MULTIGROUP DISCRIMINANT ANALYSIS USING LINEAR PROGRAMMING

Willy Gochet
Katholieke Universiteit Leuven, Leuven, Belgium

Antonie Stam
The University of Georgia, Athens, GA, USA and
International Institute for Applied Systems Analysis
Laxenburg, Austria

V. Srinivasan
Stanford University, Stanford, CA, USA

Shaoxiang Chen
Nanyang Technological University, Singapore

RR-97-16
December 1997

Reprinted from *Operations Research*, Volume 45, Number 2,
March–April 1997.

International Institute for Applied Systems Analysis, Laxenburg, Austria
Tel: +43 2236 807 Fax: +43 2236 73148 E-mail: publications@iiasa.ac.at

---

Reprinted with permission from *Operations Research*, Volume 45, Number 2, March–April 1997.

MULTIGROUP DISCRIMINANT ANALYSIS USING LINEAR PROGRAMMING

WILLY GOCHET

Katholieke Universiteit Leuven, Leuven, Belgium

ANTONIE STAM

The University of Georgia, Athens, Georgia
and
International Institute for Applied Systems Analysis, Laxenburg, Austria

V. SRINIVASAN

Stanford University, Stanford, California

SHAOXIANG CHEN

Nanyang Technological University, Singapore

(Received February 1993; revisions received November 1994, August 1995; accepted August 1995)

In this paper we introduce a nonparametric linear programming formulation for the general multigroup classification problem. Previous research using linear programming formulations has either been limited to the two-group case, or required complicated constraints and many zero-one variables. We develop general properties of our multigroup formulation and illustrate its use with several small example problems and previously published real data sets. A comparative analysis on the real data sets shows that our formulation may offer an interesting robust alternative to parametric statistical formulations for the multigroup discriminant problem.

Recently, various mathematical programming (MP)-based approaches have been proposed for solving the classification problem in discriminant analysis (Bajgier and Hill 1982; Freed and Glover 1981a, 1981b; Gehrlein 1986; Hand 1981; Smith 1968, 1969; Stam and Joachimsthaler 1989; Stam and Ragsdale 1992). There is empirical evidence that these nonparametric methods may produce more accurate classification rules than the traditional statistical methods such as Fisher's linear discriminant method (Fisher 1936) and Smith's quadratic discriminant method (Smith 1947), which are based on the assumption of multivariate normality, if this assumption is violated to a significant extent. However, the experience with MP-based methods is not uniformly positive (Nath et al. 1992; Joachimsthaler and Stam 1990). A comprehensive overview of empirical studies using MP-based approaches to classification analysis is provided by Joachimsthaler and Stam (1990). A good review of MP formulations for solving the classification problem can be found in Erenguc and Koehler (1990) and Stam (1997).

However, a major drawback of most existing MP formulations is that they are limited to the two-group case, and their extension from the two-group case to the general multigroup case is problematic at best. Gehrlein (1986) proposes a formulation for the multigroup case which unfortunately requires a multitude of binary variables in order to identify the optimal division of segments of the decision space among the various groups, rendering its implementation infeasible in practice for many real-size data sets. Freed and Glover (1981b) remark that the minimize the sum of deviations (MSD) formulation, which is one of the most widely used linear programming (LP) formulations for solving the classification problem, can easily be generalized to the multigroup classification problem by sequentially solving for the optimal separating hyperplanes between the pairs of groups. One problem with this approach, however, is that the resulting classification rules may not cover each segment of the decision space. Moreover, the pairwise estimation of hyperplanes leaves much to be desired, because it may lead to suboptimal overall classification results.

Hence, the extension to more than two groups is difficult, if it requires the introduction of a multitude of binary variables; it is ad hoc, if the composite classification scheme is determined by separate pairwise analyses of the groups. In fact, some of the previously proposed MP formulations are designed specifically for the two-group case, and cannot easily be generalized to more than two groups. Our paper provides a formulation which is applicable to the general multigroup classification problem, and is similar to the LINMAP approach for problems in multidimensional analysis of preferences (Srinivasan and Shocker 1973). We next introduce the model formulation.

Subject classification: Programming, Linear applications. Statistics; nonparametric, discriminant analysis.
Area of review: Optimization.

0030-364X/97/4502-0213 $05.00

1. THE BASIC MODEL

Consider a finite set $S = \{1, \ldots, s\}$ of populations (groups) of objects, with each object belonging to one and only one of the groups. Samples of size n_j, $j \in S$, are available from these groups, and the group membership of each sample object in the training sample is known. Let $N = \Sigma_{j \in S} n_j$ be the total sample size, and $P_j = \{1, \ldots, n_j\}$ the set of sample objects belonging to group $j, j \in S$. Each object i with either unknown or unspecified group membership is characterized by a set of K attributes contained in the $(K + 1)$-dimensional column vector $\mathbf{x}_i = (x_{i0}, x_{i1}, \ldots, x_{iK})^T$, where $x_{i0} = 1$. Denote the attribute vector for object i with known membership in group j, i.e., $i \in P_j$, by $\mathbf{x}_{ij} = (x_{ij0}, x_{ij1}, \ldots, x_{ijK})^T$, where $x_{ij0} = 1$.

We will estimate s $(K + 1)$-dimensional row vectors $\boldsymbol{\alpha}^j = (\alpha_{j0}, \alpha_{j1}, \ldots, \alpha_{jK})$, and determine linear classification scores $\boldsymbol{\alpha}^j \mathbf{x}_i$ $(j = 1, \ldots, s)$ for any object i with respect to group j, $j \in S$. The classification decision rule is to classify an object i into group m provided that

$$\boldsymbol{\alpha}^m \mathbf{x}_i = \underset{j \in S}{\text{Max}}\{\boldsymbol{\alpha}^j \mathbf{x}_i\}. \tag{1.1}$$

Hence, (1.1) assigns an object to the group for which it attains the highest classification score. The classification rule in (1.1) is comparable to the Bayesian approach where an object would be assigned to a group based on the highest posterior probability of group membership for the given vector of attributes (Anderson 1984, Johnson and Wichern 1988), or to Fisher's (1936) classical approach where group membership is determined by distances derived from linear classification scores.

The vectors $\boldsymbol{\alpha}^j, j \in S$, will be determined in a way such that the decision rule in (1.1) operates "optimally" on the sample objects according to a criterion which will be defined below, combining measures of the "goodness" and "badness" of the fit. Let us use the notation $S_{-j} = S \backslash \{j\}$ to denote the set of all groups except group j, and represent any real-valued scalar y by $y = y^+ - y^-$, where $y^+ = \text{Max}\{0; y\}, y^- = -\text{Min}\{0; y\}$. Then, the goodness of fit in the training sample for object $i \in P_r, r \in S$, can be measured by $G^i_{rj}(\boldsymbol{\alpha}^r, \boldsymbol{\alpha}^j)$ in (1.2), in which the classification score $\boldsymbol{\alpha}^r \mathbf{x}_{ir}$ of object i with respect to its own group r is pairwise compared with classification scores $\boldsymbol{\alpha}^j \mathbf{x}_{ir}$ of this object with respect to the remaining groups $j \in S_{-r}$:

$$G^i_{rj}(\boldsymbol{\alpha}^r, \boldsymbol{\alpha}^j) = (\boldsymbol{\alpha}^r \mathbf{x}_{ir} - \boldsymbol{\alpha}^j \mathbf{x}_{ir})^+, \quad i \in P_r, j \in S_{-r}, r \in S. \tag{1.2}$$

Obviously, we prefer strictly positive values for $G^i_{rj}(\boldsymbol{\alpha}^r, \boldsymbol{\alpha}^j)$, and larger values are better. Likewise, the badness of fit for object $i \in P_r$ with respect to group j can be defined as in (1.3):

$$B^i_{rj}(\boldsymbol{\alpha}^r, \boldsymbol{\alpha}^j) = (\boldsymbol{\alpha}^r \mathbf{x}_{ir} - \boldsymbol{\alpha}^j \mathbf{x}_{ir})^-, \quad i \in P_r, j \in S_{-r}, r \in S, \tag{1.3}$$

where smaller values of $B^i_{rj}(\boldsymbol{\alpha}^r, \boldsymbol{\alpha}^j)$ are preferred, and ideally $B^i_{rj}(\boldsymbol{\alpha}^r, \boldsymbol{\alpha}^j) = 0$. The aggregate goodness and badness of object $i \in P_r$ are given by $G^i_r(\boldsymbol{\alpha})$ and $B^i_r(\boldsymbol{\alpha})$ in (1.4) and (1.5), respectively:

$$G^i_r(\boldsymbol{\alpha}) = G^i_r(\boldsymbol{\alpha}^1, \ldots, \boldsymbol{\alpha}^s) = \sum_{j \in S_{-r}} G^i_{rj}(\boldsymbol{\alpha}^r, \boldsymbol{\alpha}^j), \quad i \in P_r, r \in S, \tag{1.4}$$

$$B^i_r(\boldsymbol{\alpha}) = B^i_r(\boldsymbol{\alpha}^1, \ldots, \boldsymbol{\alpha}^s) = \sum_{j \in S_{-r}} B^i_{rj}(\boldsymbol{\alpha}^r, \boldsymbol{\alpha}^j), \quad i \in P_r, r \in S. \tag{1.5}$$

Thus, the goodness and badness of all objects i in group r combined are given by $G_r(\boldsymbol{\alpha})$ and $B_r(\boldsymbol{\alpha})$ in (1.6) and (1.7), respectively:

$$G_r(\boldsymbol{\alpha}) = G_r(\boldsymbol{\alpha}^1, \ldots, \boldsymbol{\alpha}^s) = \sum_{i \in P_r} G^i_r(\boldsymbol{\alpha}^1, \ldots, \boldsymbol{\alpha}^s) = \sum_{i \in P_r} \sum_{j \in S_{-r}} G^i_{rj}(\boldsymbol{\alpha}^r, \boldsymbol{\alpha}^j), \quad r \in S, \tag{1.6}$$

$$B_r(\boldsymbol{\alpha}) = B_r(\boldsymbol{\alpha}^1, \ldots, \boldsymbol{\alpha}^s) = \sum_{i \in P_r} B^i_r(\boldsymbol{\alpha}^1, \ldots, \boldsymbol{\alpha}^s) = \sum_{i \in P_r} \sum_{j \in S_{-r}} B^i_{rj}(\boldsymbol{\alpha}^r, \boldsymbol{\alpha}^j), \quad r \in S. \tag{1.7}$$

Finally, measures of total goodness $G(\boldsymbol{\alpha})$ and total badness $B(\boldsymbol{\alpha})$ for all groups $r \in S$ are given by (1.8) and (1.9):

$$G(\boldsymbol{\alpha}) = G(\boldsymbol{\alpha}^1, \ldots, \boldsymbol{\alpha}^s) = \sum_{r \in S} G_r(\boldsymbol{\alpha}) = \sum_{r \in S} \sum_{j \in S_{-r}} \sum_{i \in P_r} G^i_{rj}(\boldsymbol{\alpha}^r, \boldsymbol{\alpha}^j), \tag{1.8}$$

$$B(\boldsymbol{\alpha}) = B(\boldsymbol{\alpha}^1, \ldots, \boldsymbol{\alpha}^s) = \sum_{r \in S} B_r(\boldsymbol{\alpha}) = \sum_{r \in S} \sum_{j \in S_{-r}} \sum_{i \in P_r} B^i_{rj}(\boldsymbol{\alpha}^r, \boldsymbol{\alpha}^j). \tag{1.9}$$

The measures of total goodness and badness of fit in (1.8) and (1.9) are conceptually similar to the "internal" and "external" deviations previously introduced by several researchers for the two-group case (Freed and Glover 1986a, Glover et al. 1988, Glover 1990, Joachimsthaler and Stam 1990).

