Working Paper

Nonparametric Two-Group Classification: Concepts and a SAS®-Based Software Package

*A. Pedro Duarte Silva**
*Antonie Stam***

WP-96-127
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

Nonparametric Two-Group Classification: Concepts and a SAS®-Based Software Package

*A. Pedro Duarte Silva**
*Antonie Stam***

WP-96-127
December 1996

*Universidade Católica Portuguesa, Faculdade de Ciencias Economicas e Empresariais, Centro Regional do Porto, Rua Diogo Botelho 1327, 4150 Porto, Portugal

**Department of Management, Terry College of Business, The University of Georgia, Athens, GA 30602, U.S.A.
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

Nonparametric Two-Group Classification: Concepts and a SAS®-Based Software Package

A. Pedro Duarte Silva[1,2]
Antonie Stam[1]

1: Management Department, Terry College of Business
The University of Georgia
Athens, GA 30602
U.S.A.

and

International Institute for Applied Systems Analysis
A-2361 Laxenburg, Austria

2: Curso de Administração e Gestão de Empresas
Universidade Católica Portuguesa
Centro Regional do Porto
Rua Diogo Botelho 1327
4100 Porto
Portugal

Nonparametric Two-Group Classification: Concepts and a SAS®-Based Software Package

ABSTRACT

In this paper, we introduce BestClass, a set of SAS macros, available in the mainframe and workstation environment, designed for solving two-group classification problems using a class of recently developed nonparametric classification methods. The criteria used to estimate the classification function are based on either minimizing a function of the absolute deviations from the surface which separates the groups, or directly minimizing a function of the number of misclassified entities in the training sample. The solution techniques used by BestClass to estimate the classification rule utilize the mathematical programming routines of the SAS/OR® software.

Recently, a number of research studies have reported that under certain data conditions this class of classification methods can provide more accurate classification results than existing methods, such as Fisher's linear discriminant function and logistic regression. However, these robust classification methods have not yet been implemented in the major statistical packages, and hence are beyond the reach of those statistical analysts who are unfamiliar with mathematical programming techniques.

We use a limited simulation experiment and an example to compare and contrast properties of the methods included in BestClass with existing parametric and nonparametric methods. We believe that BestClass contributes significantly to the field of nonparametric classification analysis, in that it provides the statistical community with convenient access to this recently developed class of methods. BestClass is available from the authors.

Keywords: Two-Group Classification Analysis, Computer Software for Statistical Analysis.

Nonparametric Two-Group Classification: Concepts and a SAS®-Based Software Package

1. Introduction

The classification problem in discriminant analysis (DA), which involves assigning (classifying) entities (observations) to exactly one of several well-defined mutually exclusive groups or classes, based on their characteristics on a set of relevant attributes, is important in almost any field of applied science. Many different approaches have been proposed for solving the classification problem in DA. Let the entities belonging to one of two mutually exclusive groups be described by the p-dimensional attribute vector $\mathbf{x}$, denote the attribute vector associated with entity i by $\mathbf{x}_i$, and indicate membership in group j by G_j.

The classical approach to classification is to first estimate the probability (density) functions $p(\mathbf{x}_i \mid G_j)$, and then derive the classification rule which minimizes either the probability of misclassification or the expected misclassification cost. Another approach is to estimate the posterior probabilities $p(G_j \mid \mathbf{x}_i)$ of group membership directly, and use a classification rule that weighs these probabilities by the appropriate misclassification costs. A third approach is to pre-specify a particular form of classification function, and then determine the parameter values of this function that optimize some accuracy criterion, *i.e.*, some measure of classification accuracy in the training sample.

The origins of the latter approach can be traced back to Fisher's Linear Discriminant Function (LDF), derived as the linear function of the attributes that maximizes the ratio between among-group squared distances and within-group variances (Fisher 1936). The choice of optimization criterion depends on the objectives of the analysis and the nature of the particular data set to be analyzed. As Fisher was more interested in maximizing group discrimination than in classification accuracy, he chose a criterion directly related to discrimination. Nevertheless, the LDF is also often used for classification purposes. In fact, Welch (1939) and Wald (1944) have shown that the LDF has optimal properties for the two-group classification problem if the attribute populations are multivariate normally distributed with a common covariance matrix. However, for non-normal data conditions, optimization criteria that – like the LDF – use distances based on the L_2-norm might not be appropriate. It is well known that criteria based on higher order norms tend to weigh larger distances more heavily than ones based on low order norms, and yield classification functions that are heavily influenced by extreme training sample entities. Examples of situations where extreme entities occur include data sets that are contaminated by outliers, and data sets with highly skewed distributions or distributions with heavy tails.

For the two-group classification problem, some authors have proposed classification models that optimize robust training sample accuracy criteria (Koford and Groner 1966; Ibaraki and Muroga 1970; Liittschwager and Wang 1978; Freed and Glover 1981a, 1981b; Bajgier and Hill 1982; Glover *et*

al. 1988; Stam and Joachimsthaler 1989; Duarte Silva and Stam 1994a). The classification performance of these methods is mixed, but somewhat promising for non-normal data conditions (Joachimsthaler and Stam 1988, 1990). Although many different criteria have been proposed, the most important ones are based on the location of the entities with respect to the surface which separates the two groups, *e.g.*, criteria based on the L_1-norm distances from this surface, on the number of misclassified cases, or on the total misclassification cost in the training sample. Since the derivation of such classification rules requires formulating and solving mathematical programming (MP) models, this approach is often referred to as the MP approach to classification.

