Improved Linear Programming Models for Discriminant Analysis*

Fred Glover
Center for Applied Artificial Intelligence, Graduate School of Business, University of Colorado, Boulder, CO 80309-0419

ABSTRACT

Discriminant analysis is an important tool for practical problem solving. Classical statistical applications have been joined recently by applications in the fields of management science and artificial intelligence. In a departure from the methodology of statistics, a series of proposals have appeared for capturing the goals of discriminant analysis in a collection of linear programming formulations. The evolution of these formulations has brought advances that have removed a number of initial shortcomings and deepened our understanding of how these models differ in essential ways from other familiar classes of LP formulations. We will demonstrate, however, that the full power of the LP discriminant analysis models has not been achieved, due to a previously undetected distortion that inhibits the quality of solutions generated. The purpose of this paper is to show how to eliminate this distortion and thereby increase the scope and flexibility of these models. We additionally show how these outcomes open the door to special model manipulations and simplifications, including the use of a successive goal method for establishing a series of conditional objectives to achieve improved discrimination.
Subject Areas: Linear Programming and Statistical Techniques.

INTRODUCTION

There is growing recognition that a variety of classical statistical problems can be approached advantageously by 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 is drawing increased attention 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 [3] [10] [13]. Artificial intelligence applications involve the challenging realm of pattern recognition, including problems of differentiating signals, diagnostic classifications, code signatures and data types [2] [9] [14] [15].

An effort to wed statistical discrimination with optimization has come about by proposals to capture the goals of discriminant analysis in a collection of linear programming formulations [4] [5] [8]. Initial forms of these models included the objectives of 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 have disclosed the "minimum sum of deviations" model to be competitive in effectiveness with the classical approach of Fisher [9] [12]. This comparative testing was applied 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 post-optimization to reweight borderline misclassified points to obtain refined solutions, one of the strategic options of the LP approaches proposed with their earliest

*This research was supported in part by the Center for Space Construction of the University of Colorado under NASA Grant NAGQ-1388.

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 sometimes gave counterintuitive and even anomalous results. Follow-up examination of specially constructed examples demonstrated that these formulations were attended by certain subtleties not found in other areas to which linear programming is commonly applied [1] [5] [12].

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 assign zero weights to all data elements. Several normalizations have been identified [5] [8] 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 [8].

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 function and LP postoptimization.

The purpose of this paper 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 has attractive properties enabling it to obtain demonstrably superior solutions. The outcomes of these properties additionally include the ability to place any desired relative emphasis on classifying particular points correctly, and to create a conditionally staged application of the model, called the Successive Goal Method, for achieving progressively more refined discrimination both for two-group and multigroup analysis.

A HYBRID LP DISCRIMINANT MODEL

We take as our starting point the hybrid LP model of [8] which integrates features of previous LP discriminant formulations [4] [5]. 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 $\mathbf{A}_i$, 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 in a later section.)

To discriminate the points of the two groups, we seek a weighting vector $\mathbf{x}$ and a scalar b, which may be interpreted as providing a hyperplane of the form $\mathbf{Ax}=b$, where $\mathbf{A}$ takes the role of representing $\mathbf{A}_i$ 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 $\mathbf{A}_i\mathbf{x} < b$ for $i \in G_1$ and $\mathbf{A}_i\mathbf{x} > b$ for $i \in G_2$.

Refining this goal as in [8], 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 (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:

$$\text{Minimize } h_o\alpha_o + \sum_{i\in G} h_i\alpha_i - k_o\beta_o - \sum_{i\in G} k_i\beta_i, \tag{1}$$

subject to

$$\mathbf{A}_i\mathbf{x} - \alpha_o - \alpha_i + \beta_o + \beta_i = b, \qquad i\in G_1, \tag{2}$$

$$\mathbf{A}_i\mathbf{x} + \alpha_o + \alpha_i - \beta_o - \beta_i = b, \qquad i\in G_2, \tag{3}$$