Clearly, by definition $G(\boldsymbol{\alpha})$ and $B(\boldsymbol{\alpha})$ are nonnegative for any $\boldsymbol{\alpha}$. The trivial solution where $\boldsymbol{\alpha}^r = \boldsymbol{\alpha}^*$, for all $r \in S$, generates $G(\boldsymbol{\alpha}) = B(\boldsymbol{\alpha}) = 0$ but does not contain any useful information in terms of classification power, as any object can be classified arbitrarily into any of the s groups. Hence, we need to rule out the trivial solution by a proper normalization. Also, a solution $\boldsymbol{\alpha}$ for which $G(\boldsymbol{\alpha}) - B(\boldsymbol{\alpha}) < 0$, i.e., a solution for which the total badness exceeds the total goodness of the fit, will in general not be satisfactory (see, e.g., Glover 1990). It can easily be verified that for any $\boldsymbol{\alpha}$, $G(\boldsymbol{\alpha}) = B(-\boldsymbol{\alpha})$ holds, so it follows that for any solution $\boldsymbol{\alpha}$ with $G(\boldsymbol{\alpha}) - B(\boldsymbol{\alpha}) = -q < 0$, $G(-\boldsymbol{\alpha}) - B(-\boldsymbol{\alpha}) = q > 0$. Hence, undesirable solutions with total badness exceeding total goodness can easily be ruled out in our proposed formulation, by using the normalization given in (1.10):

$$G(\boldsymbol{\alpha}) - B(\boldsymbol{\alpha}) = q, \tag{1.10}$$

where q is any strictly positive constant. Using this condition, we preclude solutions for which $G(\boldsymbol{\alpha}) - B(\boldsymbol{\alpha}) < 0$ and the trivial solution $\boldsymbol{\alpha}^r = \boldsymbol{\alpha}^*$, for all $r \in S$. From the definition of $G(\boldsymbol{\alpha})$ and $B(\boldsymbol{\alpha})$, (1.8) and (1.9), and the property that $y = y^+ - y^-$, it follows that the difference between $G(\boldsymbol{\alpha})$ and $B(\boldsymbol{\alpha})$ is a linear function in $\boldsymbol{\alpha}$, i.e.,

$$G(\boldsymbol{\alpha}) - B(\boldsymbol{\alpha}) = \sum_{r \in S} \sum_{j \in S_{-r}} \sum_{i \in P_r} (\boldsymbol{\alpha}^r - \boldsymbol{\alpha}^j)\mathbf{x}_{ir}.$$

The normalization in (1.10) will be investigated in more detail below. We next state the complete linear program LP^q, which determines the $\boldsymbol{\alpha}$-vectors which minimize the total badness, subject to the normalization in (1.10). The superscript q in LP^q refers to the right-hand-side value used in the normalization constraint.

Program LP^q: Min $B(\boldsymbol{\alpha})$ (1.11)

Subject to:

$$G(\boldsymbol{\alpha}) - B(\boldsymbol{\alpha}) = q \quad (1.10)$$

$\boldsymbol{\alpha}$ unrestricted in sign. (1.12)

Due to the relationship between $B(\boldsymbol{\alpha})$ and $G(\boldsymbol{\alpha})$ discussed above and propositions to be introduced later, the normalization in (1.10) does not preclude any useful classification solution from consideration, and only scales the optimal solution through the choice of the constant q (see also Proposition 6). Program LP^q can be restated as LP^q-A by explicitly introducing a set of variables β_{rj}^i and γ_{rj}^i, representing the badness $B_{rj}^i(\boldsymbol{\alpha}^r, \boldsymbol{\alpha}^j)$ and goodness $G_{rj}^i(\boldsymbol{\alpha}^r, \boldsymbol{\alpha}^j)$ of object $i \in P_r$ with respect to group $j \in S_{-r}$, respectively:

Program LP^q-A: Min $\sum_{r \in S} \sum_{j \in S_{-r}} \sum_{i \in P_r} \beta_{rj}^i$ (1.13)

Subject to:

$$\beta_{rj}^i + (\boldsymbol{\alpha}^r - \boldsymbol{\alpha}^j)\mathbf{x}_{ir} - \gamma_{rj}^i = 0,$$

for all $i \in P_r$, $j \in S_{-r}$, $r \in S$, (1.14)

$$\sum_{r \in S} \sum_{j \in S_{-r}} \sum_{i \in P_r} (\gamma_{rj}^i - \beta_{rj}^i) = q, \quad (1.15)$$

$$\beta_{rj}^i, \gamma_{rj}^i \geq 0,$$

for all $i \in P_r$, $j \in S_{-r}$, $r \in S$. (1.16)

Formulation LP^q-A is similar in concept to the Hybrid model previously proposed for the two-group case (Glover et al. 1988, Glover 1990), with the omission of the minimax deviations from their general model framework. The favorable classification results for two groups reported in a recent simulation study involving (among others) several variants of the Hybrid model (Duarte Silva and Stam 1994) indicate that our proposed multigroup formulation may give good classification results as well. Some important theoretical properties of our formulation will be derived and discussed in the next section.

From (1.14) it is clear that the $\boldsymbol{\alpha}$ vectors are determined relative to each other. Consequently, one of the $\boldsymbol{\alpha}$ vectors can be set equal to the null vector, without loss of generality (see Proposition 7).

Table I
Data for Examples 1 and 3

		Example 1		Example 3	
	Object i	x_{ij1}	x_{ij2}	x_{ij1}	x_{ij2}
Group 1	1	1	2	2	3
$j = 1$	2	2	3	5	6
	3	2	4	3	8
	4	—	—	2	1
Group 2	1	4	2	4	6
$j = 2$	2	2.5	3	1	3
	3	2.5	4	4	4.5
Group 3	1	1	1	2	4
$j = 3$	2	2	2.5	4	5
	3	4	1	—	—

Using w_{irj} as the dual variable associated with the constraint for object $i \in P_r$, $j \in S_{-r}$, $r \in S$, and w as the dual variable for the normalization constraint, the dual linear program DP^q-A of LP^q-A can be written as in (1.17)–(1.20):

Program DP^q-A: Max qw (1.17)

Subject to:

$$\sum_{j \in S_{-r}} \sum_{i \in P_r} x_{irk} w_{irj} - \sum_{j \in S_{-r}} \sum_{i \in P_j} x_{ijk} w_{ijr} + \left(\sum_{j \in S_{-r}} \sum_{i \in P_r} x_{irk} - \sum_{j \in S_{-r}} \sum_{i \in P_j} x_{ijk} \right) w = 0,$$

$k = 0, 1, \ldots, K$, and for all $r \in S$. (1.18)

$$0 \leq w_{irj} \leq 1,$$

for all $i \in P_r$, $j \in S_{-r}$, $r \in S$, (1.19)

w unrestricted in sign. (1.20)

From a computational viewpoint, the dual program DP^q-A is quite attractive, since the simplex method with bounded variables can be used to solve it, and DP^q-A contains only a relatively small number of proper constraints. While the primal problem in LP^q-A has $(s - 1)N + 1$ constraints, DP^q-A has only $s(K + 1)$ proper constraints, the remainder being upper bounds on the variables.

Example 1 is a very simple constructed data set with three groups and two (proper) attributes. Table I provides the data both for Example 1 and for Example 3, which is a special case that will be discussed in Section 2.

Table II presents optimal vectors $\boldsymbol{\alpha}^j$ obtained for Examples 1 and 3 from solving LP^q-A with $q = 10$. It should be noted that there may be alternative solutions, especially in those examples where complete linear separation of two or more groups is possible.

Example 1. The interpretation of Example 1 is straightforward, as all of the $\boldsymbol{\alpha}^j$ vectors are different. The hyperplanes which pairwise separate groups h and j are constructed by setting $\boldsymbol{\alpha}^h\mathbf{x} = \boldsymbol{\alpha}^j\mathbf{x}$, $h, j \in S$. Since the

Table II
Solution Vectors for Examples 1 and 3

Solution Vector α^j	Example 1			Example 3		
	x_{ij0}	x_{ij1}	x_{ij2}	x_{ij0}	x_{ij1}	x_{ij2}
α^1	0	0	0	3.333	0	0
α^2	−3.509	7.018	−3.509	3.333	0	0
α^3	10.526	5.848	−8.187	0	0	0
Objective Value		0			13.3333	

example has only two proper attributes (x_1, x_2), the hyperplanes are lines in $\mathbf{R}^2$. After rescaling, this leads to the following separating hyperplanes:

(1) Line separating groups 1 and 2: $2x_1 - x_2 = 1$,
(2) Line separating groups 1 and 3: $5x_1 - 7x_2 = -9$,
(3) Line separating groups 2 and 3: $x_1 + 4x_2 = 12$.

The sample points and lines of separation for Example 1 are depicted graphically in Figure 1. This figure shows that the data in this example are perfectly linearly separable, since none of the objects is misclassified. However, the separating hyperplanes do pass through three of the data points, so that the classification of these objects is ambiguous. We will discuss the implications of this issue later in Section 2.5, and will propose a slightly modified problem formulation (the ϵ-procedure) which deals with this issue.

In the next section, we derive a number of properties of program LP^q, which will provide further justification for the choice of objective function and normalization in this formulation, and to establish the usefulness of LP^q for analyzing the multigroup classification problem. A number of these properties are generalizations of similar properties previously derived, discussed and analyzed for the two-group case by, among others, Freed and Glover (1986) and Koehler (1989a, 1989b, 1990, 1991).

