Working Paper

Nontraditional Approaches to Statistical Classification: Some Perspectives on L_p-Norm Methods

Antonie Stam

WP-96-128
December 1996

IIASA International Institute for Applied Systems Analysis □ A-2361 Laxenburg □ Austria
Telephone: +43 2236 807 □ Fax: +43 2236 71313 □ E-Mail: info@iiasa.ac.at

Nontraditional Approaches to Statistical Classification: Some Perspectives on L_p-Norm Methods

Antonie Stam

WP-96-128
December 1996

Department of Management, Terry College of Business
The University of Georgia, Athens, GA 30602
and
International Institute for Applied Systems Analysis
Laxenburg, Austria

IIASA International Institute for Applied Systems Analysis □ A-2361 Laxenburg □ Austria
Telephone: +43 2236 807 □ Fax: +43 2236 71313 □ E-Mail: info@iiasa.ac.at

Nontraditional Approaches to Statistical Classification: Some Perspectives on L_p-Norm Methods

Antonie Stam
Department of Management
Terry College of Business
The University of Georgia
Athens, GA 30602

October 1996

Acknowledgment: This research was supported in part by a Terry Summer Research Fellowship from the University of Georgia.

Acknowledgement: The author thanks Ogi Asparoukhov, Pedro Duarte Silva, Carl Huberty and Cliff Ragsdale for their stimulating comments regarding the topic of this paper.

Nontraditional Approaches to Statistical Classification: Some Perspectives on L_p-Norm Methods

ABSTRACT

The body of literature on classification methods which estimate boundaries between the groups (classes) by optimizing a function of the L_p-norm distances of observations in each group from these boundaries, is maturing fast. The number of published research articles on this topic, especially on mathematical programming (MP) formulations and techniques for L_p-norm classification, is now sizable. This paper highlights historical developments that have defined the field, and looks ahead at challenges that may shape new research directions in the next decade.

In the first part, the paper summarizes basic concepts and ideas, and briefly reviews past research. Throughout, an attempt is made to integrate a number of the most important L_p-norm methods proposed to date within a unified framework, emphasizing their conceptual differences and similarities, rather than focusing on mathematical detail. In the second part, the paper discusses several potential directions for future research in this area. The long-term prospects of L_p-norm classification (and discriminant) research may well hinge upon whether or not the channels of communication between on the one hand researchers active in L_p-norm classification, who tend to have their roots primarily in decision sciences, the management sciences, computer sciences and engineering, and on the other hand practitioners and researchers in the statistical classification community, will be improved. This paper offers potential reasons for the lack of communication between these groups, and suggests ways in which L_p-norm research may be strengthened from a statistical viewpoint. The results obtained in L_p-norm classification studies are clearly relevant and of importance to all researchers and practitioners active in classification and discrimination analysis. The paper also briefly discusses artificial neural networks, a promising nontraditional method for classification which has recently emerged, and suggests that it may be useful to explore hybrid classification methods that take advantage of the complementary strengths of different methods, *e.g.*, neural network and L_p-norm methods.

Keywords: Classification Analysis, Discriminant Analysis, L_p-Norm Estimation, Mathematical Programming.

Nontraditional Approaches to Statistical Classification: Some Perspectives on L_p-Norm Methods

1. Introduction

Discriminant is almost as old as mankind. In Deuteronomy 21 of the Bible (NASB) Moses declares that[1]

> "If a slain person is found lying in the open country in the land which the Lord your God gives you to possess, and it is not known who has struck him, then your elders and your judges shall go out and measure the distance to the cities which are around the slain one. It shall be that the city which is nearest to the slain man, that is, the elders of that city, shall take a heifer of the herd, which has not been worked and which has not pulled in a yoke ..."

Thus, thousands of years ago, long before Fisher, Smith and Mahalanobis, Moses already suggested the use of distance measures for solving discriminant problems. Statistical discriminant analysis can be used for one of two purposes: (a) description or discrimination, where the goal is to identify the set of variables which maximally discriminates one group from the others, and (b) classification or prediction, where the interest is focused on correctly classifying observations into well-defined groups, based on their characteristics, when group membership is either known or unknown (Huberty 1984; Joachimsthaler and Stam 1990). The primary subject of this paper is the second purpose of discriminant analysis, *i.e.*, the classification problem of discriminant analysis.

Define the class of L_p-norm classification methods as those methods that directly estimate the boundaries of each class (group) by optimizing some function of the L_p-norm distances of a set of observations in each group from these boundaries. These boundaries are defined by surfaces that separate the groups. Since its inception in the mid 1960s, the body of research in L_p-norm classification, and particularly in mathematical programming (MP)-based formulations for solving the classification problem, has grown and matured considerably. Inspired by problems in pattern recognition, initial L_p-norm classification research focused primarily on MP-technical aspects of the formulations, rather than on comparing the classification accuracy of L_p-norm rules *vis à vis* other classification methods. The popularization of L_p-norm methods in the early 1980s led to an impetus of novel formulations, and infused new energy into the field. In the second half of the 1980s and the early to mid 1990s, the focus of research in L_p-norm classification has shifted gradually to issues related to the relative classification accuracy of various different L_p-norm methods, although the research on refined formulations and methods has continued to prosper as well.

This paper first defines the basic concepts of L_p-norm classification analysis, and reviews – within a unified framework – what the author views as some of the main trends and issues that have helped shape the current state of L_p-norm classification research. Then, the paper continues by identifying several ways in which L_p-norm classification research may be strengthened, especially from a statistical viewpoint. One issue of general concern is that, whereas it has spurred considerable discussion within the management science and engineering fields, the L_p-norm approach to classification has not attracted much attention in statistical circles, posing a serious challenge to researchers in the field of L_p-norm classification (and discrimination) to reach out and catch the interest of the statistical community. The long-term prospects of this line of research may well hinge upon whether or not the channels of communication between researchers active in L_p-norm classification and researchers in the statistical classification community will be improved. This paper offers potential reasons for the lack of communication between these groups, and suggests ways in which a bridge between these groups may be forged. The paper also discusses some recently emerged research directions that have attracted much attention, *e.g.*, promising new L_p-norm methods, efficient algorithms for solving the L_0-norm method, and nontraditional classification methods other than L_p-norm ones.

2. Review of the Most Popular L_p-Norm Methods

Morrison (1990, p. 1) notes that "every statistical analysis must be built upon a mathematical model linking observable reality with the mechanism generating the observations." Thus, a model is a conceptual abstraction of reality, and the relationships between the variables are captured in the mathematical formulae. After making relevant model assumptions, such as the optimization criterion and distributional properties, an appropriate method is selected to estimate the model parameters. Specific techniques or algorithms are then used to determine the optimal solution according to the method selected. For instance, regression methods can be used for estimation within the framework of the "general linear model," with calculations performed using some specialized algorithm. To date, almost all L_p-norm research has assumed that the classification model is known *a priori*, and has focused on methods for solving the (already known) classification model. Whereas many of the L_p-norm classification methods are conveniently formulated as MP problems, and MP optimization techniques – such as linear programming (LP), nonlinear programming (NLP) and mixed integer programming (MIP) – provide an efficient vehicle for solving these models, several L_p-norm methods can be solved using alternative (non-MP-based) algorithms as well. Thus, L_p-norm methods may or may not be formulated as MP problems and solved using MP techniques or algorithms. In fact, a few of the classification algorithms cited in this paper are not MP-based. Therefore, this paper refrains from using the potentially confusing term "MP methods," and uses "L_p-norm methods" instead.

Moreover, in an attempt to distinguish between on the one hand conceptual foundations of L_p-norm classification methods and on the other hand MP formulations, techniques and algorithms which are typically used to solve these formulations, formulations of the various L_p-norm methods are not presented in the traditional MP format.

Rather than reviewing a plethora of different L_p-norm classification methods and MP formulations used to solve these methods in detail, this section will introduce the classification problem conceptually, integrate a number of the most important formulations within a unified framework, and briefly highlight some of the differences between and similarities of these formulations. For a detailed description of each individual method and formulation, the reader is referred to the original papers in which these were first introduced, many of which are cited in this paper. Throughout, the mathematical detail is kept to a minimum.

2.1. The Classification Problem

This paper will only define the case of two-group classification explicitly. The concepts underlying the two-group case can be generalized to multiple groups (in several different ways), but the notation becomes tedious and complex. Consider the problem where an observation i is to be classified into one of two groups, G_1 or G_2, based on a q-dimensional vector of attributes $\mathbf{x}_i = (x_{i1}, \ldots, x_{iq})^{\mathrm{T}}$, such that $R_1 \cup R_2 \cup R_U = \Re^q$, $R_1 \cap R_2 = \emptyset$, and $R_j \cap R_U = \emptyset$, $j = 1, 2$, where R_j is the region of the attribute space $\Re^q$ assigned to G_j and R_U is the region for which the group assignment is undetermined (preferably, $R_U = \emptyset$), sometimes called the classification gap. Denote the classification score of observation i by $f(\mathbf{x}_i)$, where $f(\mathbf{x}_i)$ is some function of the attribute variables.

Classification Rules

After first introducing a general framework for L_p-norm classification, the major methods and corresponding MP formulations and techniques will be reviewed in detail. The rule for the two-group L_p-norm problem is defined by $R_1 = \{\mathbf{x} \mid f(\mathbf{x}) < c\}$, $R_2 = \{\mathbf{x} \mid f(\mathbf{x}) > c\}$ and $R_U = \{\mathbf{x} \mid f(\mathbf{x}) = c\}$, where the value of c is called the cut-off value. The surface defined by $f(\mathbf{x}) = c$ establishes the boundary between G_1 and G_2. In the MP problem formulation, the classification rule, augmented with appropriate deviational (distance) variables, is represented by constraints, one constraint for each training sample observation. In these constraints, most MP formulations either relax one (or both) of the strict inequalities in R_j to include the case $f(\mathbf{x}) = c$, or allow for a classification gap, defining the group assignment regions by $R_1 = \{\mathbf{x} \mid f(\mathbf{x}) \leq c\}$, $R_2 = \{\mathbf{x} \mid f(\mathbf{x}) \geq c + \epsilon\}$ and $R_U = \{\mathbf{x} \mid c < f(\mathbf{x}) < c + \epsilon\}$, with $\epsilon > 0$ (Erenguc and Koehler 1990; Koehler 1991a). In the latter case, after estimating the classification function additional rules are needed to classify observations with scores between c and $c + \epsilon$ (Ragsdale and Stam 1991; Stam and Ragsdale 1992).

In early research, only linear classification functions of the form $f(\mathbf{x}_i) = \sum_{j=1}^{q} a_j x_{ij} = c$ were considered. Similar to what is common practice in regression modeling, recently nonlinear transformations of the attributes, $y_{ir} = g(\mathbf{x}_i)$, have been used as well, implying a classification rule $h(\mathbf{y}_i) = \sum_{r=1}^{t} b_r y_{ir}$, which is nonlinear (*e.g.*, quadratic, second order, polynomial) in terms of the original attributes (Banks and Abad 1994; Duarte Silva and Stam 1994b; Rubin 1994). The components of $\mathbf{a} = (a_1, \ldots, a_q)^T$ or $\mathbf{b} = (b_1, \ldots, b_t)^T$, along with c, are the classification function coefficients to be estimated. Since $f(\mathbf{x}_i)$ is linear in the original attributes, $f(\mathbf{x}_i) = c$ defines a linear hyperplane; if $h(\mathbf{y}_i) = c$ is nonlinear in the original attributes, then it defines a (nonlinear) separating surface. For the sake of generality, in the remainder of this paper classification functions of the form $h(\mathbf{y}_i)$ will be used.