In this paper we compare the MP approach with several existing parametric and nonparametric approaches, and describe BestClass, a software package that implements the most widely used two-group classification methods in the SAS System (SAS Institute Inc. 1989a, 1989b, 1989c, 1990). In Section 2, we review the most important classical approaches to two-group classification. Section 3 describes a number of methods that are based on the MP approach. In Sections 4 and 5 we use a limited simulation experiment and an example to compare the MP and classical approaches, and Section 6 presents the conclusions. More detailed information regarding the MP formulations is provided in Appendix A. In Appendix B we highlight the BestClass classification package. A related problem, cluster analysis, where entities are assigned to groups that are not defined *a priori*, will not be discussed in this paper.

2. Classical Approaches to Two-Group Classification

The classification problem in two-group DA involves estimating a rule that assigns an entity i to one of the groups based on the observed attribute vector $\mathbf{x}_i$. Denote the prior probabilities of membership in group j by π_j, and the cost of misclassifying an entity belonging to group j by C_j. The rule that assigns entity i to the group for which $\pi_j \mathrm{p}(\mathbf{x}_i \mid G_j)$ is largest minimizes the total probability of misclassification (Welch 1939). The expected cost of misclassification is minimized by the rule that assigns entity i to the group for which $C_j \pi_j \mathrm{p}(\mathbf{x}_i \mid G_j)$ is highest (Wald 1944).

The derivation of these "optimal classification rules" requires the exact knowledge of C_j, π_j and $\mathrm{p}(\mathbf{x}_i \mid G_j)$, $j = 1, 2$. In practice, usually the $\mathrm{p}(\mathbf{x}_i \mid G_j)$, π_j and C_j are not all known, and may have to be estimated. The C_j values are usually based on previous knowledge about the problem. The π_j values can either be approximated using substantive knowledge about the problem, or estimated from the proportion of elements belonging to group j in a representative training sample, *i.e.*, a representative sample with known group membership. Parametric classification methods assume that each $\mathrm{p}(\mathbf{x}_i \mid G_j)$ follows a known probability distribution, and can be described fully by a small set of parameters that can be estimated from a training sample. For instance, the assumption that $\mathrm{p}(\mathbf{x}_i \mid G_1)$ and $\mathrm{p}(\mathbf{x}_i \mid G_2)$ follow multivariate normal distributions with different mean vectors and equal

covariance matrices implies the LDF rule, while a multivariate normal distribution with different covariance matrices leads to Smith's Quadratic Discriminant Function (QDF) (Smith 1947).

There exists a vast body of literature comparing the classical two-group classification methods. The LDF has been found to provide relatively robust classification results for problems with attributes that do not follow multivariate normal distributions, but do have similar covariance matrices across groups (Lachenbruch *et al.* 1973). Likewise, the QDF is robust with respect to small deviations from normality in the case of unequal covariance matrices across groups, but is not recommended when the training sample is "small" relative to the number of attributes (Marks and Dunn 1974; Wahl and Kronmal 1977; Bayne *et al.* 1983), as quadratic classification rules have substantially more parameters than linear ones. If the underlying distributional assumptions are violated, for instance in the case of highly skewed or heavy-tailed distributions, the LDF and QDF may not give accurate classification results (Lachenbruch *et al.* 1973; Fatti *et al.* 1982), and under these conditions classification methods based on logistic regression models tend to give better results (Press and Wilson 1978; Byth and McLachlan 1980; Bayne *et al.* 1983).

As non-normal data conditions occur frequently in practice, it is of interest to explore alternative, distribution-free (nonparametric) classification methods. One approach is to replace $p(\mathbf{x}_i \mid G_j)$ in an "optimal rule" by nonparametric estimates. The most important nonparametric methods for multivariate probability density estimation are kernel and nearest neighbor methods.

Kernel methods estimate $p(\mathbf{x}_i \mid G_j)$ by a weighted average of some function (the kernel) of distances between $\mathbf{x}_i$ and the training sample entities belonging to group j. The relative weights of these distances is usually controlled by a smoothing parameter. k–nearest neighbor methods use the volume of the region containing the k training sample entities belonging to the group j that are closest to $\mathbf{x}_i$, according to some distance norm to estimate $p(\mathbf{x}_i \mid G_j)$. If the prior probabilities and misclassification costs are equal across groups, the k-nearest neighbor method simply assigns entity i to the group with the largest number of entities among the $2k+1$ training sample vectors that are closest to $\mathbf{x}_i$. Nearest neighbor methods can be thought of as a particular class of kernel methods, in which the kernel function equals a positive constant inside a neighborhood of $\mathbf{x}_i$ and zero outside this neighborhood.

Due to their flexibility, kernel and nearest neighbor methods generally classify most effectively if the attribute distributions are highly irregular and large training samples are available. For small training samples, the relative performance of kernel and nearest neighbor methods is mixed. In some studies, these methods were found to perform better than the LDF and the QDF, even for multivariate normally distributed conditional attribute densities (Gessaman and Gessaman 1972; Van Ness 1979). However, it has been argued that the kernel functions used in these studies were particularly favorable to the data conditions analyzed. Later studies with data conditions less favorable to these functions

have not confirmed the good performance of kernel methods in small training samples. Empirical evidence of performance of kernel methods for large training samples and non-normal data conditions can be found in Remme *et al.* (1980). See Murphy and Moran (1986) and McLachlan (1992, pp. 313–318) for a discussion of the performance of these methods for small training samples.

Nearest neighbor and kernel methods with fixed smoothing parameters are available in major statistical packages such as SAS or SPSS. More sophisticated kernel methods with adaptable smoothing parameters are included in specialized nonparametric estimation packages such as X-plore (Ng and Sickles 1990) and Nkernel (Delgado and Stengos 1990). Kernel and nearest neighbor methods are widely used in artificial intelligence and pattern recognition applications. However, their use is rare in business applications and social science studies. In these fields, DA applications tend to rely either on normality-based parametric (LDF or QDF) or logistic regression methods.