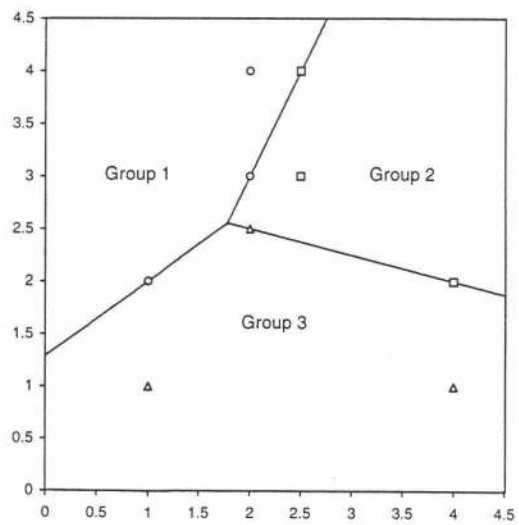

Figure 1. Classification scheme for example 1.

2. PROPERTIES OF THE BASIC MODEL

2.1. Sequential Separation

In this section, we first study the phenomenon that the classification vectors for at least two groups coincide. This situation may occur frequently in practice. Even though the normalization in (1.10) prevents that all α^j vectors are identical, it is possible that the α^j vectors in one or more subsets of S are the same. In general, suppose that LP^q-A generates an optimal solution with a partition $S_1, S_2, \ldots, S_\delta$ of S such that for all pairs $m, r \in S$, (2.1) holds:

$$m \in S_h \quad \text{and} \quad r \in S_h,\ h \in \{1, \ldots, \delta\} \Leftrightarrow \alpha^m = \alpha^r. \tag{2.1}$$

If every subset S_i is a singleton, it follows that $\delta = s$, so that condition (2.1) does not apply and we get a solution of LP^q-A where all α^j are different. If at least one subset, say S_h, contains at least two elements, e.g., m and r, then $\alpha^m - \alpha^r = \mathbf{0}$, and no separation between groups m and r is possible. A new object with attribute vector $\mathbf{x}_i$ and $\alpha^m\mathbf{x}_i = \alpha^r\mathbf{x}_i = \max_{j \in S}\{\alpha^j\mathbf{x}_i\}$ cannot be classified at this stage. In fact, this situation can occur even if perfect linear separation of groups m and r is possible, as Example 2 below will show.

In order to overcome this problem, a new linear program is solved for each subset S_h containing more than one group. This LP uses only the sample data of the groups belonging to S_h. The (incomplete) classification information from previous iterations is retained, and remains part of the final classification scheme. This process is continued until all subsets contain exactly one group, i.e., until all groups are separated. Such a process must necessarily terminate after solving at most $s - 1$ LPs, unless for a subset S_h containing at least two groups the conditions of Proposition 2 below hold. In that (unlikely) case, groups belonging to S_h cannot (and should not) be separated. Successive divisions can be represented by a tree structure, as we will show in Example 2.

Example 2. In Example 2 we solve a constructed five-group classification problem with two proper attributes. The data, optimal α-vectors and successive partitions for this example are provided in Table III. The problem and the final classification scheme are shown graphically in Figure 2.

We use the sequential separation procedure to determine the optimal classification rules for the five groups. Let us denote the α-vector associated with group r computed in iteration p by α^{pr}. Table III shows that solving the full model with all five groups yields an optimal solution where $\alpha^{11} = \alpha^{12} = \alpha^{13} = \alpha^{14} = \mathbf{0}^T$, and $\alpha^{15} = (2.907, -2.907, -2.907)^T$, leading to hyperplane (1) in Figure 2,

Table III
Data, Solutions and Successive Divisions for Example 2 (Five Groups)

Data

	Group 1		Group 2		Group 3		Group 4		Group 5	
Object i	x_{ij1}	x_{ij2}	x_{ij1}	x_{ij2}	x_{ij1}	x_{ij2}	x_{ij1}	x_{ij2}	x_{ij1}	x_{ij2}
1	0	2.6	0	1	1	1	1	0	0	0
2	1	3	0	2	2.4	2	2	0	1	0
3	1.4	1.8	1	2	3	0.6	2	0.6	0	1
4	1.4	2.6	2	2	3	1.4	3	0	0.4	0.4

Successive Divisions and Solutions

Iteration	Optimal Solution				
	S_1				S_2
1	$\boldsymbol{\alpha}^{11}$	$\boldsymbol{\alpha}^{12}$	$\boldsymbol{\alpha}^{13}$	$\boldsymbol{\alpha}^{14}$	$\boldsymbol{\alpha}^{15}$
	(0, 0, 0)	(0, 0, 0)	(0, 0, 0)	(0, 0, 0)	(2.907, −2.907, −2.907)
	S_3			S_4	
2	$\boldsymbol{\alpha}^{21}$	$\boldsymbol{\alpha}^{22}$	$\boldsymbol{\alpha}^{23}$	$\boldsymbol{\alpha}^{24}$	—
	(0, 0, 0)	(0, 0, 0)	(0, 2.358, −2.358)	(0, 2.358, −2.358)	—
3	$\boldsymbol{\alpha}^{31}$	$\boldsymbol{\alpha}^{32}$	$\boldsymbol{\alpha}^{33}$	$\boldsymbol{\alpha}^{34}$	—
	(0, 0, 0)	(66.667, 0, −33.333)	(0, 0, 0)	(13.636, 0, −22.727)	—

separating group 5 from the other groups. Hence, the partition of S consists of (S_1, S_2), where $S_1 = \{1, 2, 3, 4\}$, and $S_2 = \{5\}$. A second iteration is required to separate the four groups in S_1, resulting in the $\boldsymbol{\alpha}$-vectors $\boldsymbol{\alpha}^{21} = \boldsymbol{\alpha}^{22} = \mathbf{0}^T$, and $\boldsymbol{\alpha}^{23} = \boldsymbol{\alpha}^{24} = (0, 2.358, -2.358)^T$, thus yielding clusters $S_3 = \{1, 2\}$ and $S_4 = \{3, 4\}$ and hyperplane (2) which separates S_3 and S_4. In the third iteration it remains to solve two more linear programs, one to separate the groups in S_3, giving $\boldsymbol{\alpha}^{31} = \mathbf{0}^T$, $\boldsymbol{\alpha}^{32} = (66.667, 0, -33.333)^T$ and hyperplane (3), and another for S_4, resulting in $\boldsymbol{\alpha}^{33} = \mathbf{0}^T$, $\boldsymbol{\alpha}^{34} = (13.636, 0, -22.727)^T$ and hyperplane (4), which completes the process of successive partitioning the groups. The process of successive divisions can be represented by the tree structure as in Figure 3.

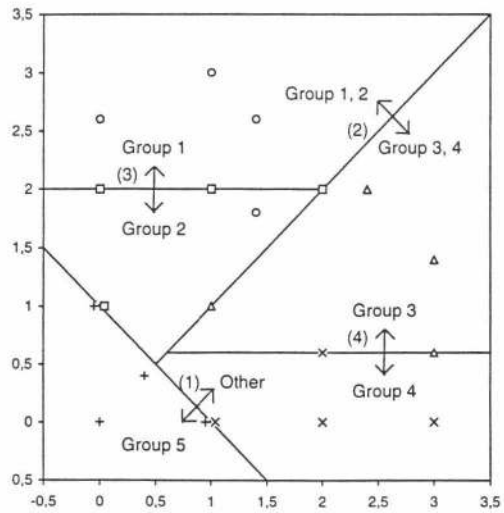

Figure 2. Classification scheme for example 2 (five-group problem).

2.2. Existence of Solutions

We next study the existence of solutions to LP^q, and show that this formulation guarantees a finite optimal solution, unless the left-hand side of the normalization constraint (1.10) is identical to zero. Without loss of generality, we will refer to the generic formulation LP^q, rather than to the equivalent formulation LP^q-A.

Proposition 1. *Program* LP^q *has a finite optimal solution for any* $q > 0$ *if and only if there exists at least one* $\boldsymbol{\alpha}$ *for which* $G(\boldsymbol{\alpha}) - B(\boldsymbol{\alpha}) \neq 0$.

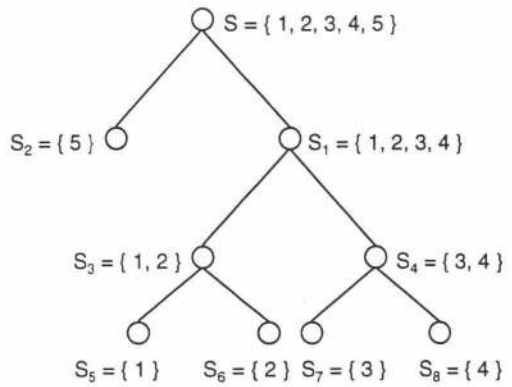

Figure 3. Tree structure and branching in the presence of coinciding **$\boldsymbol{\alpha}$-vectors for example 2.**

Proof. "⇒": It is obvious that a finite optimal solution to LP^q implies that $G(\alpha) - B(\alpha) = q \neq 0$, since $q > 0$.

"⇐": Checking LP^q, we observe that the formulation always has a finite optimal solution, if feasible. Arbitrarily select a vector α for which $G(\alpha) - B(\alpha) = p > 0$. From (1.14) and (1.15) it follows that $\beta_{rj}^{i} = \text{Max}\{0, -(\alpha^r - \alpha^j)\mathbf{x}_{ir}\}$, $i \in P_r$, $j \in S_{-r}$, $r \in S$. Hence, $\alpha^* = p^{-1}q\alpha$ and $\beta^* = p^{-1}q\beta$ form a feasible solution to LP^q, which completes the proof. □

Whenever $G(\alpha) - B(\alpha) = 0$ for all α, the linear program LP^q is infeasible and does not provide a solution to the discriminant problem. The next proposition shows under which data conditions this will happen.