In L_p-norm classification methods, the "optimal" classification rule is determined by minimizing a function of the undesirable deviations (or, in some methods, by maximizing a function of the desirable deviations) of all training sample observations from the surface separating the groups. The undesirable (external) deviations $d_i \geq 0$ and desirable (internal) deviations $e_i \geq 0$ from the separating surface are determined using the equalities in (1),

$$\left.\begin{aligned} h(\mathbf{y}_i) - d_i + e_i &= c, \text{ if observation } i \in G_1, \text{ and} \\ h(\mathbf{y}_i) + d_i - e_i &= c, \text{ if observation } i \in G_2, \end{aligned}\right\} \tag{1}$$

Thus, for given $h(\mathbf{y}_i)$ and c, an observation i is misclassified if and only if (iff) $d_i > 0$, *i.e.*, iff the observation is located on the "wrong" side of the separating surface. Similarly, e_i measures the extent to which observation i is located on the "correct" side of the separating surface (see also Ragsdale and Stam 1991; Glover *et al.* 1988; Glover 1990). If the classification criterion includes only the d_i, and not the e_i (this is the case, *e.g.*, for classification methods which focus solely on misclassification), (1) is replaced by (2),

$$\left.\begin{aligned} h(\mathbf{y}_i) - d_i &< c, \text{ if observation } i \in G_1, \text{ and} \\ h(\mathbf{y}_i) + d_i &> c, \text{ if observation } i \in G_2, \end{aligned}\right\} \tag{2}$$

where the classification of observations with $h(\mathbf{y}_i) = c$ is yet to be resolved.

Mangasarian (1968) proposes a generalization of the linear rule in (2), in which the group boundaries are defined by a multi-surface, piecewise linear separator. To date, the classification accuracy of Mangasarian's method has not been evaluated for various data conditions. As discussed below, this method is related to recently proposed artificial neural network methods for solving the classification problem (Bennett and Mangasarian 1992a).

Classification Criteria

Focusing on minimizing misclassification, the most widely used class of L_p-norm optimization criteria is given in (3),

$$\text{minimize } z_p = \Big(\sum_{i=1}^{n} (d_i)^p\Big)^{1/p}, \tag{3}$$

where n represents the number of training sample observations. In the generic MP formulation for L_p-norm classification, (3) is optimized, subject to (2) and nonnegativity constraints $d_i \geq 0$. Essentially, MP formulations for L_p-norm classification may be viewed as goal programming (GP) formulations (Ignizio 1982; Stam 1990).

The class of L_p-norm criteria includes as special cases the L_1-norm criterion, which minimizes $\sum_{i=1}^{n} d_i$ (minimize the sum of deviations, *a.k.a.* the MSD method), the L_∞-norm criterion, minimize $lim_{p\to\infty} z_p$, which minimizes $d = \max_i\{d_i\}$ (minimize the maximum deviation, *a.k.a.* the MMD method), and the L_0-norm criterion, minimize $lim_{p\downarrow 0} z_p$, which minimizes $\sum_{i=1}^{n} \delta_i$, where δ_i is a binary variable which equals 1 iff $d_i > 0$ (minimize the number of misclassifications, typically solved using MIP, *a.k.a.* the MIP method), and 0 otherwise. The rationale for considering various different values of p is that extreme observations are emphasized more heavily as p increases, so that L_p-norm methods with "low" values of p may yield potentially robust classification results in the presence of extreme observations in the training sample. A graphical representation of different L_p-norm distance measures can be found in Joachimsthaler and Stam (1990) and Stam and Joachimsthaler (1989).

In terms of computational aspects of solving L_p-norm methods, any criterion with $p < 1$ is concave, and as a result the global optimal solution is difficult to identify (this explains why the computational requirements of the L_0-norm method are substantial). L_p-norm methods with $p \in (1, \infty)$ have a convex criterion with a unique optimal solution and can be solved using NLP techniques (Stam and Joachimsthaler 1989). The L_1- and L_∞-norm methods can be solved using LP techniques, but the optimal solutions are typically non-unique. The optimal solution of the L_0-norm method is typically non-unique.

2.2. Specific Methods

The first two-group L_p-norm classification method was the MSD method, proposed in the 1960s (Minnick 1961; Koford and Groner 1966; Mangasarian 1965; Smith 1968), followed in the 1970s by the MIP method (Ibaraki and Muroga 1970; Liitschwager and Wang 1978; Warmack and Gonzalez 1973). With the exception of Koford and Groner (1966) and Warmack and Gonzalez (1973), these methods were solved using MP techniques. Conducted mostly in the fields of engineering and management science, early studies focused primarily on the geometric properties of novel L_p-norm

methods for pattern recognition, rather than on statistical issues of discrimination and classification. The methods were typically illustrated by example.

The popularity of L_p-norm classification methods experienced a substantial impetus in the early 1980s, due largely to the work by Freed and Glover (1981a, 1981b), who analyzed the MSD method and introduced the MMD method, and Bajgier and Hill (1982), who conducted extensive simulation experiments involving several different L_p-norm methods developed at that time, including the MMD, MSD and MIP. In his seminal discriminant analysis text, Hand (1981) forwarded a general L_1-norm method, of which the MSD is a special case. Early research on the MIP method also includes the little-known work by Asparoukhov (1985). While continuing to deliver refinements of existing methods and novel variants of previous MP-based methods for L_p-norm classification, research in the late 1980s and the early to mid 1990s increasingly shifted its focus to the issue of classification accuracy.

During the early 1980s, it was discovered that several of the originally proposed MP-based formulations for, *a.o.*, the MSD, MMD and MIP, were fundamentally flawed, and could easily result in unbounded ($z = \pm\infty$), unacceptable or trivial ($h(\mathbf{y}) \equiv 0$, or $\mathbf{b} = \mathbf{0}$ and $c = 0$) and improper ($h(\mathbf{y}_i) = c$, $\forall i \in G_1 \cup G_2$) solutions (Freed and Glover 1986b; Koehler 1989a, 1989b, 1990, 1991a; Markowski and Markowski 1985). Moreover, the solutions of several formulations were shown to be sensitive to data transformations (Freed and Glover 1986b; Glover *et al.* 1988; Markowski and Markowski 1985). Rubin (1990b) notes that the MSD problem may have to be solved twice, once with the original group designation and again with the groups reversed, after which the solution with the greater classification accuracy should be selected. Various modified formulations were proposed to remedy these problems. Most of these included normalization schemes, which usually involved either including a normalization constraint in the problem formulation, *e.g.*, $\sum_r b_r + c = 1$, or exploring different ways to deal with the classification gap. Unfortunately, most of the supposed cures caused other problems. For instance, the normalization constraint mentioned above precludes certain separating surfaces from consideration.

The normalization proposed by Glover (1990) in (4) resolves the problems with previous formulations, provided that the mean vectors $\bar{\mathbf{y}}_j = n_j^{-1} \sum_{i \in G_j} \mathbf{y}_i$, $j = 1, 2$, are different,

$$\left(-n_2 \sum_{i \in G_1} \mathbf{y}_i^{\mathrm{T}} + n_1 \sum_{i \in G_2} \mathbf{y}_i^{\mathrm{T}}\right) \mathbf{b} = 1, \tag{4}$$

where n_j is the number of training sample observations in G_j. If $\bar{\mathbf{y}}_1 \neq \bar{\mathbf{y}}_2$, a rule of the type of (4) cannot be an effective classifier.

Ragsdale and Stam (1991) and Stam and Ragsdale (1992) present alternative ways to deal with the classification gap, which also resolve the problems with unacceptable and improper solutions, and is insensitive to linear transformations of the data. Cavalier *et al.* (1989) propose adding a

constraint of the type $\| \mathbf{b} \| = 1$, which, although preventing unacceptable solutions, changes an LP problem into a non-convex programming problem which is hard to solve. Bennett and Mangasarian (1992b) develop a single LP formulation which generates a plane that minimizes an average sum of undesirable deviations. Their formulation does not require additional constraints which rule out certain solutions from consideration. Wanarat and Pavur (1996) note that the inclusion of second-order and cross-product terms of the attributes guarantees that the basic MSD and MIP methods are invariant to nonsingular transformations of the data, but that this is not the case if the cross-product terms are omitted. Xiao (1994a) derives necessary and sufficient conditions for unacceptable solutions in NLP classification analysis.

Further discussions of issues related to the occurrence and prevention of unbounded, unacceptable and improper solutions in LP and MIP formulations, and normalizations that can remedy some of these problems can be found in Cavalier *et al.* (1989), Erenguc and Koehler (1990), Glover (1990), Koehler (1994), Ragsdale and Stam (1991), Rubin (1989a, 1989b), and Xiao (1993, 1994b).

Computational Aspects of the L_0-Norm Method: An NP-Hard Nut to Crack

The problem of solving the L_0-norm method is NP-hard, and MIP solution algorithms are computationally very intensive. Other than the L_0-norm method and its variants which require MIP algorithms, there do not appear to be substantial computational problems with L_p-norm methods. The computational requirements in the L_0-norm method depend critically on the tightness of the constraints, *e.g.*, the value of "Big M," on the number of integer variables, *i.e.*, the number of training sample observations (Hillier and Lieberman 1990, p.467), and on the number of attributes (or functions thereof) (Duarte Silva and Stam 1996). Using standard MP packages it is virtually prohibitive to solve problems with over 100 training sample observations within reasonable CPU time, even on a mainframe (Stam and Joachimsthaler 1990). However, in real applications it is not unusual that classification problems have training samples with 1,000 observations or more – in particular in image recognition problems. Such problems are difficult to solve using MIP techniques, and require efficient special-purpose algorithms which seek to take advantage of the special structure of the L_p-norm classification problem.

Some special-purpose MP-based algorithms solve to exact optimality (Banks and Abad 1991; Duarte Silva and Stam 1996; Koehler and Erenguc 1990; Marcotte and Savard 1991; Marcotte *et al.* 1995), others are exact non-MP-based algorithms (Soltysik and Yarnold 1993, 1994; Warmack and Gonzalez 1973), or heuristic algorithms (Abad and Banks 1993; Koehler 1991b; Koehler and Erenguc 1990; Marcotte *et al.* 1995; Rubin 1990a). The MultiODA software (Soltysik and Yarnold 1994) and the Divide-and-Conquer algorithm (Duarte Silva and Stam 1996) appear to be the fastest special-purpose algorithms developed to date. The algorithm of Duarte Silva and Stam (1996), which

decomposes the overall problem, can reportedly solve problems with 1,000 training sample observations and 2 attributes on a 486 DX2 (66 Mhz) PC with 16 MB of RAM in slightly over 1 CPU minute.

Weighted Formulations and Secondary Criteria

In the class of classification criteria in (3), each deviation is weighted equally, which – depending on the application, the prior probabilities of group membership and the costs of misclassification – may not always be appropriate. It is possible to assign different weights to each component and to include both desirable and undesirable deviations in (3), in which case some measures of misclassification and correct classification are optimized simultaneously. An example of an L_p-norm criterion which minimizes a weighted function of both the desirable (d_{ij}) and desirable (e_{ij}) deviations is given in (5),

$$z = \left(\sum_{j=1}^{2} \sum_{i=1}^{n_j} \left(w_{ij}^{U} (d_{ij})^p + w_{ij}^{D} (e_{ij})^p \right) \right)^{1/p}, \tag{5}$$

where w_{ij}^{U} and w_{ij}^{D} are the weights associated with d_{ij} and e_{ij} from the separating surface of observation $i \in G_j$, respectively, and n_j is the number of training sample observations in G_j. The criterion in (5) allows for observation-specific weights, group-wise weights and weights that depend on whether an observation is classified correctly or incorrectly. An example the OSD (optimize the sum of deviations), which optimizes a weighted sum of external and internal deviations (Bajgier and Hill 1982). Although (3) and (5) are fairly general, other types of deviations and weighted criteria have been proposed as well, such as the MSID (maximize the sum of internal deviations), which simultaneously minimizes the maximum of the weighted external deviations and maximizes the sum of the weighted internal deviations (Freed and Glover 1986). In its most general form, the criterion of the Hybrid method (Glover *et al.* 1988) includes both individual and group-specific deviations, desirable as well as undesirable, each of which can be assigned a different weight. Thus, the OSD, MSID and Hybrid methods essentially extend the MSD criterion, incorporating additional information. For reasons of brevity we do not explicitly state the multitude of different variations of weighted criteria proposed in the literature.

Weighted criteria are particularly useful in the following two situations. First, a variant of the L_0-norm (MIP) method, in which the δ_i in the objective function are weighted by the relevant misclassification costs, can be used to determine the classification rule that minimizes the training sample misclassification costs *directly*. This in itself is a significant contribution to statistical classification. The weighted criterion that minimizes the training sample misclassification cost is given by (6),

$$\frac{\pi_1 C(2 \mid 1)}{n_1} \sum_{i \in G_1} \delta_i + \frac{\pi_2 C(1 \mid 2)}{n_2} \sum_{i \in G_2} \delta_i, \tag{6}$$

where π_j is the prior probability of membership in G_j, $j = 1, 2$, and $C(r \mid s)$ is the cost of classifying an observation from G_s into G_r, $r, s = 1, 2$; $r \neq s$.