The most common among methods that estimate the $p(G_j \mid \mathbf{x}_i)$ directly, without the intermediate step of first estimating $p(\mathbf{x}_i \mid G_j)$, is the linear logistic model. Anderson (1972) shows that the linear logistic model is valid when the $p(\mathbf{x}_i \mid G_j)$ belong to the exponential family of distributions, for several models with binary independent variables and mixtures of continuous and binary variables. While the LDF is more efficient than the logistic method in the case of multivariate normality with a common covariance matrix (Efron 1975), the logistic method is more robust with respect to deviations from normality than the LDF (Press and Wilson 1978; Crawley 1979). Other methods for the direct estimation of $p(G_j \mid \mathbf{x}_i)$, such as the probit (Albert and Anderson 1981) and the quadratic logistic (Anderson 1975), are less effective due to the large number of parameters to be estimated.

3. MP Approaches to Two-Group Classification

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 it handles entities of obvious origin. Hence, methods that provide the best overall approximations of $p(\mathbf{x}_i \mid G_j)$ and $p(G_j \mid \mathbf{x}_i)$ do not necessarily yield the highest classification accuracy, and it may be possible to estimate accurate classification rules by focusing on the region of group overlap where $C_1 p(G_1 \mid \mathbf{x}_i)$ and $C_2 p(G_2 \mid \mathbf{x}_i)$ are about equal, instead of estimating $p(\mathbf{x}_i \mid G_j)$ over the full attribute domain. As long as the group overlap is moderate, $C_1 p(G_1 \mid \mathbf{x}_i)$ and $C_2 p(G_2 \mid \mathbf{x}_i)$ will tend to differ substantially in the central region of the distributions, and for classification purposes a good approximation of $p(\mathbf{x}_i \mid G_j)$ is required only for the tails of the distributions where the groups overlap.

A major motivation for using MP-based classification is that these methods establish the boundaries of the regions assigned to each group, without making any assumptions about the distributional characteristics of the groups. In two-group MP-based classification, the group boundaries are described by the equation $f(\mathbf{b}, \mathbf{x}) = c$, where $\mathbf{b}$ is a vector of unknown parameters and c is a

threshold value. The equation $f(\mathbf{b}, \mathbf{x}) = c$ separates the two groups, and $f(\mathbf{b}, \mathbf{x}_i)$ represents the classification score of entity i. The classification rule is to assign entity i to group 1 if $f(\mathbf{b}, \mathbf{x}_i) < c$, and to group 2 if $f(\mathbf{b}, \mathbf{x}_i) > c$, whereas the assignment of i is undetermined if $f(\mathbf{b}, \mathbf{x}_i) = c$. The estimate of $\mathbf{b}$ (and in some methods the threshold c) optimizes some criterion directly related to classification accuracy for the training sample.

Rather than estimating conditional densities or posterior group membership probabilities, most MP-based methods use the magnitude of $|f(\mathbf{b}, \mathbf{x}_i) - c|$ as a heuristic index of "confidence" in the group assignment of entity i. $|f(\mathbf{b}, \mathbf{x}_i) - c|$ represents the external (undesirable) deviation d_i if entity i is classified incorrectly, and the internal (desirable) deviation e_i if i is classified correctly. In particular, d_i is the L_1-norm distance between $\mathbf{x}_i$ and the border of the attribute region assigned to the group to which entity i belongs. Depending on the method used, the parameter estimates c and $\mathbf{b}$ may be unique only up to a proportionality factor and need to be normalized.

Similar to regression analysis, in the MP approach the form of $f(\mathbf{b}, \mathbf{x}_i)$ is assumed to be known a priori. This restriction does not impose a serious limitation, since in practice classification rules with relatively simple functional forms perform reasonably well. For instance, the LDF and QDF imply a linear and quadratic function, respectively, and for several non-normal populations the form of the optimal rule is still linear or quadratic (McLachlan 1992, p. 238), providing a rationale for using MP-based classification methods.

We next discuss the most important issues in MP-based two-group classification: the functional form of the classification rule, the accuracy criterion and the normalization scheme for c and $\mathbf{b}$. The relevant formulae are presented in Appendix A.

3.1. The Choice of Functional Form of the Classification Rule

Until recently, research on MP-based two-group classification focused on linear classification functions. Duarte Silva and Stam (1994a) and Banks and Abad (1994) extended this approach to quadratic classification functions. The issues involved in deciding between linear and quadratic functions in the MP approach correspond exactly to the factors affecting the choice between the LDF and the QDF, and between linear and quadratic logistic models. In principle, functional forms other than linear and quadratic can be used within the MP approach as well. For instance, Rubin (1994) discussed polynomial classification functions. However, we are not aware of any research that has studied the properties of such functions.

3.2. The Choice of Accuracy Criterion

One can view MP-based two-group classification as an extension of the ideas presented in Fisher's original derivation of the LDF (1936), replacing Fisher's *discrimination* criterion by a *classification accuracy* criterion, and replacing the linear function of the LDF by one that is not necessarily linear.

The first MP-based criterion proposed may be due to Koford and Groner (1966), who proposed a linear classification rule with a fixed value of c, estimating $\mathbf{b}$ such that a weighted sum of the external deviations (d_i) is minimized. However, Koford and Groner utilized an adaptive algorithm that does not guarantee convergence to the optimal solution. Smith (1968) noted that Koford and Groner's classification rule can be determined by solving a linear programming (LP) model. Mangasarian (1965) introduces LP formulations designed to estimate linear and quadratic classification functions that correctly classify all training sample entities in the case of perfectly separable groups (*i.e.*, no group overlap). Models that optimize the criterion introduced by Koford and Groner were popularized by Freed and Glover (1981b) and are known as the MSD (minimize the sum of deviations).