Proposition 2. *$G(\alpha) - B(\alpha) = 0$, for all α, if and only if $\Sigma_{i\in P_r} \mathbf{x}_{ir} = \mathbf{x}^*$, for all $r \in S$.*

Proof.

$$G(\alpha) - B(\alpha) = \sum_{r\in S}\sum_{j\in S_{-r}}\sum_{i\in P_r}(\alpha^r - \alpha^j)\mathbf{x}_{ir}$$

$$= \sum_{r\in S}\left[\sum_{j\in S_{-r}}\sum_{i\in P_r}\alpha^r\mathbf{x}_{ir} - \sum_{j\in S_{-r}}\sum_{i\in P_r}\alpha^j\mathbf{x}_{ir}\right]$$

$$= \sum_{r\in S}\left[(s-1)\sum_{i\in P_r}\alpha^r\mathbf{x}_{ir} - \sum_{j\in S_{-r}}\sum_{i\in P_r}\alpha^j\mathbf{x}_{ir}\right]$$

$$= \sum_{r\in S}\left[s\sum_{i\in P_r}\alpha^r\mathbf{x}_{ir} - \sum_{j\in S}\sum_{i\in P_r}\alpha^j\mathbf{x}_{ir}\right] = s\sum_{r\in S}\sum_{i\in P_r}\alpha^r\mathbf{x}_{ir}$$

$$- \sum_{r\in S}\sum_{j\in S}\sum_{i\in P_r}\alpha^j\mathbf{x}_{ir} = s\sum_{r\in S}\sum_{i\in P_r}\alpha^r\mathbf{x}_{ir} - \sum_{r\in S}\sum_{j\in S}\sum_{i\in P_j}\alpha^r\mathbf{x}_{ij}$$

$$= \sum_{r\in S}\alpha^r\left[s\sum_{i\in P_r}\mathbf{x}_{ir} - \sum_{j\in S}\sum_{i\in P_j}\mathbf{x}_{ij}\right].$$

Hence, it follows that $G(\alpha) - B(\alpha) = 0$, for all α, if and only if

$$s\sum_{i\in P_r}\mathbf{x}_{ir} - \sum_{j\in S}\sum_{i\in P_j}\mathbf{x}_{ij} = \mathbf{0}, \quad \text{for } r \in S,$$

which condition can be rewritten as

$$\sum_{i\in P_r}\mathbf{x}_{ir} = \frac{1}{s}\sum_{j\in S}\sum_{i\in P_j}\mathbf{x}_{ij} = \mathbf{x}^*, \quad r \in S,$$

from which the proposition follows. □

Since x_{ir0} equals one for all $i \in P_r$, $r \in S$, it is obvious that $\Sigma_{i\in P_r} x_{ir0} = n_r$ will be identical across all groups $r \in S$ if and only if $n_j = n_h$, for all $j, h \in S$. Thus, restating Proposition 2 in a more concrete way, it shows that LP^q ($q \neq 0$) provides no feasible solution to the discriminant problem, if and only if (1) all sample sizes n_j are equal, and (2) the sum, and hence the mean, for each attribute is the same across all groups. It is unlikely that any real data set will ever satisfy these conditions. Interestingly, the parametric Bayesian approach with multivariate normal groups (Anderson 1984) and equal covariance matrices for the different groups fails to provide a solution under exactly the same conditions as those in Proposition 2, provided that the prior probability of group membership is estimated by the sample size proportions, and the group means are estimated by the respective sample means. The situation of equal sample means but unequal sample sizes is discussed in Propositions 3–6.

To correct for the different sample sizes, weights could be introduced into the normalization restriction (1.10). Probably the most justified weighted normalization would be the one given in (2.2), where the contributions to $G(\alpha) - B(\alpha)$ by observations in each group are weighted by the group's sample size,

$$\sum_{r\in S}\sum_{j\in S_{-r}}\sum_{i\in P_r} n_j(\alpha^r - \alpha^j)\mathbf{x}_{ir} = q. \tag{2.2}$$

For the case of two groups, this expression simplifies to the normalization recently proposed by Glover (1990). Proposition 2 continues to hold, provided that $\Sigma_{i\in P_r}\mathbf{x}_{ir}$ is replaced by $\frac{1}{n_r}\Sigma_{i\in P_r}\mathbf{x}_{ir}$, i.e., instead of conditions (1) and (2) below Proposition 2, the requirement for this proposition now is that the sample means for each attribute should be the same across groups. Thus, LP^q will fail to yield a feasible solution if and only if the sample means on all the attributes are equal, independent of the sample sizes. In this paper we do not investigate the use of (2.2) in LP^q further.

In Proposition 2 we derived that LP^q-A has *no* feasible solution if and only if the training sample means for the attributes are identical across all groups and the training sample sizes of all groups are identical. Under these data conditions, none of the objects will be classified into a group. We will next discuss a related special case which will rarely occur in practice, but which is nevertheless of theoretical interest. Proposition 3 shows that if the sample means of each attribute are equal across all groups, but the sample sizes are not all identical, then there exists an optimal solution for which only the constant terms α_{r0} can possibly be nonzero.

Proposition 3. *If $\bar{x}_{jk} = 1/n_j \times \Sigma_{i\in P_j} x_{ijk} = \bar{x}_k$, for all $j \in S$, $k = 1, \ldots, K$, and not all sample sizes n_j are equal, then there exists an optimal solution to LP^q-A, say $(\alpha^*, \beta^*, \gamma^*)$, such that $\alpha_{rk}^* = 0$, $k = 1, \ldots, K$, and $r \in S$.*

Proof. See the appendix.

The result of Proposition 3 is logically consistent, because one cannot expect to construct a meaningful *linear* classification function separating the groups, when the sample mean of each attribute is identical across all groups. The next proposition shows that it is possible to *directly determine* an optimal solution to LP^q-A if the conditions of Proposition 3 hold. Moreover, Proposition 4 shows in this case there exists a classification scheme which depends on the group sample sizes only. Without loss of generality, we assume that $n_1 \geq n_2 \geq \cdots \geq n_s$, with $n_1 > n_s$, i.e., we order the groups according to their size. Let $m_r = sn_r - \Sigma_{j\in S} n_j$. Then Proposition 4 is stated as follows:

Proposition 4. *Under the conditions of Proposition* 3, *an optimal solution to* LP^q-A *is given by* $\alpha^*_{jk} = 0, k = 1, \ldots, K, j \in S$; $\alpha^*_{j0} = 0, j = r^* + 1, \ldots, s$; $\alpha^*_{j0} = q/\sum_{i=1}^{r^*} m_i, j = 1, \ldots, r^*$, *where* r^* *is such that*

$$\min_{r=1,\ldots,s-1} \frac{r \sum_{j=r+1}^{s} n_j}{\sum_{i=1}^{r} m_i} = \text{Min } M_r = M^*$$

holds for $r = r^*$.

Proof. See the appendix.

Proposition 4 implies that if the sample mean for each proper attribute is the same across all groups, but the group sample sizes are not all identical, there exists an optimal solution to LP^q-A where the estimated coefficients of all proper attributes are identical to zero for all groups, while the estimated coefficient of the constant term is either a positive constant equal to α^*_{j0}, for $j = 1, \ldots, r^*$, or zero, for $j = r^* + 1, \ldots, s$. In this case any object will be classified into the cluster of groups $S_1 = \{1, \ldots, r^*\}$, and never into any of the groups in $S_2 = \{r^* + 1, \ldots, s\}$.

As remarked above, no immediate separation of objects within S_1 is possible. Further application of the sequential separation procedure described in Section 2.1 will of course eventually lead to the classification of all objects into the group with the largest training sample size. However, this phenomenon is not a shortcoming of our method. Under the data conditions of Proposition 4, the Bayesian approach will also classify all objects into the largest group of the training sample, as long as the prior group membership probabilities are proportional to the sample sizes. In the case of equal priors across groups, the posterior probabilities obtained using the Bayesian approach will be identical for each group, in other words, none of the objects will be classified into any group. Similarly, Proposition 2 states that LP^q-A does not classify any object into a group if the training sample sizes of all groups are identical.

We can extend Proposition 3 to Proposition 5, the proof of which is conceptually similar to that of Proposition 3.

Proposition 5. *Suppose that for some attribute* $t \in \{1, \ldots, K\}, x_{irt} \geq 0$, *for all* $i \in P_r, r \in S$. *If* $\tilde{x}_{jk} = \tilde{x}_k, j \in S$ *and* $k = 1, \ldots, K$, *then there exists an optimal solution to* LP^q-A, *say* $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*)$, *such that* $\alpha^*_{rk} = 0$ *and* $r \in S, k = 0, 1, \ldots, t-1, t+1, \ldots, K$.

Proof. See the appendix.

Compared with Proposition 3, the only additional requirement in Proposition 5 is that there exists an attribute $t \in \{0, 1, \ldots, K\}$ for which the training sample values x_{irt} are all nonnegative. However, Proposition 8 below shows that the training sample data can easily be transformed such that $x_{irt} \geq 0$ for all $i \in P_r, r \in S$, and any $t \in \{1, \ldots, K\}$. Moreover, this nonnegativity restriction is always satisfied for $t = 0$, since the constant terms x_{ir0} equal one for each object. Therefore, Proposition 3 is a special case of Proposition 5.

Proposition 5 shows that if the sample mean for each proper attribute (i.e., for each $k \in \{1, \ldots, K\}$) is identical across groups, but the group sample sizes are not all the same, an optimal solution exists in which the estimated values of all but one of the attribute coefficients for all groups are zero, including the coefficients of the constant terms. The nature of the optimal solution α^*_{rt} (see (A.18) in the appendix) implies that in this situation objects will always be classified into the group, say w, which has the largest coefficient $\alpha^*_{MAX} = \alpha^*_{wt}$. However, the α^*_{MAX} value may occur for more than one group, in which case the classification rule assigns each observation to the subset S^{MAX} of groups with maximal α_{it}-value. Objects are never assigned to groups with smaller α_{rt}-values, so that under the—admittedly exceptional—data condition of Propositions 3–5, alternative procedures yielding nonlinear (e.g., quadratic) classification rules are required. We relegate the extension of our linear method to the nonlinear case to future research.

Example 3 illustrates the special case described in Propositions 3–5.

Example 3. Example 3 has three groups and two proper attributes. The data and optimal solution for this example are given in Tables I and II. In the optimal solution for this example, both proper attributes x_1 and x_2 have a zero coefficient for all three groups. This implies that the only useful information from the sample data, according to the LP model, is contained in the number of training sample objects from each group. Since $\alpha^1_0 = \alpha^2_0 = 3.333$ and $\alpha^3_0 = 0$, all sample objects are classified into the cluster consisting of groups 1 and 2. No object will ever be classified into group 3, and the sequential separation procedure is needed in order to further distinguish between groups 1 and 2. Since the sample means of both attributes are identical across all groups, but the sample sizes are not ($n_1 = 4, n_2 = 3$, and $n_3 = 2$), this result is a direct application of Propositions 3 and 4. It is interesting to verify the conditions in Proposition 4 which resulted in the initial separation scheme for this example. We use the training sample sizes n_j to calculate that $m_1 = 3, m_2 = 0$ and $m_3 = -3$, so that $M_1 = 5/3$ and $M_2 = 4/3$. The minimum value M^* over $r \in \{1, \ldots, s-1\} = \{1, 2\}$ is M_2, and $r^* = 2$, resulting in an initial separation of group 3 from groups 1 and 2.

