fact shows that the Min Sum solution, where the bound $U \geq \alpha_i$ is disregarded, yields a relaxation that cannot be improved for all U that are at least as large as the maximum α_i value in the LP solution. If this value is no larger than U*, the Min Sum relaxation is as good as choosing $U = U^*$.

We now show that the IP problem can be solved without knowing (or provisionally selecting) a value for U. Consider the process of solving the IP problem by branch and bound, where an appropriate value of U is known. A branch that sets $z_i = 1$, or equivalently $\alpha_i = U$, replaces the associated constraint by

$$A_i x + \beta_i = b + U \qquad \text{if } i \in G_1$$

$$A_i x - \beta_i = b - U \qquad \text{if } i \in G_2$$

For U large, the effect of this replacement is simply to make the associated constraint redundant since α_i is not a given weight in the objective

function and takes any value necessary to product equality. Thus, in particular, the branch of setting $\alpha_i = U$ can be handled by removing the associated constraint or simply by changing the objective function coefficient of α_i to 0, which also effectively makes the constraint redundant. (The latter has the further advantage of allowing the branch and bound process to continue by primary feasible postoptimization.) Accompanying this change, the variable α_i is replaced by the objective function by the constant term U. Recording this modified form of the objective can be simply a bookkeeping formality, with no need to give a value to U.

It may also be noted that the proof of Theorem 5 implies that the Min Sum objective in fact can be replaced by any other that weights all α_i positively, still yielding a valid relaxation of the IP problem for some set of U_i values. This observation leads to the possibility of a postoptimizing strategy for modifying the LP objective function coefficients to come closer to minimizing the number of positive α_i. One approach for doing this is as follows.

Postoptimizing Heuristic to Minimize the Number of Misclassifications

1. Replace the objective for the Min Sum LP problem by Minimizing $\sum_{i \in G} h_i \alpha_i$, where all h_i are chosen to be positive (e.g., initially let all $h_i = 1$), and solve the resulting LP problem.
2. If the current LP solution yields a smaller number of positive α_i values than any solution so far, record this as the best candidate solution. (The first solution is automatically recorded as such a candidate.)
3. If $h_i\alpha_i = 0$ for all $i \in G$, the method terminates. Otherwise, select $h_p\alpha_p = \text{Max}(h_i\alpha_i: i \in G)$, and set $h_p = 0$.
4. Postoptimize to solve the resulting LP problem, and return to step 2.

The motivation for this procedure is that $x_i = 0$ must result for all i such that $h_i > 0$ if such a solution is feasible. The variable α_p may be viewed as one that most strongly resists being driven to 0. Hence, h_p is set to 0, forcing remaining variables α_i to 0. After the procedure terminates the strategy can be reversed by choosing α_p to be the smallest positive α_i such that $h_i = 0$, if any

exist, and by making h_p positive. (If the process is repeated a few steps beyond where improvement results, then the original strategy can be activated once again.) The method can be coupled with the earlier method for optimizing b, given x.

In applying the procedures of this section, it should be remembered that the Min Sum formulation may give a weight of n_2 to points of G_1 and a weight of n_1 to points of G_2 to create an equal representation relative to size, and the case of duplicate points can cause the h_i values to vary in additional ways. The foregoing discussion of the Min Sum model applies to all of these cases under the interpretation that h_i = the number of times (possibly fractional) that point i occurs. Then, in the heuristic for minimizing the number of misclassifications, candidate solutions are evaluated by reference to the sum of these original h_i values over those *i* such that $\alpha_i > 0$. A corresponding observation applies to the solution of IP problems where the h_i values represent costs of misclassification.

A SUCCESSIVE GOAL APPROACH

A particularly significant use of the model results by a successive application employing hierarchically weighted deviation terms, which were proposed for its early special cases consisting of the MMD and MSD forms in Freed and Glover (1981 and 1987) and which can now be implemented without distortion by reliance on (N). The relevance of the IP results to this process derives from the fact that each stage involves a valid relaxation of a corresponding IP formulation. Such an approach is applicable to settings where multiple groups are to be differentiated or where two groups are treated as multiple groups by redefining subsets of points improperly classified at one stage of the application as new groups to be differentiated at the next. For the multiple group case, any subset of groups can be defined to be Group 1 and the remaining subset defined to be Group 2, thus encompassing alternatives ranging from a binary tree form of separation to a "one-at-a-time" form of separation.

By this approach, when the two currently defined groups are incompletely separated at a given stage, the hyperplane dividing them may be shifted alternately in each direction (increasing and decreasing b) by an amount sufficient to include all points of each respective group. (The magnitude of the two shifts will be the same for the MMD model, which minimizes both the maximum value and the sum of these shifts.) Upon identifying the shift for a given group, all points of the alternate group that lie strictly beyond the shifted hyperplane boundary become perfectly differentiated by this means, and such perfectly differentiated points can be

segregated from the remaining points before applying the next stage. The number of stages devoted to creating perfect separation (before accepting the current hyperplane, without shifting) is a decision parameter of the process.