$$\alpha_o,\ \beta_o \geq 0, \tag{4}$$

$$\alpha_i,\ \beta_i \geq 0, \qquad i\in G, \tag{5}$$

$$\mathbf{x},\ b \qquad \text{unrestricted in sign.} \tag{6}$$

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 $b-\epsilon$ for Group 1 and by $b+\epsilon$ for Group 2, where ϵ is a sected positive constant, to pursue the goal of compelling elements of Group 1 and Group 2 to lie strictly inside the half spaces 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 $b+\epsilon$ 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 nonnegative, although it is possible to allow the coefficients of the β 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 $h_i \geq k_i$ for $i=0$ and $i\in G$. Otherwise, it would be possible to take any feasible solution and increase the value of α_i and β_i (for $h_i < k_i$) 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 minimum 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 $\mathbf{A}_i\mathbf{x} - \alpha_o + \beta_o \leq b$ for $i\in G_1$, and $\mathbf{A}_i\mathbf{x} + \alpha_o - \beta_o \geq b$ for $i\in G_2$, at the expense of enlarging the model form. By deleting the α_o and β_o variables in (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 [4].

THE NORMALIZATION ISSUE

To understand the potential difficulties that underly 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 is incomplete, for it must be amended in some fashion to avoid an optimal solution that yields the null weighting $\mathbf{x}=0$. If the two groups can be separated by a hyperplane (or nearly so) and 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 $\mathbf{x}=0$ 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 zero 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 $\mathbf{A}_i$ is replaced by the point $\mathbf{A}_i+\mathbf{t}$ for a common vector $\mathbf{t}$ [1] [12].

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 [5]. Let N denote the index set for components of the $\mathbf{x}$ vector. Then the first two proposals for alternative normalizations to remedy the problems of setting b to a constant can be written in the form:

$$b+\sum_{j\in N}\mathbf{x}_j=\text{a constant};$$

$$\sum_{j\in N}\mathbf{x}_j=\text{a constant}.$$

Of these alternatives, the latter was proved in [5] 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 the right sign was selected. Second, the variables either directly or indirectly had to be 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 $\mathbf{A}_i$ is replaced by the point $\mathbf{A}_i\mathbf{R}$, and $\mathbf{R}$ is a rotation matrix.

The most recent attempt to settle the normalization issue occurred in [8] with the "β normalization"

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

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 both for translations and rotations of the problem data. Experimentation further showed 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 also 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 β 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 $\mathbf{A}_i$ has a single coordinate, and hence the weight vector $\mathbf{x}$ 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 zero.

The relevant data for the example are given in Table 1, indicating the coordinates and the penalties for being classified in the wrong group. A graph of the points is shown in Figure 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 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 $\mathbf{A}_2$ and $\mathbf{A}_4$ are misclassified, each with a deviation of 1 unit from the value $b=0$, hence giving a total penalty cost of $1\times25+1\times25=50$.

Without a normalization constraint, the LP model falls into the trap of finding a meaningless optimal solution, $\mathbf{x}=0$ and $b=0$, which makes all external deviations zero and hence also makes the total penalty cost zero. 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 to be best by graphical analysis. Indeed, we then obtain $\mathbf{x}=1$, $b=0$ (with $\alpha_4=\alpha_2=1$, $\beta_3=\beta_6=2$, all other α_i and $\alpha_i=0$) as a feasible solution for the LP model, yielding a total penalty cost of 50, as before.

However, this solution turns out not to be optimal. Rather, the β normalization causes the inferior solution based on $\mathbf{x}=1$ and $b=-1$ to appear even better. From a graphical standpoint, the deviation variables with positive values for this solution are $\alpha_1=1$, $\alpha_2=2$, $\beta_3=1$, $\beta_5=1$, $\beta_6=3$ which yield a total penalty cost of 65. However, the sum of the β_i variables equals 5, and to rescale the solution to

Table 1: Coordinates and penalities for being classified in the wrong group.

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

Note: All $k_i = 0$.

satisfy the β normalization with right-hand side 4 the value of each variable must be multiplied by 4/5. The result is to multiply the total penalty cost of 65 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 (choosing in this case the constant to be 1), will result in correctly identifying 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.