2.3. Scaling, Linear Transformations, and Index of Fit

The next property shows that the particular choice of the positive constant q affects only the scaling of the problem.

Proposition 6. *Let* $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*)$ *be an optimal solution to program* LP^q, *with objective function value* v^q_1. *Then, for any* $t > 0$, $(\boldsymbol{\alpha}^{**}, \boldsymbol{\beta}^{**}, \boldsymbol{\gamma}^{**})$ *with* $\boldsymbol{\alpha}^{**} = tq^{-1}\boldsymbol{\alpha}^*$, $\boldsymbol{\beta}^{**} = tq^{-1}\boldsymbol{\beta}^*$ *and* $\boldsymbol{\gamma}^{**} = tq^{-1}\boldsymbol{\gamma}^*$ *is an optimal solution to* LP^t, *with objective function value* $tq^{-1}v^q_1$.

Proof. Let v_1^t the objective function value for the solution $(\boldsymbol{\alpha}^{**}, \boldsymbol{\beta}^{**}, \boldsymbol{\gamma}^{**})$ to LP^t. By the construction of $(\boldsymbol{\alpha}^{**}, \boldsymbol{\beta}^{**}, \boldsymbol{\gamma}^{**})$, it immediately follows that $v_1^t = tq^{-1}v_1^q$. If $(\boldsymbol{\alpha}^{**}, \boldsymbol{\beta}^{**}, \boldsymbol{\gamma}^{**})$ is not optimal in LP^t, then there exists a solution $(\boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}')$ with objective value v_2^t such that $v_2^t < v_1^t$. However, consider a solution $(\boldsymbol{\alpha}'', \boldsymbol{\beta}'', \boldsymbol{\gamma}'')$ with $\boldsymbol{\alpha}'' = qt^{-1}\boldsymbol{\alpha}'$, $\boldsymbol{\beta}'' = qt^{-1}\boldsymbol{\beta}'$ and $\boldsymbol{\gamma}'' = qt^{-1}\boldsymbol{\gamma}'$, which is feasible in LP^q with objective function value $v_2^q = qt^{-1}v_2^t < qt^{-1}v_1^t = qt^{-1}tq^{-1}v_1^q = v_1^q$. This is obviously a contradiction. □

Given the decision rule in (1.1) to classify an object i into group m provided that $\boldsymbol{\alpha}^m\mathbf{x}_i = \text{Max}_{j\in S}\{\boldsymbol{\alpha}^j\mathbf{x}_i\}$, it is clear that the solutions $\boldsymbol{\alpha}^*$ and $\boldsymbol{\alpha}^{**}$ as defined in Proposition 6 are equivalent in that the classification results for both vectors will be identical, and q merely scales the optimal solution. It should be noted that if LP^q has alternative optimal solutions, then LP^t has corresponding alternative optimal solutions as well.

The model described so far assigns one vector $\boldsymbol{\alpha}^j$ to each group $j \in S$. Checking the structure of LP^q, it is obvious that there is redundancy in the number of variables in $\boldsymbol{\alpha} = (\boldsymbol{\alpha}^1, \ldots, \boldsymbol{\alpha}^s)$, as $G(\boldsymbol{\alpha})$ and $B(\boldsymbol{\alpha})$ are based on the pairwise differences between the $\boldsymbol{\alpha}$-vectors. The next proposition makes this redundancy explicit.

Proposition 7. *Let $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ be an optimal solution to* LP^q. *For any fixed vector $\boldsymbol{\alpha}^0 \in \mathbf{R}^{K+1}$, define $\boldsymbol{\eta}^j = \boldsymbol{\alpha}^j + \boldsymbol{\alpha}^0$, for all $j \in S$. Then $(\boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ is also an optimal solution to* LP^q.

Proof. The pairwise difference $(\boldsymbol{\eta}^r - \boldsymbol{\eta}^j)$ is equal to $(\boldsymbol{\alpha}^r - \boldsymbol{\alpha}^0 - \boldsymbol{\alpha}^j + \boldsymbol{\alpha}^0) = (\boldsymbol{\alpha}^r - \boldsymbol{\alpha}^j)$, which reduces to the same pairwise differences as in the original formulation (1.14). Hence, if $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ is an optimal solution, $(\boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ is an optimal solution as well. □

By taking the vector $\boldsymbol{\alpha}^0$ in Proposition 7 equal to $-\boldsymbol{\alpha}^r$ for some $r \in S$, it follows that $\boldsymbol{\eta}^r = \mathbf{0}$, implying that any *one* of the vectors $\boldsymbol{\alpha}^j, j \in S$, can be set identically equal to zero without loss to the model. To preserve the symmetry of the original model form, however, we do not introduce this simplification in our paper.

An important consideration in the construction of methods for classification and discrimination is whether these methods are insensitive to rotation and/or translation of the data. To discuss this issue for our approach, we introduce some further notation. Let $\mathbf{x}_{ir}^T = (1, x_{ir1}, \ldots, x_{irK}) = (1, (\mathbf{x}_{ir}^R)^T)$, and $\boldsymbol{\alpha}^j = (\alpha_{j0}, \alpha_{j1}, \ldots, \alpha_{jK}) = (\alpha_{j0}, (\boldsymbol{\alpha}^{j,R})^T)$. Proposition 8 shows that the $\boldsymbol{\alpha}$-vectors of LP^q after a linear transformation of the data are themselves a linear transformation of the original solution, while the $\boldsymbol{\beta}$- and $\boldsymbol{\gamma}$-vectors remain unchanged.

Proposition 8. *Let $\mathbf{U}$ be a nonsingular $K \times K$ matrix, and $\mathbf{u}$ an arbitrary column vector in $\mathbf{R}^K$. Suppose that the $\mathbf{x}_{ir}$, $i \in P_r$, $r \in S$, are the original data and the transformed data are given by $\mathbf{x}_{ir}^{RR} = \mathbf{U}\mathbf{x}_{ir}^R + \mathbf{u}$, $i \in P_r$, $r \in S$. Consider program* LP^q-D, *the analogue to* LP^q-A *using the transformed data $\mathbf{x}_{ir}^{RR}$. If $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ with $\boldsymbol{\alpha} = (\boldsymbol{\alpha}^1, \ldots, \boldsymbol{\alpha}^s)$ and $\boldsymbol{\alpha}^j = (\alpha_{j0}, (\boldsymbol{\alpha}^{j,R})^T)$, $j \in S$, solves* LP^q-A, *then $(\boldsymbol{\zeta}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ solves* LP^q-D, *where $\boldsymbol{\zeta} = (\boldsymbol{\zeta}^1, \ldots, \boldsymbol{\zeta}^s)$, and $\boldsymbol{\zeta}^j = (\zeta_{j0}, (\boldsymbol{\zeta}^{j,R})^T)$, with $\zeta_{j0} = \alpha_{j0} - \boldsymbol{\alpha}^{j,R}\mathbf{U}^{-1}\mathbf{u}$ and $\boldsymbol{\zeta}^{j,R} = \boldsymbol{\alpha}^{j,R}\mathbf{U}^{-1}$.*

Proof. The stability theorem of Glover et al. (1988) can be applied directly to LP^q-A. A less direct proof can be constructed using duality theory of linear programming. □

One application of this proposition concerns solving the problem of standardized data. Let $\bar{x}_{jk}$ and s_{jk} denote the sample mean and standard deviation of attribute k in group j. Similarly, let $\bar{x}_k$ and s_k be the mean and (pooled) standard deviation of attribute k for all sample data. The original data can be standardized using the transformation in Proposition 8 by taking $\mathbf{U} = \mathbf{Diag}(s_k^{-1})$, i.e., a diagonal transformation matrix with the reciprocal of the pooled standard deviations on the main diagonal, and $\mathbf{u} = (u_1, \ldots, u_K)$, where $u_k = -\bar{x}_k s_k^{-1}$. According to Proposition 8, these standardized data generate a transformed problem with solution $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*)$, where $\alpha_{j0}^* = \alpha_{j0} + \sum_{k=1}^K \bar{x}_k\alpha_{jk}$, $\alpha_{jk}^* = s_k\alpha_{jk}$, $\boldsymbol{\beta}^* = \boldsymbol{\beta}$ and $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}$, $j \in S$, $k = 1, \ldots, K$. The coefficients α_{jk}^* can be used to identify the relative importance of the different attributes. As shown above, they can be computed directly from the coefficients α_{jk}, without re-solving the original problem.

A last basic result of our formulation for the general multigroup classification problem concerns the construction of a general index of fit. For given sample sets of objects and a set of vectors $\boldsymbol{\alpha}^j, j \in S$, such that $G(\boldsymbol{\alpha}) - B(\boldsymbol{\alpha}) > 0$, an index of fit $C(\boldsymbol{\alpha})$ is defined by (2.3):

$$C(\boldsymbol{\alpha}) = 1 - \frac{B(\boldsymbol{\alpha})}{G(\boldsymbol{\alpha})}. \tag{2.3}$$

The main properties of this index are contained in the next two propositions.

Proposition 9. *The index of fit $C_q(\boldsymbol{\alpha}^*)$ associated with an optimal solution $(\boldsymbol{\alpha}^*)$ of* LP^q *has the following properties:*

(i) $0 < C_q(\boldsymbol{\alpha}^*) \leq 1$, *with larger values of $C_q(\boldsymbol{\alpha}^*)$ indicating better classification results for the sample data.*
(ii) $C_q(\boldsymbol{\alpha}^*)$ *is independent of q and of the data transformation of Proposition 8.*
(iii) The objective function of LP^q *can be changed to maximizing $C_q(\boldsymbol{\alpha})$ without changing the solution of* LP^q.