Another L_1-norm model is the OSD (optimize the sum of deviations) (Bajgier and Hill 1982), with a criterion involving a weighted sum of the external (d_i, to be minimized) and internal (e_i, to be maximized) deviations. Freed and Glover (1986) proposed the Maximize the Sum of Internal Deviations (MSID), which simultaneously minimizes the maximum of the weighted external deviations and maximizes the sum of the weighted external deviations. Glover *et al.* (1988) proposed HYBRID, which simultaneously considers global (common to all entities) and entity-specific deviations, and suggested some variants of HYBRID that include only the most important of the subcriteria considered in the original model. Glover (1990) notes that the these deviations cannot be interpreted in the same way as the maximum and absolute deviations from the separating surface, as they are estimated simultaneously. For notational simplicity, however, we will ignore this difference in interpretation, and use the same notation for HYBRID as the other models (see Appendix A).

The OSD and HYBRID essentially extend the MSD criterion, incorporating additional information. The MSID may be viewed as an extension of either the MMD. There is some evidence (Glover *et al.* 1988, Duarte Silva and Stam 1994a) that the inclusion of additional information can improve the classification performance. Each of these methods requires subjective judgments about the relative importance of several classification criteria. In the case of HYBRID, the number of subjective judgments can be large, and it is arguable whether the potential improvement is important enough to justify the use of less intuitive criteria.

Freed and Glover (1981a) proposed to minimize the maximum external deviations (MMD). While the MSD criterion is based on an L_1-norm distance measure, the MMD uses an L_∞-norm measure. At the opposite end of the spectrum of L_p-norm measures is the criterion to directly

minimize the number of misclassifications in the training sample (Ibaraki and Muroga 1970; Asparoukhov 1985), which can be viewed as the limit of an L_p-norm measure, as p goes to 0. Glick (1976) proves that this criterion leads to a rule that, under general regularity conditions, has an expected error rate that asymptotically approaches the minimum expected error rate among all rules of the same functional form. This result is important, because no assumption is made about the $p(\mathbf{x}_i \mid G_i)$ and $p(G_j \mid \mathbf{x}_i)$. Liittschwager and Wang (1978) show how in this formulation the per unit misclassification cost in the training sample can be minimized, by incorporating prior probabilities of group membership and misclassification costs. The Liittschwager and Wang criterion includes the sum of the group-specific misclassification proportions as a special case.

Minimizing the criterion proposed by Ibaraki and Muroga (1970) and Liittschwager and Wang (1978) requires solving a mixed-integer programming (MIP) optimization model. The solution time required for solving MIP models increases exponentially with the number of training sample entities, so that – given current technology – solving these models using commercial MP software packages is practical only for small size problems, *e.g.*, problems with at most 100 entities, up to 4 attributes and a group overlap of at most 10 percent). Recently developed specialized formulations and algorithms can alleviate the computational burden of the MIP somewhat (Warmack and Gonzalez 1973; Liittschwager and Wang 1978; Koehler and Erenguc 1990; Banks and Abad 1991; Soltysik and Yarnold 1994; Duarte Silva 1995).

Stam and Joachimsthaler (1989) proposed a criterion based on a general L_p-norm measure, with $p > 0$. For p different from 0, 1 and ∞, the estimation of the classification rule requires nonlinear programming (NLP) optimization methods. Noting that the objective function is non-convex if $p < 1$, leading to convergence problems in the optimization, Stam and Joachimsthaler (1989) do not recommend using L_p-norm measures with $0 < p < 1$. Of course, the MIP (L_0-norm) criterion also implies a non-convex model, but, as noted above, computationally intensive special-purpose solvers are available to solve MIP problems to optimality, as long as the training sample is small.

Consistent with findings in L_p-norm regression, Stam and Joachimsthaler (1989) showed that L_p-norm criteria with $1 \leq p < 2$ tend to be more robust with respect to outliers and extreme deviations than the LDF, which is based on an L_2-norm measure. Models based on an L_∞-norm distance measure, like the MMD, tend to be very sensitive to extreme entities. There is ample empirical evidence confirming that the classificatory performance of the MMD and MSID on validation samples tends to be inferior to L_1-norm models (Joachimsthaler and Stam 1990).

Most studies have found that the relative classification accuracy of L_0-norm models is sensitive to training sample sizes (Stam and Joachimsthaler 1990; Stam and Jones 1990; Koehler and Erenguc 1990; Banks 1991), and improves substantially as the training sample size increases. However, in a recent study Duarte Silva (1995) does not confirm these results, and found that L_0-norm methods did

not perform as poorly with small sample sizes as reported in previous studies, perhaps because this study used a secondary objective to break ties training sample rules associated with the same error rate, thus reducing the variability in the classification performance for the L_0-norm models.

Bajgier and Hill (1982) analyzed criteria that combine L_0- and L_1-norm distance measures. One of the simplest of these criteria is the MSD/MIP model, which uses a weighted average of the criteria used in the MSD and the MIP. However, these models have not been found to be very effective and have rarely been used in practice.

As most MP classification approaches may lead to non-unique optimal solutions, yielding several non-equivalent classification rules with the same training sample misclassification error rate (or cost), it is recommend in general to include a secondary criterion to break ties among the rules which yield the same value for the primary criterion. Including a secondary criterion implies a lexicographic MP formulation which ensures that the secondary criterion will never improve at the expense of the primary criterion. Different tie-breaking schemes can be found in Warmack and Gonzalez (1973), Bajgier and Hill (1982), Koehler (1989), Erenguc and Koehler (1990), Rubin (1990a, b), Soltysik and Yarnold (1993, 1994) and Duarte Silva (1995).

