A QUADRATIC PROGRAMMING APPROACH FOR SOLVING DISCRIMINANT ANALYSIS PROBLEMS

John Maloney and Yi-Hsin Liu

University of Nebraska, Omaha

Abstract. In the past decade, there has been some interest shown in solving discriminant analysis problems using mathematical programming techniques. In this paper, we present a quadratic program which is used to obtain a linear discriminant function. This would provide an alternative to the conventional statistical approach.

1. Introduction. A linear discriminant analysis problem seeks a vector c which is used to construct a linear discriminant function $f(x) = cx$. This linear discriminant function is then used to separate the given groups of vector-valued data G_1, G_2, …, G_m and provides an allocation rule for placing future unclassified data into one of the groups.

In this paper, we assume that a subjective ranking (order relation) has been imposed on the groups G_1, G_2, …, G_m. That is, for any two distinct groups of data G_i and G_j either G_i is preferred to G_j or G_j is preferred to G_i. Without loss of generality, we may assume that G_i is preferred to G_j whenever $i > j$. This order relation is denoted by writing $G_j \prec G_i$ if $i > j$. Thus, we are given

$$G_1 \prec G_2 \prec \cdots \prec G_m$$

This assumption arises in many problems [5] and it is possible to use, in some cases, artificial intelligence techniques to determine the subjective rankings discussed above. This will not be discussed in this paper, rather we will assume that the rankings have been given.

This problem was discussed in [7] and is summarized as follows.

Problem 1.1. Given m groups of vector-valued data (that is values in E^n) such that

(1) $G_1 \prec G_2 \prec \cdots \prec G_m$

(2) $G_i = \{x^i_j \in E^n : j = 1, 2, \ldots, l_i\}$ where $i = 1, 2, \ldots, m$.

find a vector c (and hence a linear discriminant function $f(x) = cx$), and the appropriate intervals $I_i = (L_i, U_i]$, such that

i. $I_i \cap I_k = \emptyset$, $\forall\ 1 \le i, k \le n$, $i \ne k$.

ii. $f(x^i_j) \in I_i$, $\forall\ j = 1, 2, \ldots, l_i$ and $\forall\ i = 1, 2, \ldots, m$.

iii. $L_1 < U_1 \le L_2 < U_2 \le \cdots \le L_m < U_m$

Definition 1.2. The groups G_1, G_2, ..., G_m are said to be separable if there exists a linear function $f(x) = cx$ such that $f(x^i_j) \in I_i$, $\forall\ j = 1, 2, \ldots, l_i$ and $\forall\ i = 1, 2, \ldots, m$ provided that (iii) above holds. Otherwise the groups are said to be nonseparable.

2. Model and Discussion. It may happen that there is no linear function $f(x) = cx$ that will separate the data in the groups under the order relation prescribed in Problem 1.1. Of course, there could be a linear function $f(x) = cx$ that will separate the data into disjoint intervals but in such a manner that the prescribed order relation, in 1.1, is not satisfied.

We propose to split Problem 1.1 into two parts:

a. Find a linear function $f(x) = cx$, if possible, that separates the groups of data into disjoint intervals without regard to the prescribed order relation in Problem 1.1. The technique for doing this is given below and will be seen to be completely independent of the order relation in Problem 1.1.

b. Suppose that the intervals found in (a) are $I_j = [L_j, U_j]$, where $f(G_j) \subseteq I_j$. Since the order relation in Problem 1.1 was ignored in (a) above, it is not necessarily true that $L_i < L_j$ if $i < j$. To meet the order relation specified in Problem 1.1, relabel the groups, if the problem permits, to conform to the order found in (a). If the problem does not permit this then create a piecewise linear function $g(x)$ which maps each I_j into another interval $[g(L_j), g(U_j)]$ in such a manner that

$$g(L_1) < g(U_1) \le g(L_2) < g(U_2) \le \cdots \le g(L_m) < g(U_m)$$

in conformity with the order relation in Problem 1.1.

Since the function g, of (b) can always be found once the intervals from part (a) are known we confine our attention to attempting to find the intervals in part (a).