Proof. $C_q(\boldsymbol{\alpha}^*)$ is strictly greater than zero, because $C_q(\boldsymbol{\alpha}^*) = (G(\boldsymbol{\alpha}^*) - B(\boldsymbol{\alpha}^*))/G(\boldsymbol{\alpha}^*) = q/G(\boldsymbol{\alpha}^*)$, while $q > 0$ and $G(\boldsymbol{\alpha}^*) > 0$. Moreover, $C_q(\boldsymbol{\alpha}^*)$ does not exceed 1, since $G(\boldsymbol{\alpha}^*) \geq G(\boldsymbol{\alpha}^*) - B(\boldsymbol{\alpha}^*) = q$, which completes the proof of (i). To prove (ii), we let $B_q(\boldsymbol{\alpha}^*)$ and $B_t(\boldsymbol{\alpha}^{**})$ denote the objective values (badness) of the optimal solutions $\boldsymbol{\alpha}^*$ and $\boldsymbol{\alpha}^{**}$ of LP^q and LP^t, respectively. Then, $C_t(\boldsymbol{\alpha}^{**}) = t/(t + B_t(\boldsymbol{\alpha}^{**})) = t/(t + tq^{-1}B_q(\boldsymbol{\alpha}^*))$, from Proposition 6, so that $C_t(\boldsymbol{\alpha}^{**}) = q/(q + B_q(\boldsymbol{\alpha}^*)) = C_q(\boldsymbol{\alpha}^*)$. The independence of $C_q(\boldsymbol{\alpha}^*)$ from the data transformation of Proposition 8 is obvious since an optimal solution was constructed in the proposition in which the

badness vector β remained unchanged. Part (iii) follows, because minimizing badness $B(\alpha)$ is equivalent to minimizing $q + B(\alpha)$, which in turn is equivalent to maximizing $q/(q + B(\alpha)) = C_q(\alpha)$. □

2.4. Separating Hyperplanes

For any pair of solution vectors α^r and α^j, the following three cases are possible:

(i) $\alpha^r = \alpha^j$,
(ii) $\alpha^r \neq \alpha^j$, but $\alpha_{rk} = \alpha_{jk}$, $k = 1, \ldots, K$, and $\alpha_{r0} \neq \alpha_{j0}$, or
(iii) none of the above.

Given an object with score $\mathbf{x}$, and considering classification into either group r or group j, we have the following situation:

ad (i) No classification between groups r and j is possible, and the sequential procedure is to be applied (see Section 2.1).
ad (ii) No separating hyperplane between groups r and j exists. Any object will be classified into group r if $\alpha_{r0} > \alpha_{j0}$, in group j if $\alpha_{j0} > \alpha_{r0}$.
ad (iii) A separating hyperplane does exist, and the classification is as follows:

if $\alpha^r\mathbf{x} > \alpha^j\mathbf{x}$, then classify into group r,
if $\alpha^r\mathbf{x} < \alpha^j\mathbf{x}$, then classify into group j, and
if $\alpha^r\mathbf{x} = \alpha^j\mathbf{x}$, then classify into either group r or group j.

2.5. ϵ-Procedure

As remarked above, one potential drawback of LP^q-A, as well as of other previously proposed LP-based formulations for the two-group case, is that some objects in the training sample may be located exactly on the boundary between two groups, so that their classification is ambiguous. For instance, three out of nine objects in Example 1 and six of 244 objects in Example 5 below (see Section 3) are located on one or more separating hyperplanes.

This phenomenon may not pose a problem in practice, as long as the size of the training sample is large, the data are continuous, and when the classification rules are applied to validation samples. However, due to their tendency to select separating hyperplanes which cross through some of the objects in the training sample, one should be careful in interpreting the classification performance of linear programming procedures which ignore this issue—certainly on the training sample, but also on the validation sample if the populations have discrete-valued attributes, in which case some validation sample objects may be located exactly on a boundary between two or more groups.

To avoid as much as possible the case of having observations of the training sample located on the separating hyperplanes between groups, it is possible to introduce an ϵ-procedure as follows. For ϵ positive and sufficiently small, let $\beta^i_{rj} = (\alpha^r\mathbf{x}_{ir} - \alpha^j\mathbf{x}_{ir} - \epsilon)^-$ and $\gamma^i_{rj} = (\alpha^r\mathbf{x}_{ir} - \alpha^j\mathbf{x}_{ir} - \epsilon)^+$, where $i \in P_r$, $j \in S_{-r}$ and $r \in S$. Now, the set of restrictions in (1.14) is replaced by (2.4):

$$\beta^i_{rj} + \alpha^r\mathbf{x}_{ir} - \alpha^j\mathbf{x}_{ir} - \gamma^i_{rj} = \epsilon, \quad \text{for all } i \in P_r, j \in S_{-r}, r \in S. \tag{2.4}$$

Note that if $\beta^i_{rj} = \gamma^i_{rj} = 0$, object $i \in P_r$ will always be classified correctly with respect to group j. The remainder of LP^q-A remains unchanged. One choice might be to restrict ϵ to a (small) fraction of the average value of the $\alpha^r\mathbf{x}_{ir} - \alpha^j\mathbf{x}_{ir}$, for instance by applying the formula in (2.5):

$$\epsilon = \frac{1}{F(s-1)\sum_{j\in S} n_j} \sum_{r\in S} \sum_{i\in P_r} \sum_{j\in S_{-r}} (\alpha^r\mathbf{x}_{ir} - \alpha^j\mathbf{x}_{ir}), \tag{2.5}$$

where F is a large positive number denoting the fraction (e.g., $F = 1{,}000$). Defining $T = (s-1)\sum_{j\in S} n_j$ for simplicity and using (2.4), (2.5) can be written as (2.6):

$$\sum_{r\in S} \sum_{i\in P_r} \sum_{j\in S_{-r}} (\gamma^i_{rj} - \beta^i_{rj}) = TF\epsilon - T\epsilon = q, \tag{2.6}$$

or $\epsilon = q/T(F-1)$. Choosing a value of q should be guided by the principle of obtaining an optimal α-vector with components which are neither too small nor too large. Reasonable choices range from $q = T$ to $q = 1{,}000T$. It is possible to refine this ϵ-procedure, e.g., by allowing different ϵ-variables for each pair of groups. However, we will not discuss this extension in the current paper.

Table IV
Results for ϵ-Procedure Applied to the Data of Example 1

Solution Vector		
α^1	α^2	α^3
(0, 0, 0)	(−11.444,8.889,−2.444)	(5.407,6.407,6.407)
Pairwise Separating Hyperplanes		
Group 1 and Group 2	$8.889x_1 - 2.444x_2 = 11.444$	
Group 1 and Group 3	$6.407x_1 - 6.407x_2 = -5.407$	
Group 2 and Group 3	$2.482x_1 + 3.963x_2 = 16.851$	

Example 4. Recall that, even though none of the objects was misclassified, in Example 1 several of the data points were located on the boundary of the classification regions, so that the classification of these objects is ambiguous. We re-solve this example using the ϵ-procedure. The resulting optimal solution and the classification regions are given in Table IV and graphically presented in Figure 4. It appears that the classification scheme resulting from the ϵ-procedure is more attractive than that in the original scheme in Example 1, as the group boundaries are now located strictly inbetween the objects, without increasing the number of misclassified cases.

3. EVALUATION

In this section we use two real data sets that have been published previously in the literature (Rulon et al. 1967, SAS 1988) to compare the classification performance of

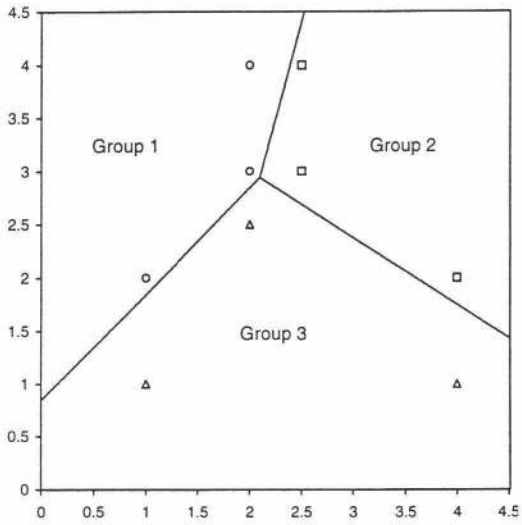

Figure 4. Classification scheme for example 4 (ϵ-procedure).

our proposed formulation with that of Fisher's linear discriminant function (Fisher 1936) and the nonparametric nearest neighbor method. Example 5 (Rulon et al. 1967) is a fairly large data set with three attributes, three groups and 244 objects. Example 6 (SAS 1988) is a small five-group problem with four attributes and 36 objects. We analyze these data sets using both the resubstitution method, where the estimated classification rules are used to classify each object of the training sample, and the Leave-One-Out (LOO) cross-validation method (Lachenbruch 1967). The resubstitution method is known to be positively biased and to underestimate the true misclassification rates, because the very same objects are used to estimate and evaluate the classification rules. The LOO method has been shown to yield almost unbiased estimates of the misclassification rates (Lachenbruch 1967, McLachlan 1992).

3.1. Example 5: Personnel Management Data Set (Rulon et al.)

Rulon et al. (1967) attempt to characterize three groups of employees, "passenger agents," "mechanics," and "operations control agents," of a particular airline company. To this purpose, 85 passenger agents, 93 mechanics and 66 operations control agents were asked to fill out an extensive questionnaire, which included various questions about each employee's preference for certain types of indoor and outdoor activities. These answers were then translated into three composite scores, measured on a ratio-scale, the first one (X_1) measuring preference for outdoor activities, the second (X_2) measuring preference for convivial activities, and the third one (X_3) measuring preference for conservative activities. For further details about the nature of these attributes and the data collection process we refer the reader to Rulon et al. (1967). The purpose of the analysis was to establish rules which would be useful in making personnel assignment decisions, answering such questions as which type of job provides the best fit with a given employee, based on the employee's questionnaire results.

Table V gives the classification results of applying our proposed nonparametric linear programming formulation, with and without the ϵ-procedure, Fisher's parametric linear discriminant function with proportional priors and with equal priors, and the nonparametric k-nearest neighbor method, with $k = 8$ and $k = 16$. The detailed information in the first part of Table V refers to the solutions obtained by applying the estimated classification rules to the training sample objects (resubstitution). From the results we see that, using the resubstitution method, all six methods classify approximately equally well for this data set, with misclassification percentages ranging from 23.8 percent for the nearest neighbor method with $k = 8$ to 27.9 percent for the nearest neighbor method with $k = 16$. Fisher's linear function and the linear programming methods yield almost identical results of between 24.2 and 25 percent misclassified. Note that the linear programming method with the ϵ-procedure correctly classifies each of the six objects from the training sample which were located on the boundary of the solution obtained without including the ϵ-procedure.

The last part of Table V presents the classification results when applying the LOO method. Again, Fisher's method gives slightly more accurate results than the linear programming formulations (25.0 and 25.4 versus 26.6 percent misclassified), while the nearest neighbor methods perform poorly (29.1 and 30.3 percent misclassified).