Several studies have found that MP methods tend to give better results than the LDF and QDF if the distributions are skewed or contaminated with outliers, whereas the LDF and QDF tend to perform better if the distributions are approximately normal (Bajgier and Hill 1982; Glorfeld and Olson 1982; Freed and Glover 1986; Srinivasan and Kim 1987; Rubin 1990b; Joachimsthaler and Stam 1990; Stam and Joachimsthaler 1990; Duarte Silva and Stam 1994a). Few studies have compared the MP approach with logistic regression methods or methods based on the nonparametric estimation of $p(\mathbf{x}_i \mid G_j)$ (Joachimsthaler and Stam 1988). Duarte Silva (1995) found that while MP methods based on L_0 distance norms (with an appropriate criterion to resolve ties) are particularly accurate in classifying for problemswith few attributes, skewed distributions and small training samples. Logistic regression methods generally tend to outperform the MP methods for problems with large training samples, for problems with skewed distributions, many attributes and large training samples.

3.3. The Choice of Normalization Scheme

Most MP models require a normalization constraint. A detailed discussion of this topic is beyond the scope of this paper. Rather, we discuss some considerations of this choice in Appendix A. For an overview of the advantages and disadvantages of different normalization schemes, see Markowski and Markowski (1985), Freed and Glover (1986), Erenguc and Koehler (1990), Koehler (1989), Mahmood and Lawrence (1987), Glover (1990), Rubin (1990a, b), Ragsdale and Stam (1991) and Stam and Ragsdale (1992).

4. Comparison of the Different Approaches

We next illustrate the relative performance of the different approaches using a limited Monte Carlo simulation experiment, using a three-attribute data condition with skewed distributions based on the multivariate log-normal distribution (Johnson and Kotz 1972, p. 20). Specifically, the attributes were generated using (1),

$$x_k^j = \exp(\beta^j + \Sigma_m \gamma_{km}^j z_m^j), \tag{1}$$

where x_k^j represents attribute k of group j, the z_m^j represent independently generated standard normal random variates, and the β^j, γ_{mk}^j are parameters, $j = 1, 2$; $k = 1, .., 3$; $m = 1, ..., 3$. In our experiment, we used the parameter value combinations in Table 1, yielding an expected error rate for the optimal classification rule of 6.67 percent; a group 2 to group 1 covariance ratio of 4 to 1; a within-group correlation between x_1^j and x_2^j, x_1^j and x_3^j, and x_2^j and x_3^j, of 0.8, 0.4 and 0.4, respectively, and a attribute skewness of attributes in group 1 of 10.

Table 1: Parameters of the Log-Normal Distributions Used in the Simulation Experiment

Group j	β^j	γ_{11}^j	γ_{12}^j	γ_{13}^j	γ_{21}^j	γ_{22}^j	γ_{23}^j	γ_{31}^j	γ_{32}^j	γ_{33}^j
1	−1.2087	1.1651	0	0	1.0271	0.5500	0	0.6588	0.1653	0.9466
2	0.9046	0.6103	0	0	0.5051	0.3426	0	0.2719	0.0835	0.5400

The classification methods compared in our experiment are described in Table 2, and include the LDF, QDF, logistic regression (LGST), 18 methods based on the nonparametric estimation of $p(\mathbf{x}_i \mid G_j)$ (10 nearest neighbor methods and 8 kernel methods), and 4 MP-based methods. Two different Mahalanobis distance metrics were employed in the nearest neighbor methods, based on the full and diagonal pooled sample covariance matrix, respectively. Two different types of kernel functions were used in the kernel method, one based on multivariate normality, the other on an Epanechnikov kernel function (Epanechnikov 1969). For each kernel function, we created four different kernels by combining pooled versus within-group covariance matrices with diagonal versus full kernel covariance matrices. The kernel smoothing parameters were determined by minimizing a leave-one-out estimate of the error rate in the first training sample (Lachenbruch and Mickey 1968).

The experiment involved estimating classification rules for each method using 50 different independently generated training samples with 15 entities in each group. The expected error rates were estimated by applying each function to an independently generated validation sample with 7,000 entities in each group. The means and standard deviations of the validation sample error rates across the 50 replications are provided in Table 3.

Table 2: Classification Methods Compared, Simulation Experiment

Method	Description
LDF	Fisher's linear discriminant function
QDF	Smith's quadratic discriminant function
LGST-R	Logistic regression
NND(k)	k-Nearest neighbor method with a Mahalanobis norm based on a diagonal covariance matrix
NNF(k)	k-Nearest neighbor method with a Mahalanobis norm based on a full sample covariance matrix
KNDP	Normal kernel method based on a diagonal pooled sample covariance matrix
KNDW	Normal kernel method based on a diagonal within-group sample covariance matrix
KNFP	Normal kernel method based on a full pooled sample covariance matrix
KNFW	Normal kernel method based on a full within-group sample covariance matrix
KEDP	Kernel method using an Epanechnikov kernel based on a diagonal pooled sample covariance matrix
KEDW	Kernel method using an Epanechnikov kernel based on a diagonal within-group sample covariance matrix
KEFP	Kernel method using an Epanechnikov kernel based on a full pooled sample covariance matrix
KEDW	Kernel method using an Epanechnikov kernel based on a full within-group sample covariance matrix
MIP	MP model minimizing the error rate of the training sample with a linear classification rule
MSD	MP model minimizing the sum of external deviations with a linear classification rule
MIP1	MP model minimizing the MIP objective as a primary criterion and the MSD objective as a secondary criterion with a linear classification rule
OSD	MP model minimizing a weighted sum of the external deviations (weight of $\alpha_1 = 3.0$) and the internal deviations (weight of $\alpha_2 = 0.5$) with a linear classification function

Table 3 shows that the error rates yielded by the parametric classification methods (LDF and QDF), 15 and 14 percent, respectively, were more than double that of the optimal rule (6.67 percent). This result is not surprising, because both the LDF and QDF are known to classify relatively poorly if the attribute distributions are highly skewed. The performance of the kernel and nearest neighbor methods varied considerably. The nearest neighbor methods with an odd number of neighbors and a diagonal covariance matrix tended to give the best results, with estimated error rates of about 10 percent, while the methods based on an Epanechnikov kernel yielded the worst results, with error rates between 18 and 28 percent. These results illustrate one of the major problems of these methods in general: although kernel methods may give excellent results when fine-tuned properly, their performance can be sensitive to the choice of parameter values, and general guidelines for the choice of the parameter values do not exist.