THE NEW NORMALIZATION

The normalization we propose is

$$(-n_2 \sum_{i \in G_1} \mathbf{A}_i + n_1 \sum_{i \in G_2} \mathbf{A}_i)\mathbf{x} = 1 \tag{7}$$

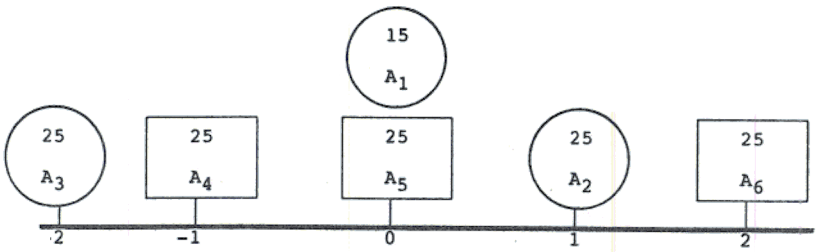
where n_1 and n_2 are, respectively, the number of elements in G_1 and G_2, and the right-hand side of 1 is an arbitrary scaling choice for a positive constant. (An alternative scaling that tends to yield x_j values closer to an average absolute value of 1 is to choose this constant to be $2n_1n_2$.) An equivalent form of this normalization occurs by adding n_2 times each equation of (1) and subtracting n_1 times each equation of (2) to yield the constraint

$$2n_1n_2(\beta_o - \alpha_o) + n_2 \sum_{i \in G_1} (\beta_i - \alpha_i) + n_1 \sum_{i \in G_2} (\beta_i - \alpha_i) = 1 \tag{8}$$

Expressing the normalization in the form (7) has certain advantages for analysis while expressing it in the form (8) is convenient for incorporation into the LP formulation (since the coefficients of the variables do not require calculation as in (7)). 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 (8), 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 (8), then a similar effect occurs in the case where it is possible to completely separate Group 1 and Group 2 points, i.e., where all α_i become zero.)

To understand the properties of the normalization given by (7) and (8), let d_i denote the net internal deviation of the point $\mathbf{A}_i$ from the hyperplane generated by the discriminant model; that is

Figure 1.



$$d_i = b - \mathbf{A}_i\mathbf{x} \qquad \text{for } i \in G_1,$$

$$d_i = \mathbf{A}_i\mathbf{x} - b \qquad \text{for } i \in G_2.$$

Hence d_i is positive (or zero) if $\mathbf{A}_i$ 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 (7) by the quantity $\epsilon(n_1+n_2)$, while leaving the constant term of the normalization (8) 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; 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), then a meaningful separation should yield a positive value for this weighted sum. We embody this observation in the following definition.

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

Theorem 1. The normalization (7) is equivalent (under scaling) to requiring a meaningful separation, and eliminates the null weighting $\mathbf{x}=0$ as a feasible solution.

The proof of the foregoing theorem and all other results of this paper appear in the Appendix. A direct consequence of the theorem is the following.

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 $\mathbf{A}_{ij}$ of each point $\mathbf{A}_i$, $i \in G$, such that

$$n_2 \sum_{i \in G_1} \mathbf{A}_{ij} \neq n_1 \sum_{i \in G_2} \mathbf{A}_{ij}.$$

Useful additional insights into the nature of (7) and its consequences for the hybrid LP discriminant formulation are provided by examining the linear programming dual of (1) through (6) with (7) 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 (7), we obtain the following result.

Dual Model Formulation

$$\text{Maximize } v_o$$

subject to

$$\mathbf{A}_o v_o - \sum_{i \in G_1} \mathbf{A}_i v_i + \sum_{i \in G_2} \mathbf{A}_i v_i = 0,$$

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

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

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

where

$$\mathbf{A}_o = -n_2 \sum_{i \in G_1} \mathbf{A}_i + n_1 \sum_{i \in G_2} \mathbf{A}_i.$$

Our interest in analyzing this dual is to determine circumstances that provide a feasible dual solution, and to 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 for certain variables of the LP discriminant formulation to be nonzero at optimality. The following is established by reference to the duality theory of linear programming.

Necessary Conditions for Bounded Optimality

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

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

Necessary Conditions for Variables to be Nonzero

$$\text{For } \beta_o\text{: } h_o < \sum_{i \in G} h_i$$

$$\text{For } \beta_o\text{: } k_o > \sum_{i \in G} k_i$$

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 that 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, (1) through (6), with the normalization, (7), is bounded for optimality whenever

$$\text{Min } (h_o/2,\ n_1 h_i : i \in G_1,\ n_2 h_i : i \in G_2)$$

is at least as large as

$$\text{Max } (k_o/2,\ n_1 k_i : i \in G_1,\ n_2 k_i : i \in G_2).$$

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_o or β_o is deleted, the corresponding term involving h_o or k_o is deleted from its statement. When both α_o and β_o 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_o) are positive. If the inequality of Theorem 2 is not satisfied, let π be the ratio of the Max term to the Min term of the theorem. Then upon replacing each πh_i by ph_i in the objective (1), the condition of the theorem is satisfied. This modification of the coefficients of (1) has the property that the relative magnitudes of the h_i coefficients, and also of the k_i coefficients, are left unchanged. Thus, the theorem shows it is possible to choose the coefficients of (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 coefficients that preserves this relative emphasis.