The second motivation for using a weighted classification criterion is that LP and MIP formulations for L_p-norm classification, such as the MSD, MMD and MIP, commonly have multiple optimal solutions, implying the existence of several non-equivalent classification functions which are optimal with respect to this (primary) classification criterion. Hence, it is recommended to use a secondary criterion to resolve ties in the optimal solution of the primary classification criterion. For example, a useful secondary criterion for the MSD and MIP which provides relevant additional information about the characteristics of the groups is to maximize correct classification (*e.g.*, to maximize $\sum_i e_i$). A secondary criterion can be included in the method by solving a weighted problem formulation. To ensure that the secondary criterion never interferes with optimizing the primary criterion, the problem should be solved as a lexicographic GP problem, with weights of the secondary criterion that are sufficiently smaller than those of the primary criterion. Different tie-breaking schemes can be found in Bajgier and Hill (1982), Erenguc and Koehler (1990) and Duarte Silva (1995).

In the Hybrid and weighted MP methods with $p>0$ it is not easy to identify appropriate weights (justified from a statistical perspective) for the elements in the objective function, and the interpretation of the results may be complex, because any relationship of these weights with misclassification costs is indirect. Glover *et al.* (1988) indicate restrictions on the criterion weights in the Hybrid method that guarantee that the problem can be solved, but offer no general guidelines on how to select these weights within these limits. As a cautionary note, methods for which the d_i (or d_{ij}) are to be minimized and the e_i (or e_{ij}) are to be maximized simultaneously can be tricky, and may easily lead to unbounded solutions, unless designed carefully.

Extensions to Multiple-Group Classification

Several researchers have suggested extensions of the two-group L_p-norm classification case to more than two groups. Freed and Glover (1981a) propose to first decompose the overall m-group classification problem into $m(m-1)/2$ two-group problems, then solve each two-group problem, and then determine classification rules based on these solutions. However, in doing so the group assignment in some of the segments in attribute space created by the $m(m-1)/2$ pairwise separating surfaces may not be clear, and the resulting classification scheme may be sub-optimal because the problem is not estimated in aggregate form.

Gehrlein (1986) proposes several generalizations of the two-group MIP formulation, in which the classification of all groups is done simultaneously. One formulation involves a single classification

function with group-specific cut-off values (intercept terms), implying that the slope of the surfaces separating the groups is the same. Thus, the separating surfaces divide the classification scores in intervals, one for each group, and observations are classified into the group associated with the interval in which its classification score falls. Gehrlein (1986) also proposes an MIP formulation with multiple classification functions, one for each group, in which each observation is assigned to the group with the largest discriminant score. Note that this classification strategy is also used in statistical multiple-group classification methods. However, Gehrlein's (1986) formulations require a considerable number of binary variables, and are computationally infeasible for medium-size or large training sample data sets, given the current state of MP software technology.

Gochet *et al.* (1996) propose an extension of the two-group MSD method to multiple groups which is related to the LINMAP method for multidimensional scaling (Srinivasan and Shocker 1973), and which uses measures of goodness and badness of fit to simultaneously estimate pairwise classification functions for each pair of groups. As in Gehrlein's (1986) method, an observation is classified into the group with the highest discriminant score. Gochet *et al.* (1996) show that their method is similar conceptually to a variant of the Hybrid method (Glover *et al.* 1988), with the minimax deviations omitted. The advantage of the formulation by Gochet *et al.* (1996) is that all groups are considered simultaneously, and the formulation does not require any integer variables. Moreover, the authors derive several results on the behavior of their method for various special cases and data conditions, and show through an analysis of real data sets that their method can yield good classification results. However, in certain cases their method involves a sequential estimation of sub-problems, and without special-purpose software the method may not be easy to apply.

Choo and Wedley (1985) develop multiple-group L_0- and L_1-norm methods to determine implicit optimal criterion weights in repetitive discrete multicriteria decision making. Pavur and Loucopoulos (1995) extend formulations for the two-group MIP method to the case of multiple groups. Both Choo and Wedley (1985) and Pavur and Loucopoulos (1995) use a single classification function, with different cut-off variables for each group.

Multiple-group methods with multiple classification functions that are general and allow for separating surfaces which intersect each other divide the attribute space into multiple different segments, each of which is assigned to exactly one group. However, multiple-group extensions which are limited to only one classification function with different cut-off variables, implying separating surfaces with equal slopes that do not intersect, limit the division of the attribute space to layers. In the case of linear classification functions, the estimation of a single function with different cut-off variables implies that the separating planes are parallel. Hence, in this method it is assumed that the attributes in the classification model define a monotonic relationship between the groups. Whereas it is easy to design a multi-group simulation experiment for which this assumption is satisfied and for which

the single-function method gives good classification results, real data sets will rarely have these characteristics, rendering this method of limited use in practice. Clearly, the approach of estimating multiple classification functions that are allowed to intersect is more flexible and general, and therefore preferable to methods that estimate a single function. However, there is usually a price to be paid for generality, and as mentioned above flexible formulations such as the those by Gehrlein (1986) and Gochet *et al.* (1996) are more complex as well.

Other Formulations

Various other creative L_p-norm classification methods have been proposed, more than can be reviewed in this paper. Here, only a representative few are highlighted. Nath (1984) derives expressions for the misclassification probabilities for several two-group L_p-norm classification methods ($p = 0,\ 1,\ \infty$), for contaminated multivariate normal attribute distributions. Lam, Choo and Wedley (1993) develop a method, solved with MP, which takes probability of misclassification into account, whereas Lam and Choo (1993) use LP to solve classification problems with nonmonotonic attributes. Lam, Choo and Moy (1996) propose an interesting MP-based method which presumes that elements of the same groups should have similar characteristics in terms of summary statistics, and minimizes the sum of deviations from the group mean. Yarnold (1996) reports that promising classification results can be achieved by applying the UniODA method (Soltysik and Yarnold 1993) within a framework of nonlinear classification tree analysis. Markowski (1990) develops formulations which take error balancing into account. Markowski (1994) proposes an adaptive statistical classification method in which, depending on which method minimizes the sum of overall classification accuracy and provides the most balanced classification results on the training sample, the LDF, QDF or a nearest neighbor method is used for evaluating validation samples. It appears that L_p-norm methods could be included in this framework as well. Markowski (1994) concludes that this adaptive procedure is an effective alternative to both statistical and L_p-norm classification methods.

Descriptive L_p-Norm Discriminant Methods

As noted above, the vast majority of L_p-norm classification methods has assumed the classification model as given, and concerned itself with selecting appropriate classification methods. With a few exceptions (Glorfeld and Olson 1982; Nath and Jones 1988) the research has focused on prescriptive discriminant analysis, rather than on issues related to descriptive discriminant analysis.

Nath and Jones (1988) develop a variable selection rule (criterion) based on the jackknife methodology to distinguish significant from non-significant attributes for use in L_p-norm methods ($p = 1,\ \infty$). By using this procedure, the analyst is able to develop a parsimonious discriminant model; obtain measures of variability of the parameter estimates, allowing for the assessment of the stability

of the estimates; and rank order the variables in terms of their discriminant ability, based on the relevant p-value. The Nath and Jones (1984) procedure is important, because issues involving the variable selection problem and descriptive discriminant analysis have not been explored much within the L_p-norm classification framework. Obviously, rigorous descriptive L_p-norm methods would strengthen its prescriptive counterpart, L_p-norm classification, and vice versa. Thus, further research in this area can be very useful. Meanwhile, practitioners wishing to use L_p-norm classification methods may need to resort to the usual statistical methods for exploratory data analysis and statistical discriminant analysis techniques in order to explore general characteristics of the problem.

3. Why Are These Methods of Interest?

3.1. Intuitive Appeal of L_p-Norm Methods for Classification

Geometric Interpretation

The geometric interpretation of L_p-norm classification methods, with their separating surfaces and distances measures from this surface, clearly has intuitive appeal. Similar to least absolute deviation (LAV) methods in regression analysis, L_p-norm methods with $p < 2$ can yield robust classification rules in the presence of outlier observations, or if the data are skewed. Whereas L_0- and L_1-norm methods do not emphasize extreme observations, methods which minimize the sum of squared errors, such as Fisher's (1936) linear discriminant function (LDF), Smith's (1947) quadratic discriminant function, and the MMD, which minimizes the L_∞-norm, are relatively sensitive to extreme observations. As discussed below in more detail, there is indeed evidence that in the case of problems with outlier-contaminated data conditions and skewed distributions, parametric statistical methods which assume normality do not classify accurately, and L_p-norm and other nonparametric methods fare substantially better. Such data conditions are common in practice, *e.g.*, in business and financial data (Eisenbeis 1977; Glorfeld 1990; Mahmood and Lawrence 1987).

Estimating Probability Distributions Vs. Estimating Separating Surfaces

McLachlan (1992, p. 16) notes that the accuracy of a classification rule depends mostly on how well it can handle entities of doubtful origin, rather than on how well it handles observations of obvious origin. Whereas statistical classification methods such as the LDF and QDF are based on assumptions about the full probability density functions of the attribute variables which describe each group, L_p-norm methods are concerned with estimating the surface separating the groups and focus on the region of group overlap, without making any assumptions about the probability density functions of the attributes. Hence, McLachlan's observation offers a powerful motivation for using L_p-norm methods, in particular in situations where it is difficult to estimate the true probability densities of the groups. Indeed, this may be the case, for instance, if the data are highly non-normal.

Mainline MP Software Can be Used for Solving Most Two-Group L_p-Norm Methods

An attractive property of L_p-norm methods, in particular L_1- and L_∞-norm methods, is that (for two-group problems) these methods are easily formulated and solved as MP problems, as long as one has access to a mainline MP software package and some basic background in optimization techniques. The user does not need to write special-purpose computer programs, except for interface programs to perform data input formatting and report generation tasks. Ironically, as will be discussed later, the very same arguments can easily be turned around and identified as possible reasons why to date statisticians have made little use of L_p-norm classification methods, and why L_p-norm classification is not used by practitioners.

3.2. Evidence of Performance

In spite of their intuitive geometric appeal, no formal general decision-theoretic justification exists for using L_p-norm classification methods, and these methods are not firmly grounded in statistical theory. As a result, it has been difficult to draw general conclusions and make definitive statements about conditions for which the L_p-norm methods are superior to other, competing methods. Like other nonparametric methods, L_p-norm methods need to be evaluated on an *ad hoc* basis, through analyses of real data sets and simulation experiments. Most evaluatory research conducted to date compares L_p-norm methods with well-known statistical classification methods, such as the LDF, the QDF and sometimes logistic regression. However, as will be discussed later, unfortunately few studies have included nonparametric statistical methods such as kernel methods, nearest neighbor methods, recursive partitioning and classification trees.

It has long been established that, if the data are approximately normally distributed, the LDF tends to give the best classification results if the dispersion across groups is similar, and the QDF tends to yield the best results if the dispersion differs substantially, as long as the training sample size is sufficiently large to estimate the QDF parameters accurately (Lachenbruch *et al.* 1973; Rawlings *et al.* 1986). Nevertheless, Freed and Glover (1986a) found the MSD to perform well and to be competitive with the LDF, even if the data are multivariate normally distributed. Whereas Joachimsthaler and Stam (1988) found the MSD to outperform the LDF, for moderate size training samples and substantially different dispersion across groups, they also found the QDF to classify more accurately than the MSD, which is not surprising because under this data condition the optimal rule is quadratic. Rubin (1990b), who compares the MSD, MMD and Hybrid methods with the LDF and QDF, finds the QDF to classify the most accurately if the data are normally distributed. He recommends that studies involving the MSD, MMD and Hybrid methods focus on non-normal data conditions, and that the behavior of these methods be evaluated as the number of attributes and the extent of group overlap are increased. In comparing several statistical, L_p-norm (*a.o.*, MSD, MMD, OSD) and rank-based

methods, Nath, Jackson and Jones (1992) conclude that, while L_p-norm methods compete well with classical statistical procedures under some experimental conditions, their performance generally lags behind that of the classical methods. In their application of classification analysis to credit granting, involving a number of parametric and nonparametric statistical methods along with the MSD and MMD, Srinivasan and Kim (1987) note that the MSD and a variant of the MMD in which group-wise maximum deviations are minimized did not classify very well for their data set.