Table 3: Validation Sample Error Rates, Simulation Experiment

Classification Method	Mean	Standard Deviation
LDF	0.152	0.035
QDF	0.143	0.024
LGST-R	0.119	0.025
NNF(3)	0.125	0.030
NNF(4)	0.178	0.042
NNF(5)	0.126	0.027
NNF(6)	0.168	0.044
NNF(7)	0.131	0.029
NND(3)	0.102	0.016
NND(4)	0.133	0.027
NND(5)	0.100	0.016
NND(6)	0.126	0.026
NND(7)	0.101	0.018
KNDP	0.126	0.027
KNDW	0.166	0.059
KNFP	0.137	0.029
KNFW	0.136	0.019
KEDP	0.182	0.030
KEDW	0.186	0.040
KEFP	0.277	0.039
KEFW	0.221	0.042
MIP	0.131	0.042
MSD	0.125	0.030
MIP1	0.125	0.030
OSD	0.128	0.032

Logistic regression (12 percent) and the MP-based methods (about 13 percent) were superior to the parametric methods, but not as good as the best nearest neighbor methods. We emphasize that the current simulation study is merely intended to illustrate the various methods, rather than providing an elaborate study evaluating classification performance.

5. An Example

Consider the example two-attribute training sample in Table 4, also displayed in Figure 1, consisting of 13 entities, 7 belonging to group 1 and 6 to group 2. The first 12 entities reveal a clear pattern: the entities belonging to group 1 tend to have lower values on both attributes than those of group 2. Entity 13 is an outlier, in that it has the highest value on both attributes, although it belongs to group 1.

Table 4: Example Problem Data Set

Entity i	x_1	x_2	True Group Membership
1	3	1	1
2	2	2	1
3	1	3	1
4	4	1	1
5	3	2	1
6	2	3	1
7	4	2	2
8	3	3	2
9	2	4	2
10	5	2	2
11	4	3	2
12	3	4	2
13	6	6	1