The quantities $\overline{x}_i$ in the objective function below are defined by

$$\overline{x}_i = \frac{\sum_{j=1}^{l_i} x^i_j}{l_i} .$$

We may think of the $\overline{x}_i$'s as being the center of mass of l_i unit masses located at x_j^i. This becomes then the center of mass of the group G_i. For this reason we shall refer to the points $\overline{x}_i$ as centroids.

The model used is

Q. P. 2.1. Maximize

$$\sum_{i=1}^{m} \sum_{j=i+1}^{m} (c\overline{x}_i - c\overline{x}_j)^2 + 2K - \sum_{i=1}^{m}(U_i - L_i)$$

subject to the constraints

$$i. \quad -K \leq L_i \leq cx_j^i \leq U_i \leq K \quad i = 1, 2, \ldots, m, \quad j = 1, 2, \ldots, l_i \ .$$

$$ii. \quad \sum_{j=1}^{n} c_j \geq 1 \ .$$

We attempt to find a $c = (c_1, c_2, \ldots, c_n)$ that will give maximum separation to the values $c\overline{x}_i$. It seems reasonable that this, if successful, will lead to the most widely separated set of disjoint intervals I_j. We would also like the intervals $[L_i, U_i]$ to be of minimum length in the sense that $cx_j^i = L_i$ for some j and $cx_j^i = U_i$ for some (different) value of j.

The first term in the objective function is maximized by giving maximum separation to the centroids. The second term in the objective function is maximized by making the intervals $[L_i, U_i]$ as small as possible subject to the constraint (i). Thus maximizing our objective function gives us the two properties that we sought above. K is an arbitrarily selected constant that guarantees a bounded solution for the intervals. In the examples that follow a value of 100 was used and seemed to work quite well.

The advantage of this form of the problem is that it determines what might be called a natural order for the intervals $[L_i, U_i]$ and removes any dependence upon the order relation prescribed for the G_i in the original problem. In most cases the first term, in the objective function, is the larger and there is no danger of c being zero. For such problems we do not need to impose the constraint in (iii). There is no danger that c will be unbounded because of constraint (i).

3. Example. Two sets of data were considered, one set was separable and the other was not. The model was solved on Gino [8].

The two groups of data were:

<u>Data Set 3.1</u>. The nonseparable data is

i. $G_1 = \{(2,4),(4,5),(6,2)\}$ $\quad \overline{x}_1 = (4, 11/3)$.

ii. $G_2 = \{(3,4),(4,3),(5,4)\}$ $\quad \overline{x}_2 = (4, 11/3)$.

iii. $G_3 = \{(5,3),(5,6),(6,7)\}$ $\quad \overline{x}_3 = (16/3, 16/3)$.

<u>Data Set 3.2</u>. The separable data is

i. $G_1 = \{(4,1),(6,1),(6,4)\}$ $\quad \overline{x}_1 = (16/3, 2)$.

ii. $G_2 = \{(3,4),(4,5),(5,4)\}$ $\quad \overline{x}_2 = (4, 13/3)$.

iii. $G_3 = \{(1,4),(2,5),(2,6)\}$ $\quad \overline{x}_3 = (5/3, 5)$.

Each of the two groups was run twice. Once using constraint (iii) and once without. The model results are presented in the two tables below.

with (iii)	*item*	*without (iii)*
$[-100, -29.4]$	$[L_1, U_1]$	$[29.4, 100]$
$[-8.8, 35.3]$	$[L_2, U_2]$	$[-35.3, 8.8]$
$[73.5, 100]$	$[L_3, U_3]$	$[-100, -73.5]$
$(-20.5, 23.5)$	(c_1, c_2)	$(20.5, -23.5)$

Table 3.1

Separable Data

with (iii)	*item*	*without (iii)*
$[-100, 100]$	$[L_1, U_1]$	$[-100, 100]$
$[0, 70]$	$[L_2, U_2]$	$[0, 70]$
$[-30, 100]$	$[L_3, U_3]$	$[-30, 100]$
$(-30, 40)$	(c_1, c_2)	$(-30, 40)$

Table 3.2

Nonseparable Data

In the following figure we show the separable data and the separating lines using the vector $c = (-20.5, 23.5)$.