Our next goal is to provide the result which establishes that the normalization (7) is stable across standard data transformations.

Theorem 3. The optimum objective function values and optimal values for the α and β deviation variables in the LP discriminant formulation are unchanged for all rotations and translations of the problem data.

Theorems 1 through 3 are likewise 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 1 and Figure 1 for the β normalization is overcome by (7). 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 zero) caused the sum of the β_i values to change, hence requiring that the solution be rescaled to satisfy the β normalization. As a result, it was impossible to hold $\mathbf{x}$ constant to find the optimal b value, given $\mathbf{x}$, since moving b forced $\mathbf{x}$ to change as well. The normalization (7) is free of this defect for the important reason that it is entirely possible to hold $\mathbf{x}$ constant and change b without any effect on the normalization constraint. Thus the normalization (7) gives the same objective function values as the graphical analysis of the example of Table 1 and Figure 1, and identifies the same solution as optimal.

MODEL MANIPULATIONS AND SIMPLIFICATIONS

Our primary goal will be to identify how the model (1) through (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; that is, avoiding the necessity of 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 $\mathbf{A}_p = \mathbf{A}_q$ 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 $\mathbf{A}_r\mathbf{x} - \alpha_o - \alpha_r + \beta_o + \beta_r = b$, where $\mathbf{A}_r$ is the common vector $\mathbf{A}_i$ 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 $\mathbf{A}_r\mathbf{x} + \alpha_o + \alpha_r - \beta_o - \beta_r = b$. In each case, assuming the h_i and k_i values are chosen in accordance with the stipulations of the preceding section, and the normalization (7) 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 these common values.

To assure that 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

$$\sum_{i \in G_1} k_i = \sum_{i \in G_2} k_i.$$

A SUCCESSIVE GOAL APPROACH

Particularly significant is the potential to use hierarchically weighted deviation terms in the successive application of the model, as proposed for its early special

cases consisting of the MMD and MSD forms in [4] [5] and which can now be implemented without distortion by reliance on (7). Such an approach is relevant 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 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 which lie strictly beyond the shifted hyperplane boundary become perfectly differentiated by this means, and such perfectly differentiated points can be segregated from 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 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 subsets 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 other, 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 in 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 convention that the group to be completely contained in its assigned region is always designated 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 $\mathbf{A}_i\mathbf{x} \leq b$ 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_o 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 can join the set of segregated points on 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 from 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 nonbasic 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 multiple group cases, and opens up interesting research possibilities for determining the best subsets of points to be segregated at each stage.

CONCLUSIONS

The LP discriminant analysis formulation (1) through (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, and it can be shown [6] that the LP formulation employing the normalization (7) is a direct relaxation of a corresponding IP problem for minimizing the number of misclassified points (or a weighted variant of this objective), a result that does not hold for the β normalization. More broadly, the use of (7) makes it possible to employ postoptimization strategies for closing potential gaps between LP and IP solutions, and for achieving other 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 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 determine 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 is studied in the context of international loan portfolios in [7].

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 [11], 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 give 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 to provide increased predictive accuracy; instead of investing in a business based simply on whether discriminant analysis classifies it 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 (7), the use of differing objective function coefficients that underlies these approaches can be applied consistently and effectively. [Received: March 10, 1989. Accepted: August 22, 1989.]

REFERENCES

[1] Bajgier, S. M., & Hill, A. V. An experimental comparison of statistical and linear programming approaches to the discriminant problem. *Decision Sciences*, 1982, *13*, 604-618.

[2] Bobrowski, L. Linear discrimination with symmetrical models. *Pattern Recognition*, 1986, *19*(1), 101-109.

[3] Charnes, A., Cooper, W. W., & Rhodes, E. Evaluating program and managerial efficiency: An application of data envelopment analysis to program follow through. *Management Science*, 1981, *27*, 668-687.