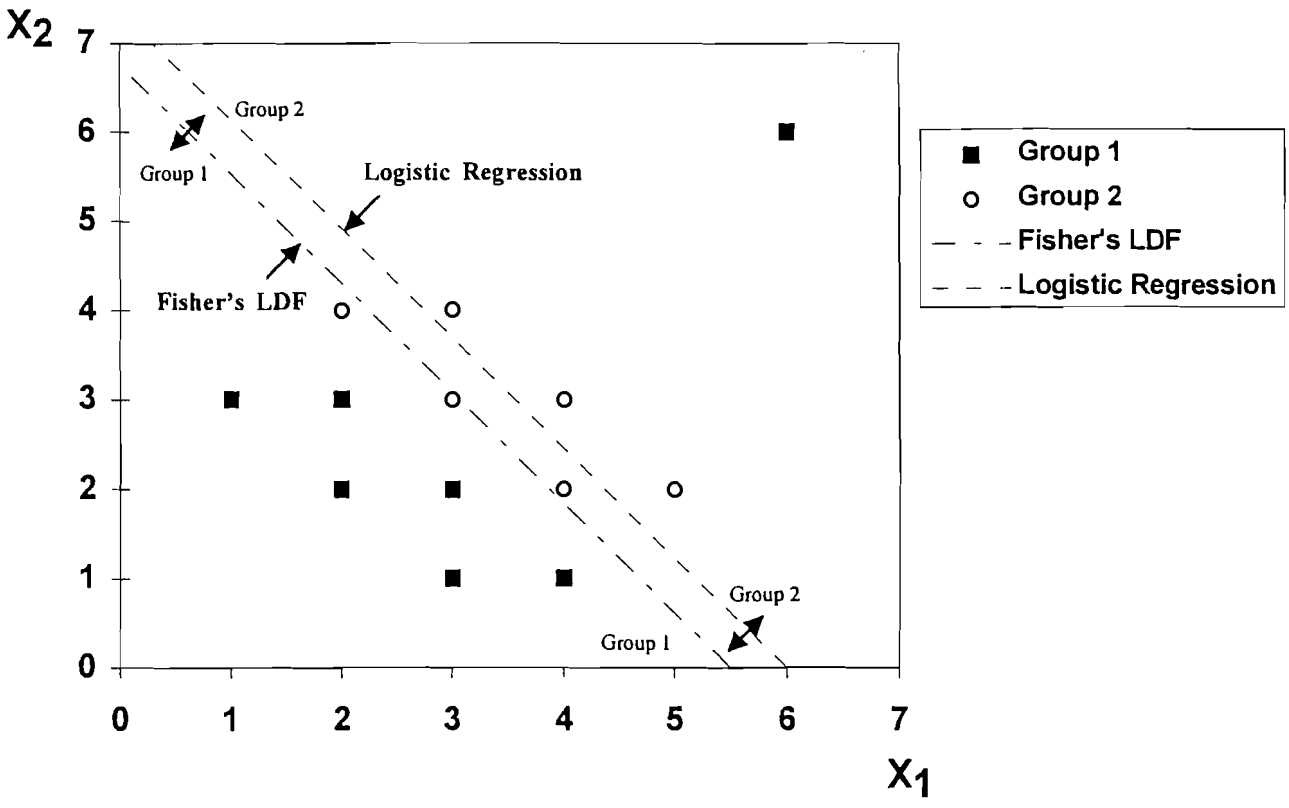

Figure 1: LDF and QDF Classification Rules, Example Problem

Table 5: Estimated Classification Rules, Example Problem

Method	Description	Normalization	b_1	b_2	c	Accuracy Criterion
1	MSD-1	$c = +1$	0.1667	0.1667	1.000	1.000
2	MSD-2	$c = -1$	−0.1667	−0.1667	−1.000	2.000
3	MSD-3	standard	0.1250	0.1250	0.7500	0.750
4	MSD-4	coefficients	0.5000	0.5000	3.0000	3.000
5	MSD-5	Glover	0.0256	0.0256	0.1538	0.154
6	EMSD	Epsilon	0.3333	0.3333	1.3333	4.667
7	MIP	Glover	0.0256	0.0256	0.1536	1.000
8	MSD/MIP	Glover	0.0256	0.0256	0.1536	1.154
9	OSD	Glover	0.0256	0.0256	0.1536	0.088
10	MMD	Glover	0.0256	0.0256	0.2308	0.077
11	MSID	Glover	0.0256	0.0256	0.2308	1.155
12	HYBRID	Glover	0.0256	0.0256	0.1410	0.293
13	LDF	–	0.2231	0.1760	1.2153	–
14	LGST-R	–	0.2677	0.2163	1.6120	–

Table 6: Values of the Deviational Variables, Example Problem

Method	1	2	3	4	5		6	10		12			
Variable	d_i	d_i	d_i	d_i	d_i	e_i	d_i^*	d_i	e_i	d_i	e_i	d_0	e_0
Entity													
1	0	0.333	0	0	0	0.051	0	0	0.128	0	0.026	0	0.013
2	0	0.333	0	0	0	0.051	0	0	0.128	0	0.026	0	0.013
3	0	0.333	0	0	0	0.051	0	0	0.128	0	0.026	0	0.013
4	0	0.167	0	0	0	0.026	0.333	0	0.103	0	0	0	0.013
5	0	0.167	0	0	0	0.026	0.333	0	0.103	0	0	0	0.013
6	0	0.167	0	0	0	0.026	0.333	0	0.103	0	0	0	0.013
7	0	0	0	0	0	0	0.333	0.077	0	0	0	0	0.013
8	0	0	0	0	0	0	0.333	0.077	0	0	0	0	0.013
9	0	0	0	0	0	0	0.333	0.077	0	0	0	0	0.013
10	0	0.167	0	0	0	0.026	0	0.051	0	0	0.026	0	0.013
11	0	0.167	0	0	0	0.026	0	0.051	0	0	0.026	0	0.013
12	0	0.167	0	0	0	0.026	0	0.051	0	0	0.026	0	0.013
13	1.0	0	0.750	3.000	0.154	0	2.667	0.077	0	0.180	0	0	0.013

Figure 1 clearly shows the influence of entity 13 on the LDF and logistic regression (LGST-R) classification rules, as both rules are shifted upward and rightward from the natural division between the groups based on the remaining 12 training sample entities, resulting in the misclassification of not only of entity 13, but also entities 7, 8 and 9, by the logistic rule. The linear classification rules estimated by the LDF, logistic and various MP methods are shown in Table 5. The accuracy criterion value for each MP-based method is included in the right-most column of Table 5. The individual d_i values for each entity i are provided in Table 6. Obviously, for each misclassified entity i, $d_i > 0$, and for each correctly classified entity, $d_i = 0$.

The first 6 MP methods listed in Table 5, MSD-1 through EMSD, are all based on the MSD accuracy criterion, but use different normalization schemes. In MSD-1 and MSD-2, the threshold value c is fixed to $+1$ and -1, respectively. From Table 5 we see that the classification rule estimated for MSD-1 assigns entity i into group 1 if $f(\mathbf{b},\ \mathbf{x}_i) = 0.167x_{i1} + 0.167x_{i2} < 1$, and into group 2 if $f(\mathbf{b},\ \mathbf{x}_i) > 1$, correctly classifying entities 1 through 6 into group 1, and entities 10, 11 and 12 into group 2. Since for entities 7, 8 and 9 $f(\mathbf{b},\ \mathbf{x}_i) = 1$, the predicted group assignment for for these entities is undetermined. Entity 13 is misclassified. The optimal accuracy criterion value for MSD-1 is $\Sigma_i d_i = 1$. In general, reversing the sign of c in the classification function may lead to entirely different results. In our example, the group assignment in MSD-1 and MSD-2 is reversed, MSD-2 yielding 9 misclassified entities and an accuracy criterion value of 2. In MSD-3 and MSD-4, c is treated as a variable and a normalization constraint is added to the problem, $b_1 + b_2 + c = 1$ in MSD-3 and $b_1 + b_2 = 1$ in MSD-4 (see also Appendix A.2). From Table 5, we note that in our example the classification rules estimated by these methods are proportional to that of MSD-1, and therefore fully equivalent. However, in general the choice of normalization can affect the classification rule estimated. To ensure that the best classification is achieved, the problem needs to be solved twice, with the group labeling reversed. Reversing the groups is equivalent to changing the sign of the right-hand side of the normalization constraint.

MSD-5 uses the Glover normalization constraint (see Appendix A.2), yielding a separating surface $f(\mathbf{b},\ \mathbf{x}_i) = 0.0256x_{i1} + 0.0256x_{i2} = 0.1536$. The corresponding rule to classify entity i into group 1 if $f(\mathbf{b},\ \mathbf{x}_i) < 0.1536$ and into group 2 if $f(\mathbf{b},\ \mathbf{x}_i) > 0.1536$, is equivalent to MSD-1, MSD-3 and MSD-4. The EMSD accuracy criterion value equals 4.667, where $d_i^* = \max(0,\ f(\mathbf{b},\ \mathbf{x}_i) - c)$ for entities in group 1, and $d_i^* = \max(0,\ -f(\mathbf{b},\ \mathbf{x}_i) + c + 1)$ for entities in group 2. Any entity i with $0 < d_i^* < 1$ is located in the classification gap. Using a threshold value $\epsilon = 0.5$, the classification rule is to classify entity i into group 1 if $f(\mathbf{b},\ \mathbf{x}_i) - c = 0.3333x_{i1} + 0.3333x_{i2} - 1.3333 < 0.5$, and into group 2 if $f(\mathbf{b},\ \mathbf{x}_i) - c > 0.5$. Thus, entities 4 through 9, which are located in the classification gap, are classified correctly ($d_i^* = 0.333 < \epsilon = 0.5$).