In a comprehensive literature review of early empirical research in L_1-norm classification, Joachimsthaler and Stam (1990) conclude that the intuitive appeal of these methods in the presence of non-normal distributions with outlier-contamination or highly skewed distributions is confirmed by empirical evidence, especially in studies comparing L_1-norm methods with the LDF and QDF. However, these authors also note that the results obtained in the studies covered in their review are not always consistent, and do not warrant strong conclusions regarding general data conditions for which L_p-norm methods yield superior results.

Stam and Joachimsthaler (1989) analyze various L_p-norm methods ($p = 1,\ 1.5,\ 2,\ 5,\ \infty$), and found the $L_{1.5}$- and L_2-norm methods to perform slightly better than the L_1-, L_5- and L_∞-norm methods for normally distributed data, which confirms similar findings in L_p-norm regression. Hosseini and Armacost (1994) study the performance of six L_p-norm methods, two linear and four nonlinear for multivariate normal data sets with equal group means, with and without outliers, and various levels of dispersion, and conclude that the L_1- and L_p-norm methods ($p = 1.5,\ 2$) perform slightly better than the classical statistical methods in the presence of outliers and if the dispersion differs across groups. However, their study does not include logistic regression, which is known to yield good classification results if the distributions are skewed (Press and Wilson 1978).

Duarte Silva (1995) finds that, while L_0- and L_1-norm methods, with an appropriate criterion to resolve ties, are particularly accurate in classifying problems with few attributes, skewed distributions and small training samples, logistic regression methods generally tend to outperform L_0- and L_1-norm methods for problems with large training samples, skewed distributions and many attributes. Yarnold and Soltysik (1991a, 1991b), Yarnold, Hart and Soltysik (1994) and Yarnold, Soltysik and Martin (1994) report that variants of the ODA and MultiODA L_0-norm methods (Soltysik and Yarnold 1993) perform better in real applications than statistical methods such as the LDF and logistic regression. While the L_0-norm method does appear to yield good classification results for non-normal data conditions, some studies have found this method to give highly variable results, especially if the training sample used to estimate the rule is small (Stam and Joachimsthaler 1990; Stam and Jones 1990). This finding, however, is not confirmed by Duarte Silva (1995), who suggests that this volatility may be due to the choice of (or lack of) secondary criterion. Duarte Silva (1995) uses $\frac{\pi_1 C(2 \mid 1)}{n_1} \sum_{i \in G_1} (d_i - e_i) + \frac{\pi_2 C(1 \mid 2)}{n_2} \sum_{i \in G_2} (d_i - e_i)$, *i.e.*, the weighted extent of misclassification

minus the weighted extent of correct classification in the training sample as the secondary criterion to be minimized.

There is ample empirical evidence confirming that – probably because these methods are strongly affected by extreme observations – the classification performance of the MMD and MSID on validation samples tends to be inferior to L_1-norm methods for almost any data condition (Joachimsthaler and Stam 1990; Markowski and Markowski 1987; Mahmood and Lawrence 1987).

The evidence regarding classification accuracy of weighted L_p-norm methods is sketchy. There is some evidence (Glover *et al.* 1988, Duarte Silva and Stam 1994b) that the inclusion of additional information, as in the Hybrid method, can improve classification performance. In situations that clearly call for a nonlinear classification rule, *e.g.*, if the dispersion matrices differ across groups and the distributions are skewed, L_p-norm methods, in particular L_1-norm methods, with classification rules that use nonlinear transformations of the attributes, may perform well (Banks and Abad 1994; Duarte Silva and Stam 1994b; Wanarat and Pavur 1996), but the inclusion of quadratic and cross-product terms can result in overfitting of the data (Rubin 1994, Wanarat and Pavur 1996).

Summarizing, although in a number of studies L_p-norm methods classified more accurately that statistical methods, not all studies have led to results favorable for L_p-norm methods, indicating that L_p-norm methods do have some merit, but the balance of evidence to date suggests that "the jury is still out." The scope of some evaluatory studies has been limited, so that questions about when to use L_p-norm methods remain. This poses a serious and ongoing challenge to researchers in the L_p-norm classification area, and it is not surprising that critics – particularly in the statistical community – have not been convinced of the potential merits of this class of methods.

4. Other Recent Nontraditional Approaches to Statistical Classification

L_p-Norm Methods for Regression and the Linear Model

Based on the L_1 distance norm, LAV regression has proven to be a robust alternative to least-squares regression methods in the presence of outlier observations (Dielman and Pfaffenberger 1982; Dodge 1987; Gonin and Money 1989; Hample *et al.* 1986; Lawrence and Arthur 1990). Inspired by the successful application of this methodology to regression problems, and noting its conceptual kinship with the MSD method, Lee and Ord (1991) developed an LAV method for solving the classification problem. Although Lau and Post (1992) argue that this classification method cannot yield optimal classification results, it still appears to be useful to investigate whether some of the robust properties of and theoretical results derived for the LAV and L_p-norm regression can be generalized to L_p-norm classification.

Excellent readings volumes, such as Dodge (1987), Gonin and Money (1989) and Lawrence and Arthur (1990) may provide valuable sources of information to this purpose. The ideas discussed in the

following landmark articles, among others, may be of interest as well. Barrodale and Roberts (1970, 1973) and Bassett and Koenker (1978) develop approximation methods for L_1-norm regression. Glahe and Hunt (1970) study small sample properties of L_1-norm methods for the estimation of simultaneous equations. Narula and Wellington (1982, 1990) derive both theoretical and empirical results for L_1-norm regression. Sposito (1990) and Koenker and Portnoy (1987) investigate properties of L_p-norm estimators in regression and linear models.

Artificial Neural Networks

Recently, several artificial neural network (ANN) methods have been applied successfully to classification and pattern recognition problems, especially multi-layer feed-forward neural networks (see, *e.g.*, Jain and Nag 1995; Kattan and Beck 1995; Lippmann 1989; Markham and Ragsdale 1995; Rypley 1994; Subramanian *et al.* 1993; Tam and Kiang 1992). The general appeal of ANN is that these methods are very flexible, do not assume an *a priori* specification of the form of the classification rule, and can represent complex mappings from input space to output space. An excellent and authoritative article, with discussion, about the use of ANN for statistical classification was published by Rypley (1994).

The feed-forward ANN (FFANN) is the most widely used ANN paradigm for classification and pattern recognition, and the remarks that follow refer to this paradigm. In the context of classification analysis, the training of certain types of FFANN has remarkable similarities with the simultaneous fitting of multiple separating surfaces. It is beyond the scope of this paper to introduce FFANNs in detail, or to discuss the classification accuracy of ANN for classification at great length. However, it is very interesting to briefly explore the link between L_p-norm and certain FFANN methods for classification. Although the introduction of some FFANN terminology cannot be avoided, and it is necessary to assume that the reader has some familiarity with FFANN concepts (or will take this opportunity to pick up a book on FFANN), an attempt is made to limit the mathematical expressions and notation to a minimum. For a good introduction of ANN, see Hertz *et al.* (1991) and Wasserman (1989).

Artificial Neural Network Methods and L_p-Norm Methods for Classification

Several researchers have combined LP techniques with the training of FFANNs. Mangasarian (1993) shows the role role of MP, particularly LP, in training FFANNs, and provides illustrations and examples of the use of ANNs in classification analysis. Roy and Mukhopadhyay (1990) introduce a novel method for pattern classification that uses LP formulations. Their method, which extends the LP formulation to obtain group separators with more general shapes, enabling the representation of complex nonlinear class boundaries, can generate FFANN type networks to take advantage of parallel

computation in the classification phase. Through its design, the method avoids certain difficulties of nonlinear optimization of complex functions.

A FFANN is composed of sequential layers, which facilitate the representation of the relationship between the elements in the input layer and the output layer, through intermediate layers of hidden nodes and connections between nodes. Simply stated, the purpose of training a FFANN is to determine the weights of the connections (arcs) and the threshold values of the nodes such that the "true" mapping of inputs to outputs is approximated as accurately as possible. Whereas FFANN methods for classification without hidden nodes can represent linear classification functions only, FFANNs with hidden layers can represent complex nonlinear classification functions. The FFANN is trained to learn the "true" mapping using example input vectors and associated desired outputs. In a FFANN designed for classification analysis, the nodes in the input layer correspond with the attributes, plus perhaps transformations of the attributes and other relevant variables. Each node in the output layer corresponds with exactly one of the groups, and the network output values associated with a given observation indicate the group membership information for this observation. As the FFANN can accomodate any number of output nodes, the FFANN method can easily be used for multiple-group classification.

Each hidden node and output node has an activation function which transforms the input signals of the node into an output signal. The aggregate input I_j into hidden (or output) node j consists of a weighted combination of signals I_{ij} from nodes i in the previous layer to node j, plus a threshold value θ_j, $I_j = \sum_i w_{ij} I_{ij} + \theta_j$, where the w_{ij} is the weight of the arc from node i to node j. Note that this notation is simplified, but it suffices for the purpose of this paper, which is intended to exemplify the connection between FFANNs and L_p-norm classification methods, rather than cover of ANN methods for classification comprehensively. The network error to be minimized is a function of the discrepancy between the desired network output and the actual network output. For instance, this error function may involve a sum of the absolute or quadratic discrepancies.

One way to view the similarities between L_p-norm classification methods and FFANN is as follows. This discussion is similar to the discussion in Wasserman (1989, pp. 29–37) on perceptrons. First, consider a FFANN with *one* hidden layer, and suppose that the activation function of each hidden node and each output node is a step function, such that the output of node j, O_j, equals $O_j = 1$ if $I_j \geq 0$, and $O_j = 0$ if $I_j < 0$. The output signal O_j of hidden node j is binary, and indicates which side of the surface defined by $I_j = 0$ the input vector is most likely located. The output nodes perform the logical "and" function, taking the value 1 in a specific convex segment of the attribute space. For instance, in the case of a two-group classification problem, a FFANN with one hidden layer consisting of m nodes and the step-wise activation function described above can represent piece-wise linear separators with m segments. The propagation of signals in a FFANN with *more than one* hidden

layer, and with the above step-wise activation functions, allows for any non-convex division of the attribute space into areas for each group.

Of course, typically the hidden and output nodes have more general activation functions, *e.g.*, sigmoidal or logistic. Sigmoidal activation functions for the hidden nodes allow "fuzziness" into the division of the attribute space, introducing additional flexibility and complexity into the mapping, and sigmoidal activation functions for the output nodes imply that these nodes can take any value between 0 and 1. In this case, the link with the LP approach is less evident. Sigmoidal activation functions are much more useful, in general, than step functions. For instance, Cybenko (1989) shows that FFANN with hidden layers and sigmoidal activation functions are capable of approximating any input-output relation to any desired degree of accuracy, provided that a sufficient number of hidden nodes is used.

The network training set corresponds directly with the training sample in statistical and L_p-norm classification analysis. In each training pattern, the desired value of the output node corresponding to the group to which the training sample observation belongs equals 1, and the desired value of each other output node equals 0. During the training process, the conflict between desired output (true group membership value) and the actual output (predicted group membership value) determined within the ANN for the training sample observations is minimized, according to some optimization criterion. The backpropagation algorithm (Rumelhart, Hinton and Williams 1986) or a variant thereof is often used to train the FFANN to optimality.

Once the ANN has been trained, the FFANN classification function is implicitly defined by a complex (usually nonlinear) function embodied by the network structure, weights, thresholds and connections. The FFANN classification rule is to assign an observation to the group for which the output node value is the highest. In the case of two groups, this rule reduces to assigning an observation to the group for which the output node value exceeds 0.5. Thus, one interpretation of the FFANN method is that it seeks to approximate the probability of correct classification, but without assuming probability densities for the attributes. Another interpretation is that the output node values provide a "balance of evidence," or fuzzy group membership values.