3.2. Example 6: Remote-Sensing Data on Crops (SAS 1988)

The real data of Example 6 are used in the *SAS/Stat User's Guide* (SAS 1988) to illustrate Fisher's linear discriminant analysis. In this example, four measures of remote-sensing data are used to classify observations as one of five different crops: clover, corn, cotton, soybeans and sugar beets. The training sample consists of 11, 7, 6, 6, and 6 observations in these groups, respectively, for a total sample size of $n = 36$. Given the small number of training sample objects, we limit the nearest neighbor analysis to $k = 8$. Table VI gives the summary classification results using the resubstitution analysis and the LOO analysis.

From Table VI, we see that the misclassification rates are high, no matter which linear classification rule is used. Misclassification rates of over 50 percent are not as surprising in the five-group case as in two-group classification, as in our current example there are multiple ways of misclassifying objects. When re-substituting the training sample, the linear programming approach with $\epsilon > 0$ is the most accurate with a misclassification rate of 27.8 percent, followed at a distance by the linear programming approach with $\epsilon = 0$ (43.0 percent), Fisher's method (50 percent), and the nearest neighbor method with 53.2 percent misclassified. Analyzing the data using the LOO method,

Table V
Solution of Example 5 (Rulon et al. 1967), Plus Comparison with Other Methods

Linear Programming Method (LP^q-A) (ReSubstitution)

$\epsilon = 0$

	Classified into Group: 1	2	3	On Boundary
From Group: 1	66	13	3	3
2	15	63	13	2
3	3	12	50	1

$\epsilon = 0.001$

	Classified into Group: 1	2	3
From Group: 1	69	13	3
2	15	65	13
3	3	12	51

Fisher's Linear Discriminant Function (ReSubstitution)

Proportional Priors

	Classified into Group: 1	2	3
From Group: 1	68	13	4
2	16	67	10
3	3	13	50

Equal Priors

	Classified into Group: 1	2	3
From Group: 1	70	11	4
2	16	62	15
3	3	12	51

***k*-Nearest Neighbor Method (ReSubstitution)**

$k = 8$

	Classified into Group: 1	2	3
From Group: 1	72	7	6
2	17	62	14
3	4	10	52

$k = 16$

	Classified into Group: 1	2	3
From Group: 1	68	11	6
2	20	56	17
3	4	10	51

Summary Classification Results for Example 5

Method	Percentage Misclassified: ReSubstitution Method	Leave-One-Out Method
LP^q-A, $\epsilon = 0$	24.2	26.6
LP^q-A, $\epsilon = .001$	24.2	26.6
Fisher's LDF, Proportional Priors	24.2	25.0
Fisher's LDF, Equal Priors	25.0	25.4
k-Nearest Neighbor, $k = 8$	23.8	29.1
k-Nearest Neighbor, $k = 16$	27.9	30.3

Fisher's method with proportional priors yields the best results, with 63.9 percent misclassified, closely followed by Fisher's method with equal priors and the linear programming approach (66.7 percent). The difference of about three percent between the misclassification rates of these three methods corresponds with a difference of only one misclassified object. As when using resubstitution method, the nearest neighbor method gives the poorest classification results (72.2 percent misclassified).

Table VI
Summary Classification Results for Example 6 (SAS 1988)

Method	Percentage Misclassified: ReSubstitution Method	Leave-One-Out Method
LP^q-A, $\epsilon = 0$	43.0	66.7
LP^q-A, $\epsilon = .001$	27.8	66.7
Fisher's LDF, Proportional Priors	50.0	63.9
Fisher's LDF, Equal Priors	50.0	66.7
k-Nearest Neighbor, $k = 8$	53.2	72.2

4. CONCLUSIONS

Our proposed multigroup LP approach for solving classification problems appears to greatly enhance the types of problems that can be analyzed systematically using nonparametric LP-based methods. The example problems and the analysis of real data sets presented in this paper clearly show that our multigroup LP procedure is indeed capable of providing good classification results, which can compete with both Fisher's parametric method and the nonparametric *k*-nearest neighbor method. The purpose of our paper is to introduce the novel problem formulation and study a number of important properties of the formulation. Of course, future research should further investigate the robustness of the proposed multigroup LP classification method with respect to various data conditions, much like it has already been done—with mixed success—for the two-group case.

APPENDIX

Proof of Proposition 3. From Propositions 1 and 2 it follows that LP^q-A has a finite optimal solution, say $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$. Consider $\alpha^*_{rk} = 0$, $k = 1, \ldots, K$ and $r \in S$; $\alpha^*_{r0} = \alpha_{r0} + \sum_{k=1}^{K} \bar{x}_k \alpha_{rk}$, $r \in S$; $\beta^{i}_{rj}{}^* = \frac{1}{n_r} \sum_{t \in P_r} \beta^t_{rj}$, for all $i \in P_r$, $j \in S_{-r}$

and $r \in S$; and $\gamma_{rj}^{i*} = \frac{1}{n_r} \sum_{t \in P_r} \gamma_{rj}^t$, for all $i \in P_r$, $j \in S_{-r}$ and $r \in S$. We will show that $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*)$ is both feasible and optimal for LP^q-A.

(1) *Feasibility*: From the nonnegativity constraints $\beta_{rj}^i \geq 0$, $\gamma_{rj}^i \geq 0$, $i \in P_r$, $j \in S_{-r}$, $r \in S$, of the original problem in (1.16), it is obvious that $\beta_{rj}^{i*} \geq 0$ and $\gamma_{rj}^{i*} \geq 0$, $i \in P_r$, $j \in S_{-r}$, $r \in S$. Checking whether $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*)$ is feasible with respect to (1.14), we derive that

$$\begin{aligned}
&\beta_{rj}^{i*} + (\boldsymbol{\alpha}^{r*} - \boldsymbol{\alpha}^{j*})\mathbf{x}_{ir}^* - \gamma_{rj}^{i*} = \beta_{rj}^{i*} + \alpha_{r0}^* - \alpha_{j0}^* - \gamma_{rj}^{i*} \\
&= \frac{1}{n_r} \sum_{t \in P_r} \beta_{rj}^t + \alpha_{r0} + \sum_{k=1}^{K} \bar{x}_k \alpha_{rk} - \alpha_{j0} - \sum_{k=1}^{K} \bar{x}_k \alpha_{jk} - \frac{1}{n_r} \sum_{t \in P_r} \gamma_{rj}^t \\
&= \frac{1}{n_r} \Bigg[\sum_{t \in P_r} \beta_{rj}^t + n_r \alpha_{r0} + \sum_{k=1}^{K} \sum_{t \in P_r} x_{trk} \alpha_{rk} \\
&\qquad - n_r \alpha_{j0} - \sum_{k=1}^{K} \sum_{t \in P_r} x_{trk} \alpha_{jk} - \sum_{t \in P_r} \gamma_{rj}^t \Bigg] \\
&= \frac{1}{n_r} \Bigg[\sum_{t \in P_r} \beta_{rj}^t + \sum_{t \in P_r} \boldsymbol{\alpha}^r \mathbf{x}_{tr} - \sum_{t \in P_r} \boldsymbol{\alpha}^j \mathbf{x}_{tr} - \sum_{t \in P_r} \gamma_{rj}^t \Bigg].
\end{aligned}$$

This last expression equals zero, because $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ is a feasible (and optimal) solution to LP^q-A, so that (1.14) implies that $\beta_{rj}^t + \boldsymbol{\alpha}^r \mathbf{x}_{tr} - \boldsymbol{\alpha}^j \mathbf{x}_{tr} - \gamma_{rj}^t = 0$. Hence, $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*)$ satisfies the constraint set (1.14) in LP^q-A as well. Verifying the feasibility of the normalization constraint (1.15) with respect to $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*)$, we see that

$$\begin{aligned}
&\sum_{i \in P_r} \sum_{j \in S_{-r}} \sum_{r \in S} (\gamma_{rj}^{i*} - \beta_{rj}^{i*}) = \sum_{i \in P_r} \sum_{j \in S_{-r}} \sum_{r \in S} \frac{1}{n_r} \sum_{t \in P_r} \\
&\cdot (\gamma_{rj}^i - \beta_{rj}^i) = \sum_{i \in P_r} \sum_{j \in S_{-r}} \sum_{r \in S} (\gamma_{rj}^i - \beta_{rj}^i),
\end{aligned}$$

which expression indeed equals q, using (1.15). This completes the proof that $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*)$ is a feasible solution to LP^q-A, and hence to LP^q.

(2) *Optimality*: We know that $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ is an optimal solution to LP^q, with an objective function value of $z^q = \sum_{i \in P_r} \sum_{j \in S_{-r}} \sum_{r \in S} \beta_{rj}^i$. The objective function value of $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*)$ is given by:

$$\begin{aligned}
z^* &= \sum_{r \in S} \sum_{j \in S_{-r}} \sum_{i \in P_r} \beta_{rj}^{i*} = \sum_{r \in S} \sum_{j \in S_{-r}} \sum_{i \in P_r} \frac{1}{n_r} \sum_{t \in P_r} \beta_{rj}^t \\
&= \sum_{r \in S} \sum_{j \in S_{-r}} \sum_{i \in P_r} \beta_{rj}^i = z^q,
\end{aligned}$$

which completes the proof. □

Proof of Proposition 4. From Proposition 3 it follows that we can set $\alpha_{rk}^* = 0$, $k = 1, \ldots, K$, and $r \in S$. Thus, LP^q-A simplifies to the LP-problem in (A.1)–(A.4):

$$\text{Min} \sum_{r \in S} \sum_{j \in S_{-r}} \sum_{i \in P_r} \beta_{rj}^i, \tag{A.1}$$

Subject to:

$$\beta_{rj}^i + \alpha_{r0} - \alpha_{j0} \geq 0, \quad i \in P_r, j \in S_{-r}, r \in S, \tag{A.2}$$

$$\sum_{i \in S} \left(sn_i - \sum_{j \in S} n_j \right) \alpha_{i0} = q, \tag{A.3}$$

$$\beta_{rj}^i \geq 0, \quad i \in P_r, j \in S_{-r}, r \in S. \tag{A.4}$$

Note that, for an optimal solution to this problem, $\beta_{rj}^i = \text{Max}(0; -\alpha_{r0} + \alpha_{j0})$, for all $i \in P_r$, so that we can set $\beta_{rj}^h = \beta_{rj}^i$ for all $h, i \in P_r$. Using this information, and letting $m_i = sn_i - \sum_{j \in S} n_j$, the above linear program can be simplified to (A.5)–(A.8):

$$\text{Min} \sum_{r \in S} \sum_{j \in S_{-r}} n_r \beta_{rj}, \tag{A.5}$$

Subject to:

$$\beta_{rj} + \alpha_{r0} - \alpha_{j0} \geq 0, \quad j \in S_{-r}, r \in S, \tag{A.6}$$

$$\sum_{i \in S} m_i \alpha_{i0} = q, \tag{A.7}$$

$$\beta_{rj} \geq 0, \quad j \in S_{-r}, r \in S. \tag{A.8}$$

Substituting the β_{rj} in the objective function, this problem can in turn be rewritten as (A.9)–(A.10):

$$\text{Min} \sum_{r \in S} \sum_{j \in S_{-r}} n_r [\text{Max}(0, \alpha_{j0} - \alpha_{r0})], \tag{A.9}$$

Subject to:

$$\sum_{i \in S} m_i \alpha_{i0} = q. \tag{A.10}$$

Notice that $\sum_{i \in S} m_i = 0$. It follows that if α_{r0}^*, $r \in S$, is optimal, the solution $\alpha_{r0}^* + c$, $r \in S$, where c is any constant, is also optimal. Hence, we can assume that an optimal solution exists such that the α_{r0}^* take on $\pi + 1$ different values, $\pi + 1 \leq s$, say $v_0, v_1, \ldots, v_\pi$, such that $0 = v_0 < v_1 < \cdots < v_\pi$.