It is important in such a process, if a superior set of differentiating hyperplanes is sought, to retain points in the model that have been segregated as perfectly differentiated, rather than dropping them from consideration during subsequent stages. To reflect the fact that these segregated points should not inhibit the goal of differentiating among remaining points, their deviation terms are assigned objective function weights that are hierarchically of a lower order than are those assigned to points not yet segregated. The relative magnitudes of these lower order weights may reasonably be scaled to become progressively smaller for points segregated earlier in time. (In addition, to reduce problem size, a subset of the points most recently segregated may be discarded at each stage, where this subset is identified to consist of points lying beyond a chosen *magnified shift* of b. It is easy to shift b, for example, to a depth that excludes any selected percentage of most recently segregated points belonging to a specified group.)

We call this approach the *successive goal method* because the introduction of hierarchical differences in deviation weights, with diminishing weights for points segregated earlier, constitutes a natural partitioning of problem points into subset by reference to prioritized goals. Furthermore, the ability to manipulate weights within a given goal level (or to split out additional hierarchies) makes it possible to treat the two groups of points that remain unsegregated at a given stage in a nonsymmetric manner.

This leads to an approach that characteristically is able to generate a stronger set of hyperplanes, at the expense of approximately doubling the overall computational effort. The basis of this nonsymmetric approach rests on creating successive objectives to exclude a maximum segment of one group from a region that contains all of the others in a series of alternating hierarchies.

The *alternating hierarchy method* that results has the property of adapting successive hyperplanes to more closely match the distributions of the groups and generally increases the frequency with which earlier hyperplanes are permitted to be discarded as redundant. The procedure consists of solving two problems at each stage. Each of the two groups of currently unsegregated points is chosen in turn to be the one that lies completely within the region assigned to it by the current hyperplane, with the associated (subordinate) goal of excluding the maximum portion of the other group from this region.

The structure of the goals for each problem gives rise to the "alternating hierarchy" characterization of this procedure. Specifically, we adopt the conviction that the group to be completely contained in its assigned region is always designed to be Group 1. Then the problem goals are ordered as

follows. At the highest level, only the external deviations of unsegregated Group 1 points are incorporated into the objective (which is equivalent to imposing the condition $A_i x \leq b$ for these points). At the next level, the external deviations of unsegregated Group 2 points are assigned corresponding lower-order weights in the objective, thus respecting the dominance of the level preceding. For the points of this second level, the b term is replaced by $b + \epsilon$ to seek strict separation. (Alternatively, a restricted β_0 variable, which appears only in the equations for the second-level points, may be incorporated with a positive weight.) At the third and fourth levels, respectively, external deviations of segregated Group 1 and Group 2 points receive weights reflecting their associated position in the hierarchy (or a single third level may treat these segregated points uniformly). Finally, two concluding levels incorporate internal deviations of both groups, first for unsegregated points and then for segregated points. These last levels are relevant to enhancing the differentiation between those groups, which are in fact separable, and may be expected to have diminished relevance after generating the first few hyperplanes.

The portion of unsegregated Group 2 points that are perfectly differentiated from unsegregated Group 1 points at a given stage, and hence that can join the set of segregated points in the stage following, may vary substantially depending on which group is chosen to be Group 1. In fact, one of the two choices for Group 1 may fail to differentiate any of the unsegregated Group 2 points (i.e., all such points may lie in the half space required to include the unsegregated Group 1 points). When the sets of points differentiated by the two choices differ significantly in size, the smaller set can be excluded form joining the segregated points on the next stage—an exclusion that, in effect, will occur automatically if the smaller set is empty. If both sets are empty, the process stops. Because of the alternating dominance of the two groups in each of the problems solved, no shifting of hyperplanes is needed in this approach. (For added refinement, after a forward pass of generating a selected set of hyperplanes, a reverse pass can be applied to improve the differentiation.)

From a practical standpoint, the hierarchical levels of this approach can be handled with greater efficiency by dividing the solution process into stages. At the first stage, attention is restricted to the objective function associated with the highest level until that objective is optimized. Then, following a process analogous to that employed by Phase 1/Phase 2 LP methods, nonbasic variables with nonzero reduced costs are fixed at their current values, and the objective appropriate to the next level is introduced and optimized. The process repeats until all levels are treated or all remaining basic variables receive fixed values (thus implicitly determining solutions for

levels not yet examined). This approach requires notably less computational effort than an implementation which relies on large coefficient differences to control the treatment of hierarchies. Independent of implementation details, the approach provides an opportunity to achieve progressively improved differentiation of the original group in both the two-group and the multiple-group cases and opens up interesting research possibilities for determining the best subset of points to be segregated at each stage.

CONCLUSIONS

The LP discriminant analysis formulation (1) - (6) is susceptible to a variety of uses as a result of the ability to handle different discriminant analysis goals by varying the coefficients of the objective function. Such uses range from accommodating inherent differences in the need to classify specific points correctly to employing strategies for producing greater refinement in classification (as by the successive goal method).