Although a detailed discussion of the caveats of FFANNs is beyond the scope of this paper, it is important to mention some common technical drawbacks of FFANN training, besides the potential occurrence of local minima. First, there is a danger that during the training process the FFANN gets trapped in a local optimal solution, thus providing inferior classification results. Second, it is possible to overtrain a FFANN, in which case it memorizes training sample patterns, thus reducing the ability to generalize. The FFANN is particularly vulnerable to overtraining if the training sample is relatively small. Third, FFANN are susceptible to network paralysis, which occurs when the weights grow to very large values, without improving the classification performance.

Mukhopadhyay *et al.* (1993) and Roy *et al.* (1995) propose novel ANN-like LP-based methods

that use memory, storing training sample patterns for learning. These authors argue that, in contrast to traditional ANN training, where the network design is usually fixed by the analyst prior to the network training phase, the network design should take place during the training phase, as this corresponds more closely with actual learning in the brain. Their algorithms reflect these ideas, and as such constitute a quite different and creative approach to classification, extending elements of both the LP and FFANN approach to classification.

Comparative Studies Involving Artificial Neural Network Methods for Statistical Classification

Numerous papers have been published on the classification performance of ANN methods, for a review see Rypley (1994). Several studies have compared FFANN methods directly with L_p-norm and statistical methods, with good results. Archer and Wang (1993) and Yoon *et al.* (1993) compare FFANN methods with the LDF, and report results positive for the FFANNs. In studies by Benediktsson *et al.* (1990) and Fischer *et al.* (1994), ANNs outperform statistical methods in applications of classification involving remote sensing data. In a comparison of the ANN with the LDF, QDF and MSD, Patuwo *et al.* (1993) find that ANN methods performs as well or better on training samples, but slightly worse on validation samples. Of course, the findings are difficult to compare across studies, because different studies use different ANN architectures, training schemes and network parameter settings – in addition to different data conditions and validation schemes.

However, whatever the comparative results, like the L_p-norm methods the neural network approach has intuitive appeal, but lacks a decision-theoretic justification. Hence, although flexible, without a theoretical framework the classification performance of ANN methods should be evaluated on a case-by-case basis. The trained ANN has been compared with a "black box," since it is not easy to assign a meaningful interpretation to the multitude of network weights and parameters which together constitute the classification rule. In MP-based L_p-norm classification, the interpretation of the estimated model is not a problem, as the rule defined by R_1, R_2 and R_U is explicit and has a simple form. Therefore, while certainly promising and powerful, ANN classification methods should be interpreted with some caution, just like any other nonparametric method, and it is advisable to conduct a statistical analysis as well, in parallel, in order to assess relative classification accuracy. It is especially important to test the generalizability of ANN methods using validation samples.

Are Artificial Neural Network Methods Appropriate for all Statistical Classification Problems?

There are additional considerations affecting the decision whether or not to use ANN methods that have not been mentioned in many research studies, in that – given the current state of technology and software development – building an ANN is simply not feasible or efficient for all classification applications. The effort of constructing and training an appropriate ANN is time- and expertise-

intensive. ANN methods are certainly attractive if there is sufficient time for constructing and training the network, if the appropriate software is available, and if an analyst with the necessary neural network building expertise is at hand. All of these requirements are met, *e.g.*, in the case of a large bank wanting to develop a screening system for credit applicants, or an investment company seeking to predict turning points in the stock market. Such applications are characterized by a frequent usage of the model, a substantial project development budget and the luxury of a relatively long development period. Large companies often have an in-house R&D team, and can afford to expend the man-power needed for building effective ANNs. Often, these ANN models are embedded in a larger decision support system. Once developed, in such applications the basic ANN structure often remains intact for an extended period of time, but taking advantage of the adaptivity of ANNs the networks are updated frequently, through additional training as new data become available.

With the currently available software technology it is not feasible to build an ANN from scratch if a quick turnaround time is essential, due to the careful effort and considerable time commitment that are required. Small companies with a limited budget may find the use of ANNs prohibitively expensive. For infrequent classification decisions, the effort of building a neural network model may not be worthwhile either, even if the money is available. In those cases, existing statistical methods may be preferred, or L_p-norm methods if the software is available.

It is no coincidence that at present real-life ANN applications are mostly limited to large companies. Considering the tremendous effort of constructing and training separate ANNs for each replication, it is also no coincidence that there exist only a few evaluatory simulation studies involving ANNs, and that most of these are based on few replications or a limited holdout sample analysis. Once reliable, automated, self-structuring ANN packages will become available, the classification analyst may no longer need to build his/her own ANN models. As soon as such products become available at a reasonable price, which may happen in the near future, the impediments to a wide-spread use of ANN methods for classification in practice are bound to disappear.

Combining Artificial Neural Networks with Other Methods

In addition to the combination of ANNs and LP, several interesting hybrid ANN methods have been proposed. Particularly intriguing are approaches that to seek to combine the strengths of statistical methods and ANNs, based on the premise that certain kinds of statistical information, such as an observation's distance from the group centroids, might provide useful input into a FFANN for classification. Markham and Ragsdale (1994) propose a FFANN method in which, in addition to the original attributes, the Mahalanobis distances from the group centroids serve as inputs into the FFANN. This method is similar in concept to that forwarded by Lam, Choo and Moy (1996), who however do not use ANNs. Markham and Ragsdale (1994) note that the predictions from the

Mahalanobis distance method are equivalent to the LDF, and report that their combined method yields more accurate composite predictions on two real data sets than either of the individual methods. Wang (1996) first pre-processes the training sample data using linear discriminant analysis, and then uses a combination of self-organizing feature maps to detect clusters of misclassifications.

5. Trends: What Might be in Store for the L_p-Norm Classification Field?

Undoubtedly, the literature on L_p-norm classification is interesting and forms a worthwhile contribution to the field of classification. Given their geometric rather than decision-theoretic foundation, it is not surprising that to date many different L_p-norm classification methods have been proposed. Some of these have proven to perform better than others, but none dominates across the board for all data conditions, and much work is to be done to establish the L_p-norm methods *vis à vis* competing methods. As the research in this area continues to mature, the field is approaching an important crossroad, well worth reflecting about. It may well be that the direction of research in this area over the next five to ten years will be pivotal, in terms of whether or not in the long run L_p-norm methods will be used by practitioners and will have an impact on the field of statistical classification. In this section, a number of promising directions for future research are identified, and issues are discussed with are of vital importance to the long-term outlook of the field of L_p-norm classification.

5.1. Why Have Statisticians Rarely Used L_p-Norm Methods?

In order to identify promising research directions, it is necessary to examine in detail why statisticians have largely ignored L_p-norm classification, and which lessons can be learned from this. The research conducted on L_p-norm classification has had little impact in statistical circles, and many statisticians do not seem to be familiar with the L_p-norm line of research. For instance, although many of the over 1,000 papers cited in McLachlan's (1992) seminal text on statistical discriminant analysis deal with nonparametric classification methods, only two of them are papers on L_p-norm classification and discriminant analysis. In his book on applied discriminant analysis, Huberty (1994) cites only a handful of L_p-norm papers. Of course, most L_p-norm classification articles have been published within the last ten to fifteen years, and the time lag effect may play a role. However, this can provide only a partial explanation. Other reasons why statisticians have not used L_p-norm methods for classification – some of which the author learned about in personal discussions with statisticians – are introduced below, along with suggestions on how each issue might be remedied.

Communication, Promotion and Terminology

First, researchers in L_p-norm classification may not have promoted their work effectively to the statistical community. For instance, few L_p-norm classification papers have been presented at major

statistical meetings, such as the Annual Joint Statistical Meetings organized by the ASA, ENAR, WNAR, IMS and SSC, and the Annual Meeting of the Classification Societies. Of course, the reverse is true as well – few statisticians have presented their classification work at DSI and INFORMS meetings. Second, most L_p-norm classification papers are packaged in a way familiar to readers in the management science community, but the terminology used is not familiar to most statisticians, thus representing a communication gap which inhibits the adoption of L_p-norm methods by statisticians, especially practitioners. By necessity, much of the early L_p-norm classification research focused on issues related to refining mathematical modeling aspects of the MP methodology, rather than on statistical aspects. This may have contributed to the low exposure of L_p-norm methods to statisticians as well.

Communication can be improved by adopting terminology in L_p-norm classification papers which corresponds more closely with that used in mainstream statistical circles, and avoiding unnecessary MP-related details (an exception, of course, being those papers which relate directly with MP-algorithmic issues), making L_p-norm classification research more accessible to statistical researchers and practitioners. In addition, L_p-norm classification methods can be promoted in various ways, *e.g.*, by submitting solid research articles to leading statistical journals, presenting research findings at professional meetings of statistical organizations, and making a serious attempt to address those issues which statisticians perceive as weaknesses of L_p-norm methods. Some of these perceived problems with L_p-norm methods are discussed next.

Software Availability

There is a real need for easily accessible L_p-norm classification software packages, both user-friendly stand-alone packages and add-on software that can be used in conjunction with mainline statistical packages such as SAS (1990), SPSS (1990) and BMDP (1990). Most researchers in L_p-norm classification use their own software programs. To the knowledge of the author, a handful of software packages are available to interested analysts (Lam and Choo 1991; Duarte Silva and Stam 1994a; Stam and Ungar 1995; Soltysik and Yarnold 1993, 1994). Among these, only the ODA software by Soltysik and Yarnold (1993) is available commercially – the others can be obtained upon request from the respective authors. Among these packages, BestClass (Duarte Silva and Stam 1994a) is the only one that can be used as an add-on to a major statistical package, SAS (1990). There is no doubt that a wider availability and more commercial-quality software products for L_p-norm classification will stimulate the use of these methods, by researchers and practitioners alike. It would very helpful to develop a central repository of L_p-norm classification analysis software, perhaps on the World Wide Web, with easy access for any analyst.

Relative Accuracy of L_p-Norm Classification Methods: Ad Hoc Studies

As noted above, most simulation studies have compared the accuracy of L_p-norm methods with the LDF, QDF and logistic regression, but not (or not enough) with other nonparametric methods, such as nearest neighbor, kernel, classification tree and recursive partitioning methods. This is unfortunate, since literally hundreds of studies in the statistical literature have – not surprisingly – found nonparametric methods to outperform the LDF, QDF and logistic regression for data conditions similar to those in the studies involving MP-based methods. For a more detailed review of the findings in these studies, see, *e.g.*, Dillon and Goldstein (1978), Fatti *et al.* (1982), Goldstein and Dillon (1978), Hand (1982, 1993), Huberty (1994), Krzanowski (1988), McLachlan (1992) and Press and Wilson (1978).

Thus, from a statistical viewpoint evaluations of L_p-norm methods that are limited to a comparison with classical statistical methods are less than interesting. One may view the early studies which evaluated L_p-norm methods *vis à vis* the LDF and QDF as preliminary scouting work, in order to establish that L_p-norm methods are at least viable. However, in the long run such research is bound to have a decreasing marginal impact. Since it appears that the viability of L_p-norm methods has indeed been established, it is now much more interesting to turn the attention to assessing how L_p-norm methods compare with the most successful nonparametric statistical methods. For example, little is known on how the different nonparametric methods compare for skewed distributions, various different misclassification cost schemes, different training sample sizes, and various numbers of attributes. It is known that nonparametric statistical methods, such as kernel and nearest neighbor methods, perform well if the distributions are skewed and the training samples are large, but not as well – *i.e.*, not necessarily better than the LDF and QDF – if the training samples are small (Remme, Habbema and Hermans 1980). Therefore, it is interesting to compare the performance of these methods with L_p-norm classification methods for these data conditions. It is imperative that these studies lay to rest the concerns that statisticians have regarding the relative standing of L_p-norm methods.

It is worthwhile to mention some other issues that have not received much attention in L_p-norm classification research, nor in many of the classification studies that focus on statistical methods. Almost all L_p-norm studies use the proportion (rate) of misclassified (or correctly classified) observations in the training or validation sample as the only measure of classification accuracy. Of course, the misclassification rate is the most widely used accuracy measure, but it is not always the most telling one. For instance, if the misclassification costs differ across groups, a simple error count does not provide an accurate measure of classification performance. The misclassification rate is a measure of overall accuracy. It would also be of interest to conduct studies involving micro-level measures, for instance investigating whether or not different methods tend to misclassify the same

observations, in order to develop a better understanding of the reasons why L_p-norm methods yield (or do not yield) improved classification rules for specific data conditions.