[4] Freed E., & Glover, F. Simple but powerful goal programming models for discriminant problems. *European Journal of Operational Research*, 1981, *7*(1), 44-60.

[5] Freed, E., & Glover, F. Resolving certain difficulties and improving the classification power of the LP discriminant analysis procedure. *Decision Sciences*, 1986, *17*, 589-595.

[6] Glover, F. Exploiting links between linear and integer programming formulations for discriminant analysis. Working paper (CAAI 89-1). University of Colorado, 1989.

[7] Glover, F., Gordon, K., & Palmer, M. LP discriminant analysis for international loan portfolio management. Working paper (CAAI 89-3). University of Colorado, 1989.

[8] Glover, F., Keene, S., & Duea, B. A new class of models for the discriminant problem. *Decision Sciences*, 1988, *19*, 269-280.

[9] Jurs, P. C. Pattern recognition used to investigate multivariate data in analytical chemistry. *Science*, 1986, *232*(6), 1219-1224.

[10] Kazmier, L. *Statistical analysis for business and economics*. New York: McGraw Hill, 1967.

[11] Mahmood, M. A., & Lawrence, E. C. A performance analysis of parametric and nonparametric discriminant approaches to business decision making. *Decision Sciences*, 1987, *18*, 308-326.

[12] Markowski, E. P., & Markowski, C. A. Some difficulties and improvements and applying linear programming formulations to the discriminant problem. *Decision Sciences*, 1985, *16*, 237-247.

[13] Spurr, W., & Bonini, C. *Statistical analysis for business decision*. Homewood, IL: Richard D. Irwin, 1967.

[14] Tou, J. T, & Gonzalez, R. C. *Pattern recognition principles*. Reading, MA: Addison-Wesley, 1974.

[15] Watanabe, S. *Methodologies of pattern recognition*. New York: Academic Press, 1969.

APPENDIX

Proof of Theorem 1. First, (7) reduces to an inconsistent equation when $\mathbf{x}=0$, and hence renders the null solution infeasible. To see that (7) 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

$$n_2 \sum_{i \in G_1} (b - \mathbf{A}_i\mathbf{x}) + n_1 \sum_{i \in G_2} (\mathbf{A}_i\mathbf{x} - b) > 0.$$

Algebraic manipulation and reduction permits this inequality to be reexpressed in the form

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

whose left-hand side corresponds to the left-hand side of (7). 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 (7). Hence, allowing for scaling, the solutions are equivalent. Similarly, any feasible solution that satisfies (7) automatically satisfies the definition of a meaningful separation. This completes the proof.

Proof of the Corollary. The corollary is a direct consequence of Theorem 1 and the form of (7).

Proof of Theorem 2. Replace (7) by (8) in the primal formulation, whereon the Dual Problem becomes

Maximize v_o

subject to

$$-\sum_{i\in G_1} \mathbf{A}_i v_i + \sum_{i\in G_2} \mathbf{A}_i v_i = 0,$$

$$h_o \geq -2n_1n_2v_o + \sum_{i\in G} v_i \geq k_o,$$

$$h_i \geq -n_2v_o + v_i \geq k_i$$

$$h_i \geq -n_1v_o + v_i \geq k_i$$

Here v_o 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_o. Expressing these as bounds on $-n_1n_2v_o$ in each case, and then comparing terms, yields the inequalities stated in the theorem. This completes the proof.

Proof of Theorem 3. We omit this proof, noting that the result can be established by reference to the Stabilty Theorem of [8], using the (8) form of the normalization.

Fred Glover is the US West Chair in System Science and Research Director of the Center for Applied Artificial Intelligence at the University of Colorado, Boulder. He has authored or coauthored more than two hundred published articles in the fields of mathematical optimization, computer science and artificial intelligence, with particular emphasis on practical applications in industry and government. In addition to holding editorial posts for journals in the U.S. and abroad, Dr. Glover has been featured as a National Visiting Lecturer by the Institute of Management Science and the Operations Research Society of America and has served as a host and lecturer in the U.S. National Academy of Sciences Program of Scientific Exchange. He has served on the boards of directors of several companies and is cofounder of Analysis and Research and Computations, Inc., Management Science Software Systems, and the nonprofit research organization, Decision Analysis and Research Institute.