Among the settings of practical relevance, situations in which there are real dollar costs for misclassifications can be modeled in a natural and highly appropriate manner by such a model. Many applications gain additional realism by an integer programming interpretation. The fact that the LP formulation employing the normalization (N) is a direct relaxation of the corresponding IP problem, and lends itself to convenient strategies for closing potential gaps between LP and IP solutions, gives further motivation for using this type of model. Related forms of postoptimizing strategies can be applied to achieve additional goals, such as diminishing the effects of outliers (whose identities are disclosed by the initial solution) without the risk of being driven to "wrong solutions" when objective function coefficients are thereby modified.

Postoptimization is also useful in the "ϵ version" of the model to identify values of the ϵ that yield different separation effects. In particular, this model version is equivalent to introducing a translation of the β_o variable by the lower bound $\beta_o \geq \epsilon$. Thus, standard sensitivity analysis on the LP solution with β_o, included in the model can precisely determining the outcome of increasing β_o, hence ϵ, up to the point where a new optimal basis results, and a postoptimization step can then move to this new basis, allowing the analysis to repeat for larger ϵ values. Such a mapping of the effects of different ϵ values provides an interesting area for optimization, and has been studied in the context of international loan portfolios in Glover, Keene, and Duea, 1988.

From another perspective, the ability to weight the internal and external deviations differently for different points, and to encompass tradeoffs between such deviations and "minmax" and "maxmin" objectives, provides a direct way to handle issues that are often troubling in classical discriminant analysis. A prominent example is the type of problem in which Type I and Type II errors deserve different emphasis. As pointed out in Mahmood and Lawrence (1987), in the context of identifying firms that succumb to bankruptcy, it may be more important to be assured that a firm classed as financially strong will in fact escape bankruptcy than to be assured that a firm classed as financially weak will become insolvent.

Indeed, by the capacity to giver higher weights to firms that are dramatically successful and unsuccessful, the LP formulation will tend to position the "sure bets" more deeply inside their associated half spaces. The advantage of this is that it provides increased predictive accuracy: instead of investing in a business simply on the basis of whether discriminant analysis classifies it as financially strong or financially weak, greater confidence may be gained by investing in a firm that lies well within the financially strong region. The successive goal method provides an opportunity to additionally improve the discrimination in such cases. By the ability to remove distortion with the normalization (N), the uses of different objective function coefficients that underlie these approaches can be applied consistently and effectively.

NOTE

This chapter was prepared based on the author's earlier article, "Improved Linear Programming Models for Discriminant Analysis," in *Decision Sciences*, Volume 21, Number 4, Fall 1990, pp. 771–785, by elaborating on several aspects of the paper in further detail and adding new results on integer programming models for discriminant analysis.

REFERENCES

Bajgier, S. M., and Hill, A. V., "An Experimental Comparison of Statistical and Linear Programming Approaches to the Discriminant Problem," *Decision Sciences* 13, no. 4 (October 1982): 604-618.

Bobrowski, L., "Linear Discrimination with Symmetrical Models," *Pattern Recognition* 19, no. 1 (1986): 101-109.

Charnes, A., Cooper, W. W., and Rhodes, E., "Evaluating Program and Managerial Efficiency: An Application of Data Envelopment Analysis to Program Follow Through," *Management Science* 27 (1981): 668-687.

Freed, E. (Ned), and Glover, F., "Simple but Powerful Goal Programming Models for Discriminant Problems," *European Journal of Operational Research* 7, no. 1 (May 1981): 44-60.

Freed, E. (Ned), and Glover, F., "Resolving Certain Difficulties and Improving the Classification Power of the LP Discriminant Analysis Procedure," *Decision Sciences* 17 (1987): 589-595.

Glover, F., Gordon, K., and Palmer, M., "LP Discriminant Analysis for International Loan Portfolio Management," CAAI 89-3, University of Colorado, April 1989.

Glover, F., Keene, S., and Duea, B., "A New Class of Models for the Discriminant Problem," *Decision Sciences* 19 (1988): 269-280.

Jurs, P. C., "Pattern Recognition Used to Investigate Multivariate Data in Analytical Chemistry," *Science* 232, no. 6 (June 1986): 1219-1224.

Kazmier, L., *Statistical Analysis for Business and Economics*, McGraw Hill, New York, 1967.

Mahmood, M. A., and Lawrence, E. C., "A Performance Analysis of Parametric and Nonparametric Discriminant Approaches to Business Decision Making," *Decision Sciences* 19, no. 2 (Spring 1987): 308-326.

Markowski, E. P., and Markowski, C. A., "Some Difficulties and Improvements and Applying Linear Programming Formulations to the Discriminant Problem," *Decision Sciences* 16, no. 3 (Summer 1985): 237-247.

Spurr, W., and Bonini, C., *Statistical Analysis for Business Decision*, Richard D. Irwin, Homewood, IL, 1967.

Tou, J. T., and Gonzalez, R. C., *Pattern Recognition Principles*, Addison-Wesley, Reading, MA, 1974.

Watanabe, S. *Methodologies of Pattern Recognition*, Academic Press, New York, 1969.