One way to enhance simulation studies for classification is to use data conditions for which the "true" optimal classification rule (*e.g.*, the Bayes rule) is known, the advantage being that the accuracy of the optimal rule can serve as a benchmark for measuring the absolute classification performance of each method under consideration, providing valuable information in addition to (or, instead of) the usual relative classification accuracy measures. The Bayes optimal classification rule can be derived for a number of different probability distributions (Duarte Silva 1995).

Summarizing, the goal of convincing statisticians that the L_p-norm class of methods has definite merits can be achieved by showing the classification accuracy of L_p-norm methods using rigorous, well-designed experimental studies. While still *ad hoc*, studies with carefully selected data conditions and factors (*e.g.*, skewness, extreme observations, number of attributes, group overlap, prior probabilities as reflected in the balance of the group sizes, misclassification costs, training sample size) in the experimental design that closely reflect reality, a legitimate statistical analysis of the results (*e.g.*, using MANOVA and perhaps T-tests), a sufficiently large number of replications, appropriate measures of classification accuracy, and a competitive set of alternative methods used in the comparison – especially nonparametric rivals – will provide excellent insights. Such studies will lend additional credibility to the class of L_p-norm classification methods and will capture the attention of those statisticians who are active in nonparametric classification.

Accuracy of L_p-Norm Classification Methods: Decision Theoretic Justification

One needs to be careful in generalizing results obtained in simulation studies. Of course, an advantage of simulation studies is that the distributional characteristics of the populations in the experimental design can be controlled exactly. However, by the nature of the process by which the data were generated, simulation results can be unrealistically "clean," and a generalization to classification problems with real data may be tenuous. For example, many simulation studies have assumed independent attribute variables, a condition which is rarely met in reality.

Ideally, studies evaluating the classification accuracy of a given method would be supplemented by a decision-theoretic justification for using this method. The advantage of exploring decision-theoretic properties is that, once such properties have been established, extensive simulations to show the relevance of the method are no longer needed – the decision-theoretic foundation provides this justification. Of course, a decision-theoretic justification does not guarantee an accurate classification rule, but it does improve the odds that the estimated rule is a good one; besides, it is always preferred to use a method that has a proven theoretical foundation.

It is unrealistic to expect that a general decision-theoretic foundation exists for the class of L_p-

norm methods, but certain L_p-norm methods do have one for specific data conditions that are realistic in practice. For instance, Asparoukhov and Stam (1996) derive an L_0-norm method for binary variable classification problems which yields an optimal Bayesian rule. Binary variable classification has many applications, *a.o.*, in medical disease diagnosis, where a given symptom is either present or absent. It is likely that decision-theoretic L_p-norm methods exist for other specific types of classification problems as well, for instance in the case of mixed and discrete variable problems. An important development in the direction of a formal foundation for L_0-norm discrimination is the research by Soltysik and Yarnold (1993) and Yarnold and Soltysik (1991a, 1991b), who derive several fundamental properties of the UniODA and MultiODA methods.

As mentioned previously, the L_0-norm method is a promising classification tool for yet another important reason, in that it can be used to minimize the training sample misclassification costs *directly*, if the objective function components are weighted appropriately as in (6), according to their prior probabilities and misclassification costs. As statistical methods for minimizing the misclassification costs are not fully satisfactory if the group-wise attribute probability distributions are difficult to establish, this is a real contribution to the field of statistics. Due to the importance of methods that minimize misclassification costs, it appears useful to continue to study all aspects of the L_0-norm method, and to evaluate its accuracy for various data conditions and various levels of misclassification costs for each group. Note that it is much more complicated to reflect misclassification costs in the case of weighted variants of L_p-norm ($p > 0$) and Hybrid methods. Moreover, the MIP appears to have interesting asymptotic properties (Glick 1976), which may be warrant further study, *e.g.*, in the context of misclassification costs.

In sum, as the proponents of these methods it is the responsibility of researchers in L_p-norm classification to expose statisticians to and convince them of the merits of these methods. Clearly, a stronger link with statisticians and exchange of views with statisticians are of vital importance in terms of the long-term prospects of research in L_p-norm classification, as are rigorous comparative studies.

5.2. Other Important Research Topics

Computational Aspects

The need for more efficient methods to solve the MIP problem is obvious. Any serious methodology should allow for the analysis, within reasonable computational time, of several thousand training sample observations, and at least 10 attributes. From previous research it appears that, besides the size of the training sample, the number of attributes affects the computational performance of L_0-norm methods critically, and that the most promising improvements of exact special-purpose algorithms for the L_0-norm method may be found by decomposing the problem formulation into

smaller sub-problems that can be solved more easily. In turn, efficient problem decompositions offer an open invitation for the development of parallel processing techniques. Other ways to take advantage of the special structure of the L_0-norm classification problem should be explored in further detail as well. In addition to exact algorithms for the L_0-norm method, it is of interest to develop efficient (and effective) special-purpose heuristics, such as genetic algorithms (Koehler 1991b) and Tabu Search algorithms, for solving the L_0-norm method. Special-purpose algorithms for L_p-norm methods, $p > 0$, do not appear to be as worthwhile, since powerful software exists for solving these methods.

Generality of Scope: Multi-Group Methods and Nonlinear Transformations of Attributes

As discussed above, most multi-group L_p-norm classification methods have serious drawbacks, limiting either their applicability in practice or their generalizability. Gehrlein's extension for the L_0-norm method with intersecting hyperplanes is methodologically rigorous and generalizable, but requires many binary variables and is difficult computationally, even for small training samples. The extension by Gochet *et al.* is also general, conceptually simple and methodologically rigorous, but fairly involved in terms of implementation. Multi-group methods based on parallel separating surfaces may provide accurate classification results on a given data set and for certain data conditions, but are not sufficiently general. While the multi-group methods have greatly expanded the scope of L_p-norm methods, using them in practice requires software that is not readily available or easily accessible. To date, few studies have evaluated the classification performance of these methods. Clearly, more research is needed to evaluate multi-group methods, but this research should not restrict itself to methods based on a single function with parallel hyperplanes.

Another topic of interest is to study under which conditions L_p-norm formulations with nonlinear transformations of the attributes yield good classification results. Early L_p-norm studies ignored such nonlinear transformations, although the data conditions analyzed would seem to justify the use of nonlinear rules. In these studies, the deck was stacked against the methods with linear rules, and the question arises whether some of the earlier studies that considered only L_p-norm rules which were linear in the original attribute values should be re-done, in particular if the data conditions clearly called for a classification rule which was nonlinear in the attributes.

The evaluation of nonlinear L_p-norm rules is particularly interesting because it is not a given that a nonlinear rule classifies more accurately than a linear one, even if the theoretically optimal rule is nonlinear. As noted above, for small training samples the QDF tends to be less accurate than the LDF, even if the populations are normally distributed and the group-wise dispersions are clearly different. It is an open question whether L_p-norm formulations which involve quadratic (and perhaps cross-product) terms are as sensitive to sample size as the QDF. It would be a contribution to the field of L_p-norm classification to establish guidelines about appropriate training sample sizes for the use of

nonlinear rules, and to develop an analogon for L_p-norm methods to the concept of degrees of freedom in statistical analysis. The related phenomenon of overfitting has not been studied thoroughly in the context of L_p-norm classification analysis either, and warrants additional research, not only in the context of nonlinear classification rules, but also as it relates to multi-group classification. In both of these cases, the number of parameters to be estimated can be substantial.

L_p-Norm Methods for Descriptive Discriminant Analysis

The focus in this paper – as in the field at large – has been on prescriptive issues involving the classification problem in discriminant analysis. Perhaps it is time to expand the horizon to include issues in descriptive discriminant analysis, such as variable (attribute) selection (Glorfeld and Olson 1982; Nath and Jones 1988). As it stands now, almost all research in L_p-norm classification and discrimination has taken the number of attributes as given. A reasonable L_p-norm variable selection methodology which complements (but does not replace, of course) traditional descriptive discriminant methods would add another dimension to L_p-norm discrimination, and would develop this area more fully.

Combination with other Methodologies

As noted above, Roy and Mukhopadhyay (1990) and Mangasarian (1993) offer interesting links between L_p-norm classification methods and Artificial Intelligence, showing how MP techniques can be used for machine learning, and how MP and ANN can be combined into powerful classifiers. The arguments forwarded by Mukhopadhyay *et al.* (1993) and Roy *et al.* (1995), who argue in favor of an approach which retains training patterns in memory and accommodates a flexible network design that can be adapted during the training process, are intriguing. Their approach, which involves both LP and ANN-like networks, appears very promising. Although they have been illustrated by individual examples, the classification accuracy of these methods has not been put to sufficient testing through comprehensive, systematic statistical comparisons with competing methods. As noted previously, a rigorous statistical evaluation of ANN methods involving numerous replications may be cumbersome, but it is definitely a worthwhile effort. It also appears useful to explore other hybrid methods that combine ANNs with, *e.g.*, with statistical methods and L_p-norm methods. Spiegelhalter and Knill-Jones (1984) combine statistical methods with Expert Systems models for classification in the medical field. The use of statistical classification methods within an Expert Systems framework is also used in other areas, *e.g.*, in the field of finance. As these models combine statistical evidence with expert knowledge that may not be easy to analyze quantitatively, classification models which combine Expert Systems with statistical, L_p-norm, ANN methods and classification trees, appear fertile ground for future research.

6. Conclusions

This paper highlights previous research in L_p-norm classification, and suggests directions for future research. Above all, it is argued that there is a need to forge a link between researchers active in statistical discriminant analysis and researchers in the area of L_p-norm classification. Such a link would be beneficial for both groups. Particularly, L_p-norm classification may well be of considerable interest to researchers in areas where nonparametric classification analysis is traditionally used successfully, such as discrete variable classification, mixed variable classification, and in application areas which are often susceptible to data analytical problems, such as medical diagnosis, psychology, marketing, financial analysis, engineering and pattern recognition. Without reaching out, the L_p-norm classification field will remain limited to a small group of researchers with interesting new methodologies that are hardly used where they may be most needed.

In order to improve the channels of communication, closer ties with the Society of Classification and similar organizations would be helpful, as would be the availability of and easy access to software for L_p-norm classification, the development and evaluation of more general methods for L_p-norm classification that can handle nonlinear classification rules and multiple groups, and the development of faster algorithms for solving the L_0-norm method, which is attractive in that it allows for directly minimizing the training sample misclassification costs. In addition, there is a need for rigorous simulation experiments that should establish beyond any doubt for which general data conditions L_p-norm methods perform well, and which are not limited to comparisons with well-known but not always robust statistical methods such as the LDF and QDF, but which take on the best performing statistical nonparametric methods directly.

Summarizing, the area of L_p-norm classification appears have great potential, but the future of this research field depends on the ability to catch the attention of the statistically oriented research community. This enhances the international standing of this research area. Without an effort in the direction of more statistically oriented and motivated papers and rigorous studies that prove beyond a doubt when the L_p-norm methods are most appropriate, the general interest in this area may ebb away, perhaps to re-emerge ten or twenty years down the road in a different form. Such a scenario is not far-fetched, it is a simple and plain fact that this is part of many research area's life cycles. An encouraging sign is that, whereas most of the earlier papers originated in North America, recently there has been an increasing flow of publications from Europe and the Pacific Basin.

Footnote 1: I am grateful to Dr. Cliff T. Ragsdale for bringing this passage to my attention.

References

Abad, P. L. and Banks, W. J., "New LP Based Heuristics for the Classification Problem," *European Journal of Operational Research*, **67**, 1993, 88–100.

Archer, N. P. and Wang, S., "Application of the Back Propagation Neural Network Algorithm With Monotonicity Constraints for Two-Group Classification Problems," *Computers & Operations Research*, **24**, 1993, 60–75.

Asparoukhov, O. K., *Microprocessor System for Investigation of Thromboembolic Complications*, Unpublished Ph.D. Dissertation, Technical University of Sofia, Bulgaria (in Bulgarian), 1985.

Asparoukhov, O. K. and Stam, A., "Mathematical Programming Formulations for Two-Group Classification With Binary Variables," Working Paper 96–92, International Institute for Applied Systems Analysis, Laxenburg, Austria, 1996.

Bajgier, S. M. and Hill, A., "An Experimental Comparison of Statistical and Linear Programming Approaches to the Discriminant Problem," *Decision Sciences*, **13**, 1982, 604–618.