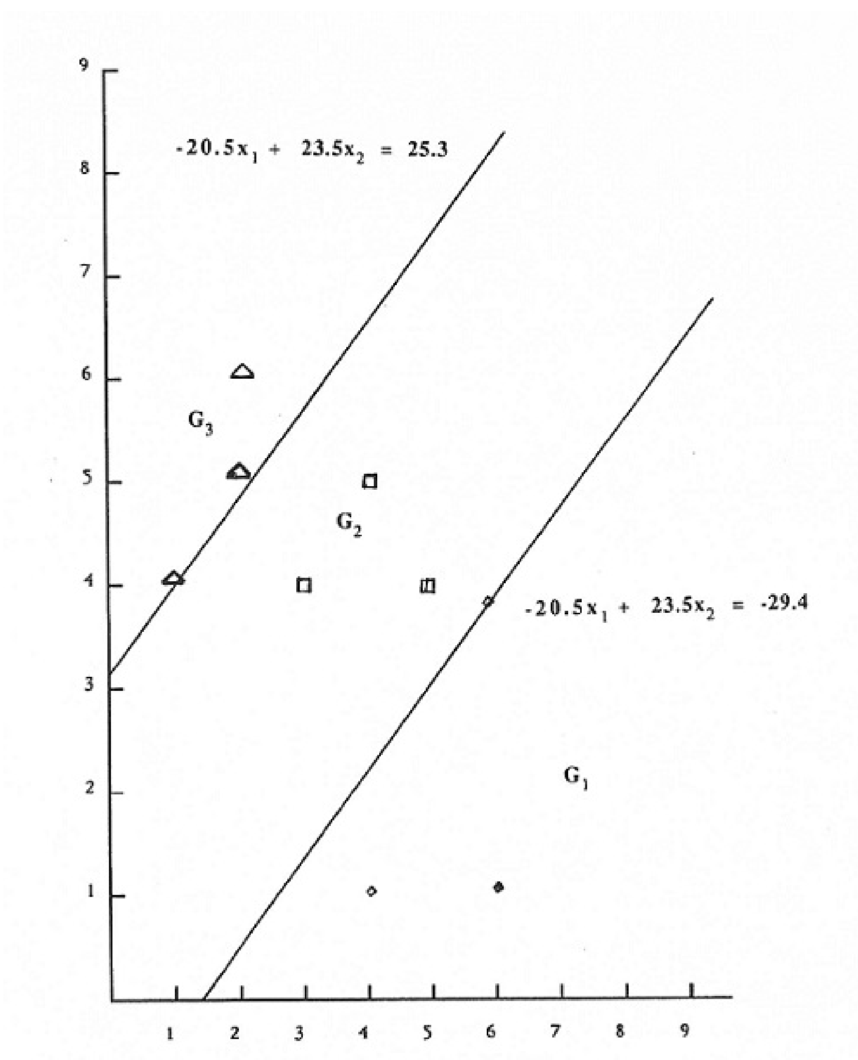

Figure 1

References

1. T. M. Cavalier, J. P. Ignizio, and A. L. Soyster, "Discriminant Analysis via Mathematical Programming: Certain Problems and Their Causes," *Computer and Operations Research*, 16 (1989), 353–362.

2. N. Freed and F. Glover, "A Linear Programming Approach to the Discriminant Problem," *Decision Sciences*, 12 (1981), 67–84.

3. N. Freed and F. Glover, "Simple But Powerful Goal Programming Models for Discriminant Problems," *European Journal of Operational Research*, (1989), 353–362.

4. R. A. Johnson and D. W. Wichen, *Applied Multivariate Statistical Analysis*, Prentice Hall, Englewood Cliffs, N.J., 1988.

5. J. Liebman, L. Lasdon, L. Schrage, and A. Waren, *GINO*, The Scientific Press, San Francisco, California, 1986.

6. Y-H. Liu and Z. Chen, "Rule Base Assisted Consistence Checking for Decision Making," *Proceedings of the Twentieth Annual Model and Simulation Conference*, May 1989.

7. Y-H. Liu and J. Maloney, "Discriminant Analysis Using a Bicriteria Linear Program," *Missouri Journal of Mathematical Sciences*, 4 (1992), 61–69.

8. P. A. Rubin, "Evaluating the Maximum Distance Formulation of the Discriminant Problem," *European Journal of Operational Research*, 41 (1989), 240–248.