Notes and Communications

RESOLVING CERTAIN DIFFICULTIES
AND IMPROVING THE CLASSIFICATION POWER
OF LP DISCRIMINANT ANALYSIS FORMULATIONS

Ned Freed
School of Business, University of Portland, Portland, OR 97203

Fred Glover
School of Business, Box 419, University of Colorado, Boulder, CO 80309

ABSTRACT

In certain settings, difficulties arise that limit the effectiveness of LP formulations for
the discriminant problem. Explanations and possible remedies have been offered, but these
have had only limited success. We provide a simple way to overcome these problems based
on an appropriate use and interpretation of normalizations. In addition, we demonstrate
a normalization that is invariant under all translations of the problem data, providing a
stability property not shared by previous approaches. Finally, we discuss the possibility of
using more general models to improve discrimination.

Subject Areas: Goal Programming, Linear Programming, and Statistical Techniques.

INTRODUCTION

In a recent paper, Markowski and Markowski [6] identified potential difficulties
that may arise with the linear programming (LP) approaches to discriminant analysis
first introduced by us in [3] [4]. Citing a set of illustrative cases, the authors ex-
pressed concern that placement of the origin can seriously affect the model's capacity
to discriminate properly. Bajgier and Hill [1] also reported cases in which the clas-
sification capability of LP models seems to break down but attributed this to the
presence of extreme outliers in the subject data groups rather than to origin
placement.

In pursuit of a broader explanation, Markowski and Markowski attempted
to isolate specific characteristics of those data sets which seem to resist effective
classification under LP formulations. The types of difficulties they encountered
may be expressed under the headings of "degeneracy" and "stability." They con-
cluded that the appearance of negative-valued data or, more generally, the ap-
ppearance of data points in all four quadrants, promotes breakdown of the formulation
by forcing an overly restrictive origin requirement on the LP discriminant func-
tion. To compensate, they provided an interesting analysis that leads to data trans-
formations designed to shift the origin to a more advantageous position. Unfor-
unately, as their analysis showed, problems remain that resist effective treatment
in this fashion.

SOURCES OF DIFFICULTIES

We will show that the types of difficulties identified by Markowski and Mar-
kowski can be resolved by reference to the "normalization" the model employs,
which can take more than one form. The normalization used in most previous studies sets a boundary value \( b \) to a positive constant. (Common values are 1 and 10, the difference being simply one of scale in the final result.)

We note as a first step that the simplest manifestation of difficulties in LP discriminant formulations can arise from nothing more than the restriction of \( b \) to a positive value, a restriction that in turn constrains the search for an effective set of discriminant weights—sometimes fatally.

Such a situation can be demonstrated using the two-group, two-dimensional data set below:

<table>
<thead>
<tr>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>((a_{i1}, a_{i2}))</td>
<td>((a_{i1}, a_{i2}))</td>
</tr>
<tr>
<td>-1, 3</td>
<td>0, 2.5</td>
</tr>
<tr>
<td>-1, 4</td>
<td>2, 3</td>
</tr>
<tr>
<td>1, 3</td>
<td>-4, -9</td>
</tr>
<tr>
<td>4, 9</td>
<td>3, 4</td>
</tr>
<tr>
<td>2, 7</td>
<td>-1, -3</td>
</tr>
</tbody>
</table>

We illustrate by referring to the elementary LP discriminant model that seeks discriminant weights \( x_1 \) and \( x_2 \) in order to

Minimize \( z \)

subject to

\[ a_{i1}x_1 + a_{i2}x_2 - z \leq b \quad \text{for all } i \text{ in } G_1 \]

and

\[ a_{i1}x_1 + a_{i2}x_2 + z \geq b \quad \text{for all } i \text{ in } G_2 \]

where \( G_1 \) and \( G_2 \) are the index sets for group 1 and group 2 and the variables \( x_1 \), \( x_2 \), and \( b \) are unrestricted in sign. If the variable \( z \) is treated as nonnegative, the model corresponds to the one from [4] that seeks to minimize the maximum deviation of points in groups 1 and 2 that fail to satisfy the inequality in which \( z \) is removed. We consider this situation first.