Banks, W. J. and Abad, P. L., "An Efficient Optimal Solution Algorithm for the Classification Problem," *Decision Sciences*, **22**, 1991, 1008–1023.

Banks, W. J. and Abad, P. L., "On the Performance of Linear Programming Heuristics Applied on a Quadratic Transformation in the Classification Problem," *European Journal of Operational Research*, **74**, 1994, 23–28.

Barrodale, I. and Roberts, F. D. K., "Applications of Mathematical Programming to L_p Approximations," in *Nonlinear Programming*, Rosen, J. B., Mangasarian, O. L. and Ritter, K. (Eds.), Academic Press, New York, NY, 1970, 447–464.

Barrodale, I. and Roberts, F. D. K., "An Improved Algorithm for Discrete L_1 Approximation," *SIAM Journal on Numerical Analysis*," **10**, 1973, 839–848.

Bassett, G., Jr. and Koenker, R., "The Asymptotic Distribution of the Least Absolute Error Estimator," *Journal of the American Statistical Association*, **73**, 1978, 618–622.

Benediktsson, J. A., Swain, P. H. and Ersory, O. K., "Neural Network Approaches Versus Statistical Methods in Classification of Multisource Remote Sensing Data," *IEEE Transactions on Geoscience and Remote Sensing*, **28**, 1990, 540–551.

Bennett, K. P. and Mangasarian, O. L., "Neural Network Training via Linear Programming," in *Advances in Optimization and Parallel Computing*, Pardalos, P. M. (Ed.), North Holland, Amsterdam, 1992a, 56–67.

Bennett, K. P. and Mangasarian, O. L., "Robust Linear Programming Discrimination of Two Linearly Separable Sets," *Optimization Methods and Software*, **1**, 1992b, 23–34.

BMDP, *BMDP Statistical Software Release 7, Volumes 1, 2*, 1990.

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

Choo, E. U. and Wedley, W. C., "Optimal Criterion Weights in Repetitive Multicriteria Decision-Making," *Journal of the Operational Research Society*, **36**, 1985, 983–992.

Cybenko, G., "Approximations by Superpositions of a Sigmoidal Function," *Mathematics of Control, Signals, and Systems*, **2**, 1989, 303–314.

Dillon, W. R. and Goldstein, M., "On the Performance of Some Multinomial Classification Rules," *Journal of the American Statistical Association*, **73**, 1978, 305–313.

Dodge, Y. (Ed.), *Statistical Data Analysis Based on the L_1-Norm and Related Methods*, North-Holland, Amsterdam, 1987.

Duarte Silva, A. P., *Minimizing Misclassification Costs in Two-Group Classification Analysis*, Unpublished Ph.D. Dissertation, The University of Georgia, 1995.

Duarte Silva, A. P. and Stam, A., "BestClass: A SAS-Based Software Package of Nonparametric Methods for Two-Group Classification," Working Paper 94–396, Terry College of Business, The University of Georgia, Athens, GA, 1994a.

Duarte Silva, A. P. and Stam, A., "Second Order Mathematical Programming Formulations for Discriminant Analysis," *European Journal of Operational Research*, **74**, 1994b, 4–22.

Duarte Silva, A. P. and Stam, A., "An Efficient Mixed Integer Programming Algorithm for Minimizing the Training Sample Misclassification Cost in Two-Group Classification," Working Paper 96–93, International Institute for Applied Systems Analysis, Laxenburg, Austria, 1996.

Erenguc, S. S. and Koehler, G. J., "Survey of Mathematical Programming Models and Experimental Results for Linear Discriminant Analysis," *Managerial and Decision Economics*, **11**, 1990, 215–226.

Fatti, L. P., Hawkins, D. M. and Raath, E. L., "Discriminant Analysis," in *Topics in Applied Multivariate Analysis*, D. M. Hawkins (Ed.), Cambridge University Press, Cambridge, England, 1982, 1–71.

Fischer, M., Gopal, S., Staufer, P. and Steinocher, K., "Evaluation of Neural Pattern Classifiers for a Remote Sensing Application," Working Paper, Department of Geography, Wirtschaftsuniversität Wien, Vienna, Austria, 1994.

Fisher, R. A., "The Use of Multiple Measurements in Taxonomy Problems," *Annals of Eugenics*, **7**, 1936, 179–188.

Freed, N. and Glover, F., "A Linear Programming Approach to the Discriminant Problem," *Decision Sciences*, **12**, 1981a, 68–74.

Freed, N. and Glover, F., "Simple But Powerful Goal Programming Formulations for the Discriminant Problem," *European Journal of Operational Research*, **7**, 1981b, 44–60.

Freed, N. and Glover, F., "Evaluating Alternative Linear Programming Models to Solve the Two-Group Discriminant Problem," *Decision Sciences*, **17**, 1986a, 151–162.

Freed, N. and Glover, F., "Resolving Certain Difficulties and Improving the Classification Power of LP Discriminant Analysis Formulations," *Decision Sciences*, **17**, 1986b, 589–595.

Gehrlein, W. V., "General Mathematical Programming Formulations for the Statistical Classification Problem," *Operations Research Letters*, **5**, 1986, 299–304.

Glahe, F. R. and Hunt, J. G., "The Small Sample Properties of Simultaneous Equation Least Absolute Estimators Vis-a-Vis Least Squares Estimators," *Econometrica*, **38**, 1970, 742–753.

Glick, N., "Sample-Based Classification Procedures Related to Empiric Distributions," *IEEE Transactions on Information Theory*, **22**, 1976, 454–461.

Glorfeld, L. W., "A Robust Methodology for Discriminant Analysis Based on Least-Absolute-Value Estimation," *Managerial and Decision Economics*, **11**, 1990, 267–277.

Glorfeld, L. W. and Olson. D. L., "Variable Selection for Robust Discriminant Analysis Based on the L_1 Metric," *Proceedings of the American Institute of the Decision Sciences*, 1982, 745–747.

Glover, F., "Improved Linear Programming Models for Discriminant Analysis," *Decision Sciences*, **21**, 1990, 771–785.

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

Gochet, W., Stam, A., Srinivasan, V. and Chen, S., "Multi-Group Discriminant Analysis Using Linear Programming," *Operations Research*, Forthcoming, 1996.

Goldstein, M. and Dillon, W. R., *Discrete Discriminant Analysis*, Wiley, New York, NY, 1978.

Gonin, R. and Money, A. H. (Eds.), *Nonlinear L_p-norm Estimation*, Marcel Dekker, 1989.

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

Hand, D. J., *Kernel Discriminant Analysis*, Wiley, New York, NY, 1982.

Hand, D. J., "Discriminant Analysis for Categorical Data," in *Lecture Notes and Program of the 4th European Courses in Advanced Statistics Program: Analysis of Categorical Data Theory and Application*, Leiden, The Netherlands, 1993, 135–174.

Hertz, J., Krogh, A. and Palmer, R. G., *Introduction to the Theory of Neural Computation*, Addison-Wesley, Redwood City, CA, 1991.

Hillier, F. S. and Lieberman, G. J., *Introduction to Operations Research, Fifth Edition*, McGraw-Hill, New York, 1990.

Hosseini, J. H. and Armacost, R. L., "The Two-Group Discriminant Problem With Equal Group Mean Vectors: An Experimental Evaluation of Six Linear/Nonlinear Programming Formulations," *European Journal of Operational Research*, **77**, 1994, 241–252.

Huberty, C. J., "Issues in the Use and Interpretation of Discriminant Analysis," *Psychological Bulletin*, **95**, 1984, 156–171.

Huberty, C. J., *Applied Discriminant Analysis*, Wiley, New York, NY, 1994.

Ibaraki, T. and Muroga, S., "Adaptive Linear Classifier by Linear Programming," *IEEE Transactions on Systems, Science and Cybernetics*, **SSC-6**, 1970, 53–62.

Ignizio, J. P., *Linear Programming in Single and Multiple Objective Systems*, Prentice Hall, Englewood Cliffs, NJ, 1982.

Jain, B. A. and Nag, B. N., "Artificial Neural Network Models for Pricing Initial Public Offerings," *Decision Sciences*, **26**, 1995, 283–302.

Joachimsthaler, E. A. and Stam, A., "Four Approaches to the Classification Problem in Discriminant Analysis: An Experimental Study," *Decision Sciences*, **19**, 1988, 322–333.

Joachimsthaler, E. A. and Stam, A., "Mathematical Programming Approaches for the Classification Problem in Two-Group Discriminant Analysis," *Multivariate Behavioral Research*, **25**, 1990, 427–454.

Kattan, M. W. and Beck, J. R., "Artificial Neural Networks for Medical Classification Decisions," *Archives of Pathology and Laboratory Medicine*, **119**, 1995, 672–677.

Koehler, G. J., "Characterization of Unacceptable Solutions in LP Discriminant Analysis," *Decision Sciences*, **20**, 1989a, 239–257.

Koehler, G. J., "Unacceptable Solutions and the Hybrid Discriminant Model," *Decision Sciences*, **20**, 1989b, 844–848.

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

Koehler, G. J., "Improper Linear Discriminant Classifiers," *European Journal of Operational Research*, **50**, 1991a, 188–198.

Koehler, G. J., "Linear Discriminant Functions Determined by Genetic Search," *ORSA Journal on Computing*, **3**, 1991b, 345–357.

Koehler, G. J., "A Response to Xiao's 'Necessary and Sufficient Conditions of Unacceptable Solutions in LP Discriminant Analysis:' Something is Amiss," *Decision Sciences*, **25**, 1994, 331–333.

Koehler, G. J. and Erenguc, S. S., "Minimizing Misclassifications in Linear Discriminant Analysis," *Decision Sciences*, **21**, 1990, 63–85.

Koenker, R. and Portnoy, S., *L-Estimation for Linear Models," Journal of the American Statistical Association*, **82**, 1987, 851–857.

Koford, J. S. and Groner, G. F., "The Use of an Adaptive Threshold Element to Design a Linear Optimal Pattern Classifier," *IEEE Transactions on Information Theory*, **IT-12**, 1966, 42–50.

Krzanowski, W. J., *Principles of Multivariate Analysis*, Clarendon Press, Oxford, England, 1988.

Lachenbruch, P. A., Sneeringer, C. and Revo, L. T., "Robustness of the Linear and Quadratic Discriminant Function to Certain Types of Non-Normality," *Communications in Statistics*, **1**, 1973, 39–57.

Lam, K. F. and Choo, E. U., "Software Package for Linear Programming in Classification Problems," Faculty of Business Administration, Simon Fraser University, Burnaby, BC, Canada, 1991.

Lam, K. F. and Choo, E. U., "A Linear Goal Programming Model for Classification With Non-Monotonic Attributes," *Computers & Operations Research*, **20**, 1993, 403–408.

Lam, K. F., Choo, E. U. and Wedley, W. C., "Linear Goal Programming in Estimation of Classification Probability," *European Journal of Operational Research*, **67**, 1993, 101–110.

Lam, K. F., Choo, E. U. and Moy, J. W., "Minimizing Deviations From Group Mean: A New Linear Programming Approach for the Classification Problem," *European Journal of Operational Research*, **88**, 1996, 358–367.

Lau, K.-N. and Post, G. V., "A Note on Discriminant Analysis Using LAD," *Decision Sciences*, **23**, 1992, 260–265.

Lawrence, K. D. and Arthur, J. L. (Eds.), *Robust Regression: Analysis and Application*, Marcel Dekker, New York, NY, 1990.

Lee, C. K. and Ord, J. K., "Discriminant Analysis Using Least Absolute Deviations," *Decision Sciences*, **21**, 1990, 86–176.

Liitschwager, J. M. and Wang, C., "Integer Programming Solution of a Classification Problem," *Management Science*, **24**, 1978, 1515–1525.

Lippmann, R. P., "Pattern Classification Using Neural Networks," *IEEE Communication Magazine*, 1989, **27**, 47–64.

Mahmood, M. A. and Lawrence, E. C., "A Performance Analysis of Parametric and Nonparametric Discriminant Approaches to Business Decision Making," *Decision Sciences*, **18**, 1987, 308–326.

Mangasarian, O. L., "Linear and Nonlinear Separation of Patterns by Linear Programming," *Operations Research*, **13**, 1965, 444–452.