Let $S_j = \{i: \alpha_{i0}^* = v_j\}$, $\bar{S}_j = \{i: i \in S \backslash S_j\}$ and let $|S|$ denote the cardinality of set S. Further, denote the minimal objective function value of LP^q-A by z_q, for any $q > 0$. Let ϵ be such that

$$0 < \epsilon < \min\left(v_\pi - v_{\pi-1}, \text{abs}\left(\frac{q}{\sum_{i \in S_\pi} m_i} \right) \right),$$

where abs(y) is the absolute value of y, and let

$$\alpha_{i0}^{\text{new}} = \begin{cases} \alpha_{i0}^* - \epsilon, & i \in S_\pi, \\ \alpha_{i0}^*, & i \in \bar{S}_\pi, \end{cases}$$

$$\beta_{rj}^{\text{new}} = \begin{cases} \beta_{rj}^* - \epsilon, & i \in S_\pi, r \in \bar{S}_\pi, \\ \beta_{rj}^*, & \text{all other } (r, j) \in S \times S. \end{cases}$$

Next, consider the program LP^u, where $u = q - \epsilon \sum_{i \in S_\pi} m_i$, by the choice of ϵ. Note that $\epsilon < v_\pi - v_{\pi-1}$ and $u > 0$. It is easy to check that the above solution is feasible for LP^u, and its objective value equals v^{new} in (A.12):

$$v^{\text{new}} = z_q - \epsilon |S_\pi| \sum_{j \in S_\pi} n_j. \tag{A.12}$$

From (A.12) it follows immediately that

$$v^{\text{new}} < z_q. \tag{A.13}$$

By the definition of z_u we also have $z_u \leq v^{\text{new}}$, and by Proposition 6, $z_u = u/q\, z_q \leq v^{\text{new}}$ or, substituting $u = q - \epsilon \sum_{i \in S_\pi} m_i$, we have (A.14):

$$z_q \left(1 - \frac{1}{q} \epsilon \sum_{i \in S_\pi} m_i \right) \leq v^{\text{new}}. \tag{A.14}$$

From (A.13) and (A.14) it follows that $\Sigma_{i\in S_\pi} m_i > 0$, and from (A.12) and (A.14) we derive that (A.15) holds,

$$z_q\left(1 - \frac{1}{q}\,\epsilon \sum_{i\in S_\pi} m_i\right) \leq z_q - \epsilon|S_\pi| \sum_{j\in \bar{S}_\pi} n_j. \tag{A.15}$$

As $\Sigma_{i\in S_\pi} m_i > 0$, (A.15) implies (A.16):

$$z_q \geq \frac{q|S_\pi| \sum_{j\in \bar{S}_\pi} n_j}{\sum_{i\in S_\pi} m_i}. \tag{A.16}$$

Since the solution proposed in Proposition 4 has an objective value z_q equal to the lower bound of (A.16), and the sets S_π and $\bar{S}_\pi$ are such that this lower bound itself is at its minimum, this solution must be optimal. □

Proof of Proposition 5. Similar to Proposition 3, it can be shown that, if (α, β, γ) is an optimal solution to LP^q-A, then $(\alpha^*, \beta^*, \gamma^*)$ is also an optimal solution to LP^q-A, where:

$$\alpha^*_{rk} = 0,\ r \in S;\quad k = 0, 1, \ldots, t-1, t+1, \ldots, K, \tag{A.17}$$

$$\alpha^*_{rt} = \frac{1}{\bar{x}_t}\left(\alpha_{r0} + \sum_{k=1}^{K} \bar{x}_k \alpha_{rk}\right),\quad r \in S, \tag{A.18}$$

$$\beta^{i}_{rj}{}^* = \frac{x_{irt}}{n_r \bar{x}_t} \sum_{i\in P_r} \beta^i_{rj},\quad i \in P_r, j \in S_{-r}, r \in S, \tag{A.19}$$

$$\gamma^{i}_{rj}{}^* = \frac{x_{irt}}{n_r \bar{x}_t} \sum_{i\in P_r} \gamma^i_{rj},\quad i \in P_r, j \in S_{-r}, r \in S. \quad \square \tag{A.20}$$

REFERENCES

ANDERSON, T. W. 1984. *Introduction to Multivariate Statistical Analysis*. Second Edition, Wiley, New York.

DUARTE SILVA, A. P. AND A. STAM. 1994. Second Order Mathematical Programming Formulations for Discriminant Analysis. *European J. Opnl. Res.* **72,** 4–22.

ERENGUC, S. S. AND G. J. KOEHLER. 1990. Survey of Mathematical Programming Models and Experimental Results for Linear Discriminant Analysis. *Managerial and Decision Economics*, **11,** 215–225.

FISHER, R. A. 1936. The Use of Multiple Measurements in Taxonomy Problems. *Ann. Eugenics*, **7,** 179–188.

FREED, N. AND F. GLOVER. 1981a. A Linear Programming Approach to the Discriminant Problem. *Decision Sci.* **12,** 68–74.

FREED, N. AND F. GLOVER. 1981b. Simple but Powerful Goal Programming Formulations for the Discriminant Problem. *European J. Opnl. Res.* **7,** 44–60.

FREED, N. AND F. GLOVER. 1986. Resolving Certain Difficulties and Improving the Classification Power of the LP Discriminant Analysis Procedure. *Decision Sci.* **17,** 589–595.

GEHRLEIN, W. V. 1986. General Mathematical Programming Formulations for the Statistical Classification Problem. *O. R. Lett.* **5,** 299–304.

GLOVER, F., S. KEENE AND B. DUEA. 1988. A New Class of Models for the Discriminant Problem. *Decision Sci.* **19,** 269–280.

GLOVER, F. 1990. Improved Linear Programming Models for Discriminant Analysis. *Decision Sci.* **21,** 771–785.

GOCHET, W. AND V. SRINIVASAN. 1983. Two Group Classification and Linear Programming, *Liber Amicorum*. Verzamelde Ekonomische Studiën, F. Van Winckel (ed.), Faculty of Economics and Applied Economic Sciences, Katholieke Universiteit Leuven, Belgium, 91–99.

HAND, D. J. 1981. *Discrimination and Classification*. Wiley, New York.

JOACHIMSTHALER, E. A. AND A. STAM. 1990. Mathematical Programming Approaches for the Classification Problem in Two-Group Discriminant Analysis. *Multivariate Behavioral Res.* **25,** 427–454.

JOHNSON, R. A. AND D. W. WICHERN. 1988. *Applied Multivariate Statistical Analysis*. Second Edition, Prentice-Hall, Englewood Cliffs, NJ.

KOEHLER, G. J. 1989a. Characterizations of Unacceptable Solutions in LP Discriminant Analysis. *Decision Sci.* **20,** 239–257.

KOEHLER, G. J. 1989b. Unacceptable Solutions and the Hybrid Discriminant Model. *Decision Sci.* **20,** 844–848.

KOEHLER, G. J. 1990. Considerations for Mathematical Programming Models in Discriminant Analysis. *Managerial and Decision Economics*, **11,** 227–234.

KOEHLER, G. J. 1991. Improper Linear Discriminant Classifiers. *European J. Opnl. Res.* **50,** 188–198.

LACHENBRUCH, P. A. 1967. An Almost Unbiased Method of Obtaining Confidence Intervals for the Probability of Misclassification in Discriminant Analysis. *Biometrics*, **23,** 525–534.

MCLACHLAN, G. J. 1992. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York.

NATH, R., W. M. JACKSON AND T. W. JONES. 1992. A Comparison of the Classical and the Linear Programming Approaches to the Classification Problem in Discriminant Analysis. *J. Statist. Comput. Simul.* **41,** 73–93.

RAGSDALE, C. T. AND A. STAM. 1991. Mathematical Programming Formulations for the Discriminant Problem: An Old Dog Does New Tricks. *Decision Sci.* **22,** 296–306.

RULON, P. J., D. V. TIEDEMAN, M. M. TATSUOKA AND C. R. LANGMUIR. 1967. *Multivariate Statistics for Personnel Classification*. Wiley, New York, NY.

SAS INSTITUTE, INC. 1988. *SAS/STAT User's Guide, Release 6.03 Edition*. SAS Institute, Carey, NC.

SRINIVASAN, V. AND A. SHOCKER. 1973. Linear Programming Techniques for Multi-Dimensional Analysis of Preferences. *Psychometrica*, **38,** 337–369.

SMITH, C. A. B. 1947. Some Examples of Discrimination. *Ann. Eugenics*, **13,** 272–282.

SMITH, F. W. 1968. Pattern Classifier Design by Linear Programming. *IEEE Trans. Comput.* **C-17,** 367–372.

SMITH, F. W. 1969. Design of Multicategory Pattern Classifiers with Two-Category Classifier Design Procedures. *IEEE Trans. Comput.* **C-18,** 548–551.

STAM, A. 1997. MP Approaches to Classification: Issues and Trends. *Ann. Operations Res.*, forthcoming.

STAM, A. AND E. A. JOACHIMSTHALER. 1989. Solving the Classification Problem in Discriminant Analysis via Linear and Nonlinear Programming Methods. *Decision Sci.* **20,** 285–293.

STAM, A. AND C. T. RAGSDALE. 1992. On the Classification Gap in Mathematical Programming-Based Approaches to the Discriminant Problem. *Naval Res. Logist.* **39,** 545–559.