With the normalization constraint \( b = 10 \), the model produces a trivial (degenerate) solution in which the discriminant weights, \( x_1 \) and \( x_2 \), both are zero. The difficulty can be portrayed graphically as in Figure 1.

The cause of the problem is apparent. Figure 1 discloses that an effective discriminant function will be produced only if \( b \) takes a negative value. As constructed above, the LP model seeks the set of discriminant weights \( x_1 \) and \( x_2 \), which will produce weighted scores generally less than or equal to \( b \) for group 1 members and greater than or equal to \( b \) for group 2 members. Given the position of the origin relative to the required boundary \( b \), an efficient discriminant solution with these characteristics is impossible unless \( b \) is allowed to take on a negative value.
FIGURE 1
Effect of Different Normalization Signs

(The reader should note that the vector \( x \), sketched to represent the optimal discriminant weighting scheme \((x_1, x_2)\), suggests that \( x_1 \) must be positive while \( x_2 \) must be negative with \( b \), as argued, clearly required to take on a negative value.) By setting a normalization constraint that forces \( b=10 \), we have ensured that the method surrenders to degeneracy.

Two fairly straightforward remedies suggest themselves. A simple reversal of sign on \( b \) (e.g., let \( b=-10 \)) produces a far more attractive discriminant solution in which \( x_1=2 \) and \( x_2=-4 \)—a solution which reduces the overlap variable \( z \) to zero (establishing perfect group separation). Equivalently, we can reverse the inequality constraints so that

\[
a_{it}x_1 + a_{it}x_2 + z \geq b
\]

for all group 1 members.
and
\[ a_{11}x_1 + a_{12}x_2 - z \leq b \]
for all group 2 members

while retaining the condition that \( b = +10 \).

This latter is the same as simply reversing the sequence in which the experimental data points are assigned to the two groups, which has been noted previously in [1] and [4] to affect the quality of the resulting classifications.

**ALTERNATIVE NORMALIZATIONS**

The preceding remedy by itself, however, is not sufficient to handle the range of complications that can arise. To handle broader concerns, the simple normalization that sets \( b \) to a constant, whether positive or negative, must be replaced by a more effective normalization. One candidate for such a normalization, proposed in [5], consists of allowing \( b \) to remain a variable unrestricted in sign and requiring the sum of \( b \) and the \( x_j \) weights to equal a constant. Each of the variables in the sum is bounded from above and below. (Alternatively, an upper bound can be placed on the sum of the absolute values of these variables. This can be accomplished by expressing each variable as the difference of two nonnegative variables and bounding the sum of these substitute variables.) Another candidate for a more effective normalization requires the sum of the \( x_j \) values to equal a constant without reference to \( b \).

Several properties of these two normalizations may be observed immediately: (1) both allow for the case where \( b = 0 \), which can be only approximated in the simpler normalization by letting the absolute magnitude of one or more other variables tend to infinity; (2) the first allows for the case where the sum of the \( x_j \) values equals 0, which can be approximated only by the second; (3) the second eliminates the possibility of the degenerate solution in which all \( x_j \) are 0; (4) except for the asymptotic conditions noted in (1) and (2) above, both are able to differentiate any two groups that can be strictly separated by a hyperplane for any objective that dominantly weights the deviation variables measuring incorrect classification, provided the right sign of the normalization constant is selected.