Mangasarian, O. L., "Multi-Surface Method of Pattern Separation," *IEEE Transactions on Information Theory*, **IT-14**, 1968, 801–807.

Mangasarian, O. L., "Mathematical Programming in Neural Networks," *ORSA Journal on Computing*, **5**, 1993, 349–360.

Marcotte, P., Marquis, G. and Savard, G., "A New Implicit Enumeration Scheme for the Discriminant Analysis Problem," *Computers & Operations Research*, **22**, 1995, 625–639.

Marcotte, P. and Savard, G., "Novel Approaches to the Discriminant Problem," *Zeitschrift für Operations Research*, **36**, 1991, 517–545.

Markham, I. S. and Ragsdale, C. **T.**, "Combining Neural Networks and Statistical Predictions to Solve the Classification Problem in Discriminant Analysis," *Decision Sciences*, **26**, 1995, 229–242.

Markowski, C. A., "On the Balancing of Error Rates for LP Discriminant Methods," *Managerial and Decision Economics*, **11**, 1990, 235–241.

Markowski, C. A. and Markowski, E. P., "An Experimental Comparison of the Discriminant Problem With Both Qualitative and Quantitative Variables," *European Journal of Operational Research*, **28**, 1987, 74–78.

Markowski, C. A., "An Adaptive Statistical Method for the Discriminant Problem," *European Journal of Operational Research*, **73**, 1994, 480–486.

Markowski, E. P. and Markowski, C. A., "Some Difficulties and Improvements in Applying Linear Programming Formulations to the Discriminant Problem," *Decision Sciences*, **16**, 1985, 237–247.

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

Minnick, R. C., "Linear-Input Logic," *IEEE Transactions on Electronics and Computers*, **EC-10**, 1961, 6–16.

Morrison, D. F., *Multivariate Statistical Methods*, Third Edition, McGraw-Hill, New York, NY, 1990.

Mukhopadhyay, S., Roy, A., Kim, L. S. and Govil, S., "A Polynomial Time Algorthm for Generating Neural Networks for Pattern Classification – Its Stability Properties and Some Test Results," *Neural Computing*, **5**, 1993, 225–238.

Narula, S. C. and Wellington, J. F., "The Minimum Sum of Absolute Error Regressions: A State of the Art Review," *International Statistical Review*, **50**, 1982, 317–326.

Narula, S. C. and Wellington, J. F., "On the Robustness of the Simple Linear Minimum Sum of Absolute Errors Regression," in Lawrence, K. D. and Arthur, J. L. (Eds.), *Robust Regression: Analysis and Application*, Marcel Dekker, New York, NY, 1990, 129–141.

Nath, R. R., "Estimation of Misclassification Probabilities in the Linear Programming Approaches to the Two-Group Discriminant Problem," *Decision Sciences*, **15**, 1984, 248–252.

Nath, R., Jackson, W. M. and Jones, T. W., "A Comparison of the Classical and the Linear Programming Approaches to the Classification Problem in Discriminant Analysis," *Journal of Statistical Computation and Simulation*, **41**, 1992, 73–93.

Nath, R. R. and Jones, T. W., "A Variable Selection Criterion in the Linear Programming Approaches to Discriminant Analysis," *Decision Sciences*, **19**, 1988, 554–563.

Patuwo, E., Hu, M. Y. and Hung, M. S., "Two-Group Classification Using Neural Networks," *Decision Sciences*, **24**, 1993, 825–845.

Pavur, R. and Loucopoulos, C., "Examining Optimal Criterion Weights in Mixed Integer Programming Approaches to the Multiple-Group Classification Problem," *Journal of the Operational Research Society*, **46**, 1995, 626–640.

Press, S. J. and Wilson, S., "Choosing Between Logistic Regression and Discriminant Analysis," *Journal of the American Statistical Association*, **73**, 1978, 699–705.

Rawlings, R. R., Faden, V. B., Graubhard, B. I. and Eckardt, M. J., "A Study of Discriminant Analysis Techniques Applied to Multivariate Lognormal Data," *Journal of Statistical Computing and Simulation*, **26**, 1986, 79–100.

Ragsdale, C. T. and Stam, A., "Mathematical Programming Formulations for the Discriminant Problem: An Old Dog Does New Tricks," *Decision Sciences*, **22**, 1991, 296–307.

Remme, J., Habbema, J. D. F. and Hermans, J., "A Simulative Comparison of Linear, Quadratic and Kernel Discrimination," *Journal of Statistical Computing and Simulation*, **11**, 1980, 87–106.

Roy, A., Govil, S. and Miranda, R., "An Algorithm to Generate Radial Basis Function (RBF)-Like Nets for Classification Problems," *Neural Networks*, **8**, 1995, 179–202.

Roy, A. and Mukhopadhyay, S., "Pattern Classification Using Linear Programming," *ORSA Journal on Computing*, **3**, 1990, 66–80.

Rubin, P. A., "Separation Failure in Linear Programming Discriminant Models," *Decision Sciences*, **22**, 1989a, 519–535.

Rubin, P. A., "Evaluating the Maximize Minimum Distance Formulation of the Linear Discriminant Problem," *European Journal of Operational Research*, **41**, 1989b, 240–248.

Rubin, P. A., "Heuristic Solution Procedures for a Mixed-Integer Programming Discriminant Model," *Managerial and Decision Economics*, **11**, 1990a, 255–266.

Rubin, P. A., "A Comparison of Linear Programming and Parametric Approaches to the Two Group Discriminant Problem," *Decision Sciences*, **21**, 1990b, 373–386.

Rubin, P. A., "A Comment Regarding Polynomial Discriminant Analysis," *European Journal of Operational Research*, **72**, 1994, 29–31.

Rumelhart, D. E., Hinton, G. E. and Williams, R. J., "Learning Internal Representations by Error Propagation," in *Parallel Distributed Processing, Volume 1: Foundations*, Rumelhart, D. E., McClelland, J. L. and the PDP Group (Eds.), MIT Press, Cambridge, MA, 1986, 318–362.

Rumelhart, D. E., McClelland, J. L. and the PDP Group (Eds.), *Parallel Distributed Processing, Volume 1: Foundations*, MIT Press, Cambridge, MA, 1986.

Rypley, B., "Neural Networks and Related Methods for Classification," *Journal of the Royal Statistical Society*, **B**, **56**, 1994, 409–456.

SAS Institute, *SAS/STAT User's Guide, Version 6, First Edition*, SAS Institute, Inc., Cary, NC, 1990.

Smith, C. A. B., "Some Examples of Discrimination," *Annals of Eugenics*, **13**, 1947, 272–282.

Smith, F. W., "Pattern Classifier Design by Linear Programming," *IEEE Transactions on Computers*, **C-17**, 1968, 367–372.

Soltysik, R. C. and Yarnold, P. R., *ODA 1.0: Optimal Discriminant Analysis for DOS*, Optimal Data Analysis, Chicago, IL, 1993.

Soltysik, R. C. and Yarnold, P. R., "The Warmack-Gonzalez Algorithm for Linear Two-Category Multivariate Optimal Discriminant Analysis," *Computers & Operations Research*, **21**, 1994, 735–745.

Spiegelhalter, D. J. and Knill-Jones, R. P., "Statistical and Knowledge-Based Approaches to Clinical Decision-Support Systems, With an Application to Gastraenterology," *Journal of the Royal Statistical Society Series A*, **147**, Part 1, 35–77.

Sposito, V. A., "Some Properties of L_p-Estimators," in *Robust Regression: Analysis and Application*, Lawrence, K. D. and Arthur, J. L. (Eds.), Marcel Dekker, New York, NY, 1990, 23–58.

SPSS, *Statistical Package for the Social Sciences, SPSS 6.1 Base System, Users Guide, Parts 1 and 2*, Prentice Hall, Englewood Cliffs, NJ, 1990.

Srinivasan, V. and Kim, Y. H., "Credit Granting: A Comparative Analysis of Classification Procedures," *Journal of Finance*, **42**, 1987, 665–683.

Srinivasan, V. and Shocker, A., "Linear Programming Techniques for Multi-Dimensional Analysis of Preferences," *Psychometrica*, **38**, 1973, 337–369.

Stam, A., "Extensions of Mathematical Programming-Based Classification Rules: A Multicriteria Approach," *European Journal of Operational Research*, **48**, 1990, 351–361.

Stam, A. and Joachimsthaler, E. A., "Solving the Classification Problem in Discriminant Analysis via Linear and Nonlinear Programming Methods," *Decision Sciences*, **20**, 1989, 285–293.

Stam, A. and Joachimsthaler, E. A., "A Comparison of a Robust Mixed-Integer Approach to Existing Methods for Establishing Classification Rules for the Discriminant Problem," *European Journal of Operational Research*, **46**, 1990, 113–120.

Stam, A. and Jones, D. G., "Classification Performance of Mathematical Programming Techniques in Discriminant Analysis: Results for Small and Medium Sample Sizes," *Managerial and Decision Economics*, **11**, 243–253.

Stam, A. and Ragsdale, C. T., "On the Classification Gap in MP-Based Approaches to the Discriminant Problem," *Naval Research Logistics*, **39**, 1992, 545–559.

Stam, A. and Ungar, D. R., "RAGNU: A Microcomputer Package for Two-Group Mathematical Programming-Based Nonparametric Classification," *European Journal of Operational Research*, **86**, 1995, 374–388.

Subramanian, V., Hung, M. S. and Hu, M. Y., "An Experimental Evaluation of Neural Networks for Classification," *Computers & Operations Research*, **20**, 1993, 769–782.

Tam, K. Y. and Kiang, M. Y., "Managerial Applications of Neural Networks: The Case of Bank Failure Predictions," *Management Science*, **38**, 1992, 926–947.

Wanarat, P. and Pavur, R., "Examining the Effect of Second-Order Terms in Mathematical Programming Approaches to the Classification Problem," *European Journal of Operational Research*, Forthcoming, 1996.

Wang, S., "Self-Organizing Approach to Managerial Nonlinear Discriminant Analysis: A Hybrid Method of Linear Discriminant Analysis and Neural Networks," *INFORMS Journal of Computing*, **8**, 1996, 118–124.

Warmack, R. E. and Gonzalez, R. C., "An Algorithm for the Optimal Solution of Linear Inequalities and its Application to Pattern Recognition," *IEEE Transactions on Computers*, **C-22**, 1973, 1065–1075.

Wasserman, P. D., *Neural Computing, Theory and Practice*, Van Nostrand Reinhold, New York, NY, 1989.

Xiao, B., "Necessary and Sufficient Conditions of Unacceptable Solutions in LP Discriminant Analysis," *Decision Sciences*, **24**, 1993, 699–712.

Xiao, B., "Necessary and Sufficient Conditions for Unacceptable Solutions in NLP Discriminant Analysis," *European Journal of Operational Research*, **77**, 1994a, 404–412.

Xiao, B., "Decision Power and Solutions of LP Discriminant Models: Rejoinder," *Decision Sciences*, **25**, 1994b, 335–336.

Yarnold, P. R., "Discriminating Geriatric and Non-Geriatric Patients Using Functional Status Information: An Example of Classification Tree Analysis via UniODA," *Educational and Psychological Measurement*, **66**, 1996, 656–667.

Yarnold, P. R., Hart, L. A. and Soltysik, R. C., "Optimizing the Classification Performance of Logistic Regression and Fisher's Discriminant Analysis," *Educational and Psychological Measurement*, **54**, 1994, 73–85.

Yarnold, P. R. and Soltysik, R. C., "Theoretical Distributions of Optima for Univariate Discrimination of Random Data," *Decision Sciences*, **22**, 1991a, 739–752.

Yarnold, P. A. and Soltysik, R. C., "Refining Two-Group Multivariable Classification Models Using Univariate Optimal Discriminant Analysis," *Decision Sciences*, **22**, 1991b, 1158–1164.

Yarnold, P. R., Soltysik, R. C. and Martin, G. J., "Heart Rate Variability and Susceptibility for Sudden Cardiac Death: An Example of Multivariable Optimal Discriminant Analysis," *Statistics in Medicine*, **13**, 1994, 1015–1021.

Yoon, Y., Swales, G. and Margavio, T. M., "A Comparison of Discriminant Analysis Versus Artificial Neural Networks," *Journal of the Operational Research Society*, **44**, 1993, 51–60.