This fourth point, of course, presupposes the lower and upper bounds on the problem variables are “sufficiently large” positive and negative numbers. Indeed, it is evident that such bounds are irrelevant for the simple model illustrated previously since an unbounded optimum is impossible given the stipulation that \( z \) is non-negative. In other models, such as the one employed by Markowski and Markowski [6] (where \( z \) is permitted to be negative), unbounded optima may occur if the feasible region is not bounded. Such a condition does not arise from any pathology in either the data or the model but simply reflects the fact that the groups are strictly separable and the measure of separation can be increased indefinitely by allowing the values of the variables to scale upward (provided this can be accomplished while satisfying the normalization constraint). Each type of normalization, including the standard that sets \( b \) to a constant, will eliminate some cases of unboundedness
permitted by the others. In general, the simple expedient of bounding the feasible region, if the bounds are not too tight, will remove unbounded optimality without negative effects on the ability to discriminate. (This condition also can be handled by bounding the objective function from below.)

We note that instances of degeneracy, where the $x_j$ values all are zero, still can occur in the normalization that includes $b$ in its sum; but, as with "unboundedness," these cases represent nothing pathological. Such an outcome merely signals that the groups cannot be separated and that the form of group overlap confounds any reasonable "partial separation" with the type of model employed. We emphasize these last words because the models studied empirically thus far are among the simpler types of LP discriminant formulations. Some of the more general alternatives offer greater latitude for group differentiation.

A STABILITY PROPERTY

We might imagine that the normalization that sets the sum of the $x_j$ variables to a constant will be less flexible than one that includes $b$ in the sum, in view of condition (2) noted earlier. However, the normalization that excludes $b$ has a stability property not shared by the other normalizations. To express this property, assume an upper bound is imposed on the sum of the absolute values of the $x_j$ weights (or a lower bound is imposed on the value of the objective to be minimized) while $b$ is left unrestricted. Then the following result applies to every two-group LP discrimination analysis formulation described in [4, Section 2.4].

Theorem. The normalization that constrains the sum of the $x_j$ weights to a constant yields the same optimal solution values for the objective function and all variables except $b$, for every translation of the data points.

To establish this result, first express it in a more general form that allows simplified notation. Assume slack and surplus variables are added to convert inequalities into equations, and all problem variables (except $b$) are represented as components of a single vector $\mathbf{z}$. Then a problem definition that includes the formulations covered by the theorem is the following:

Minimize $f(\mathbf{z})$: $\mathbf{B}_i \mathbf{z} = b$, $i \in M$ and $\mathbf{z} \in Z$.

The condition $\mathbf{z} \in Z$ may incorporate an arbitrary set of restrictions and hence encompass various types of normalizations involving the $x_j$ components of $\mathbf{z}$. Each $\mathbf{B}_i$ is a row vector whose elements are constants, and $b$ is a scalar variable.

We note that a translation of the vector $\mathbf{B}_i$ includes, as a special case, a translation relative to those components associated with the $x_j$ weights. Representing such a translation by a row vector $\mathbf{T}$, we may write the problem that results from the translation as

Minimize $f(\mathbf{z})$: $(\mathbf{B}_i + \mathbf{T}) \mathbf{z} = b$, $i \in M$ and $\mathbf{z} \in Z$.

Relative to this representation it is immediately clear that if $(\mathbf{z}^*,b^*)$ is optimal for the first (second) problem, then $[(\mathbf{z}^*,b^*) + \mathbf{T}^*(\mathbf{z}^*,b^* - \mathbf{z}^* \mathbf{T})]$ is optimal for the
second (first) problem. (By contradiction, the existence of a better solution for either problem implies that \((z^*, b^*)\) could not have been optimal for the original.)

Markowski and Markowski have called attention to the fact that the stability property expressed by the preceding theorem does not hold for the simple normalization that sets \(b\) to a constant. Duea [2] showed that it also does not hold for the normalization that includes \(b\) in the sum. Duea further found that the normalization that excludes \(b\) generally performs better than the others, suggesting that stability relative to data translations has empirical as well as theoretical significance.

We may note an element in these observations that runs counter to intuition. It might be thought that stability for the second normalization would imply stability for the first, since any solution available to the second is available to the first on allowing for scaling of the \(x_j\) weights. However, reflection discloses that scaling implicitly affects the objective function evaluation of the solution, and hence, for any situation where the sum of the \(x_j\) weights is appropriately nonzero, the second normalization is less restrictive than the first.

**ISSUES OF MODEL GENERALITY**

We have seen that the “unnatural” difficulties (such as degeneracy) that have been associated with LP discriminant formulations can be removed simply by utilizing a more appropriate normalization, provided this is accompanied by allowing for normalization constants of both signs and bounding the feasible region. The use of limited models, however, still may be expected to exhibit a limited degree of discrimination power. Empirical studies have demonstrated the performance characteristics of the simpler models, and thus the groundwork has been laid for studies that examine broader considerations such as differential weightings for “internal” and “external” deviation variables associated with each data point (the \(d_i\) and \(a_i\) variables, respectively, of [4]). As long as the objective function weight for an external deviation variable has greater absolute value than the oppositely signed weight for the associated internal deviation variable, at least one of the two variables will be zero in any LP solution, which is a desirable feature in such a model.

Analysis further suggests that appropriate weights for different groups should generally bear an inverse relation to the numbers of elements in these groups (or to the ratios of sample sizes to population sizes for the groups), and that a process of iterative postoptimization that decreases weights associated with outliers should tend to increase the number of points classified correctly. Among the more interesting phenomena inviting investigation (and resolution by appropriate parameter settings) are “trivial discriminations”—that is, those in which points of both groups lie on the same hyperplane and hence are not strictly separate. (One possibility in pursuit of strict separation is to shift \(b\) by a constant for one of the groups, although the numerical effects of choosing such a constant may be subtle.)

Another factor of interest derives from the fact that extreme-point solutions in the presence of multiple optima generally will yield hyperplanes that can be shifted to a more “central” location by changing \(b\), corresponding implicitly to the identification of an optimum which is a convex combination of extreme points. Such a shift can be calculated readily by reference to the LP solution values of internal and external deviation variables (which may be viewed as components of \(L_1\) norm
distances). A simple analogous calculation, exploiting the fact that the problem is reduced to a single dimension when only $b$ changes, can be used to shift $b$ progressively farther (in its two available directions) to identify an optimal value for minimizing the number of misclassifications, given the values of the $x_j$ variables.

Beyond these considerations, however, is a need to transcend the classical view of discrimination by amending the criterion that minimizes the number of misclassifications as an implicit measure of the merit of alternative approaches. (Such a goal could, of course, be expressed in an integer programming formulation,\(^1\) provided the amount of improvement in classification warranted the extra computational burden.) It is not difficult, for example, to conceive of situations where the underlying goal should properly dictate an increase in the number of misclassifications in order to obtain a stronger differentiation of points classified correctly. Investigations of such concerns provide an opportunity to make discriminant analysis increasingly relevant to intriguing areas of practical application, such as the realm of pattern recognition in artificial intelligence. [Received: September 30, 1985. Accepted: January 7, 1986.]

REFERENCES


Ned Freed is Associate Professor of Management at the University of Portland. He received his doctorate in management science from the University of Colorado and holds an M.B.A. from the Wharton School. Dr. Freed has published a number of articles on the application of mathematical programming techniques to statistical problems in such journals as Decision Sciences, International Journal of Operations Research, and Communications in Statistics.

Fred Glover is Professor of Management Science at the University of Colorado. He has been sponsored as a National Visiting Lecturer by TIMS and ORSA and has served in the U.S. National Academy of Sciences Program of Scientific Exchange. In addition to holding editorial and publication committee posts for Management Science and Operations Research, Dr. Glover has published more than 150 articles in the field of mathematical and computer optimization, particularly in the areas of integer programming, networks, and large-scale systems. Dr. Glover has received awards and honorary fellowships from the American Association for the Advancement of Science, NATO, Decision Sciences Institute, Colorado Energy Research Institute, Ford Foundation, American Association of Collegiate Schools of Business, Miller Research Institute, and TIMS College of Practice. He also consults widely for industry and government.