The classification rules estimated by MSD-1 through EMSD are displayed in Figure 2. Except for MSD-2, all methods yield a natural division of the attribute space, with only one misclassified entity, entity 13. The inferior rule fitted by MSD-2 is due to suboptimal labeling of the groups. In each of these methods, entity 13 has less influence on the estimated classification rule than in the LDF and LGST-R, as the optimization is based on minimizing the absolute values d_i, rather than least squares. In MSD-1, MSD-3, MSD-4 and MSD-5, the surface separating the groups did move toward entity 13, but fell short of misclassifying entities 7, 8 and 9. Interestingly, in the EMSD the separating surface did not pass through any of the entities.

Methods 7 through 12 in Table 5 are MP formulations based on accuracy criteria other than MSD (see Section 3.2 and Appendix A.1), in each case with the Glover normalization constraint. The classification rules estimated for the MIP, MSD/MIP (with $\alpha_1 = \alpha_2 = 1$) and the OSD (with $\alpha_1 = 3$, $\alpha_2 = 0.5$) were equivalent to that of the MSD-1. The MMD misclassified entities 9 through 13, with a

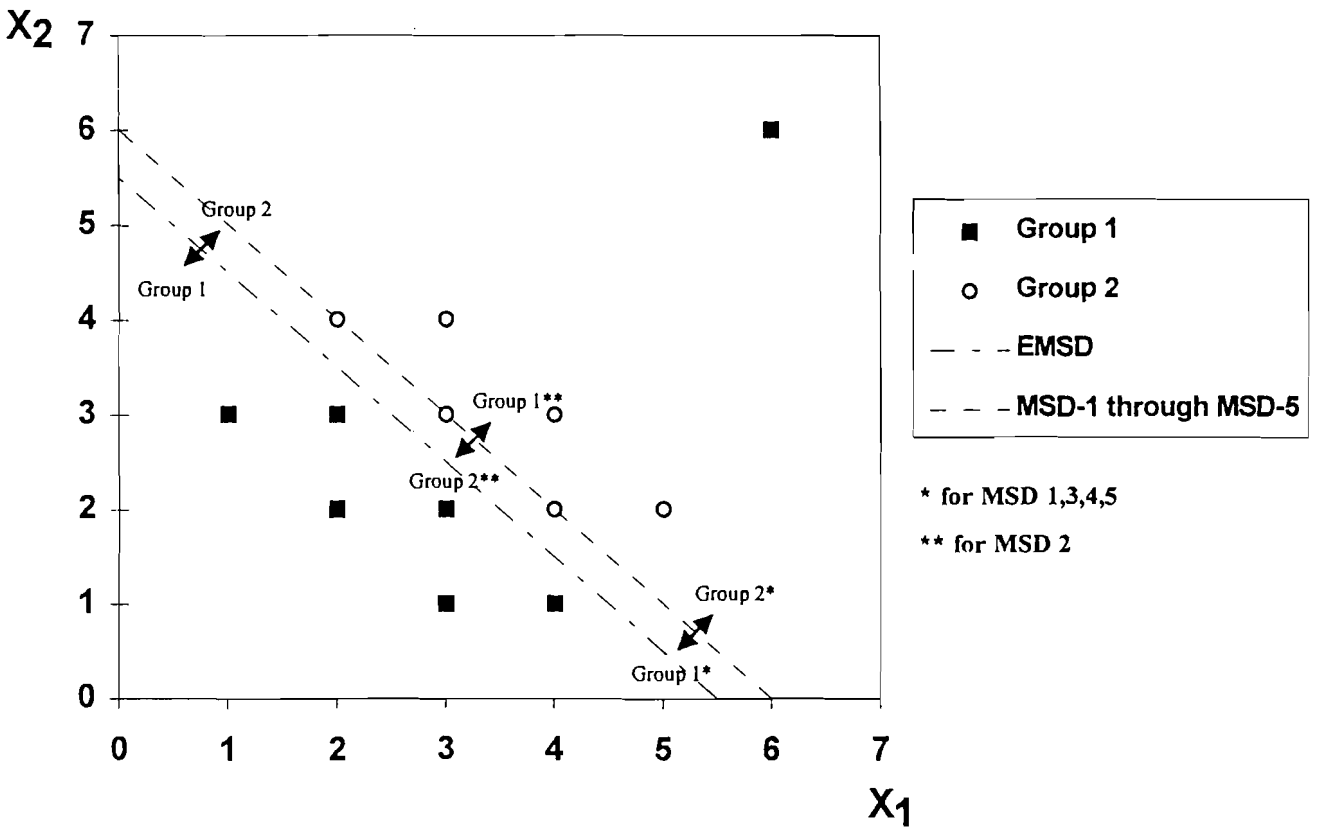

Figure 2: Classification Rules for MP Methods 1 Through 6, Example Problem

maximum external deviation of $d = 0.077$. The rule fitted by the MSID (with $\alpha_1 = 1.5 \times n = 19.5$, $\alpha_2 = 0.5$) was the same as that for the MMD, with an accuracy criterion value of 1.155. HYBRID (with $\alpha_1 = 1.5 \times n$, $\alpha_2 = 1 \times n = 13$, $\alpha_3 = 3$, $\alpha_4 = 0.5$) correctly classified all entities except 13, with an accuracy criterion value of 0.293.

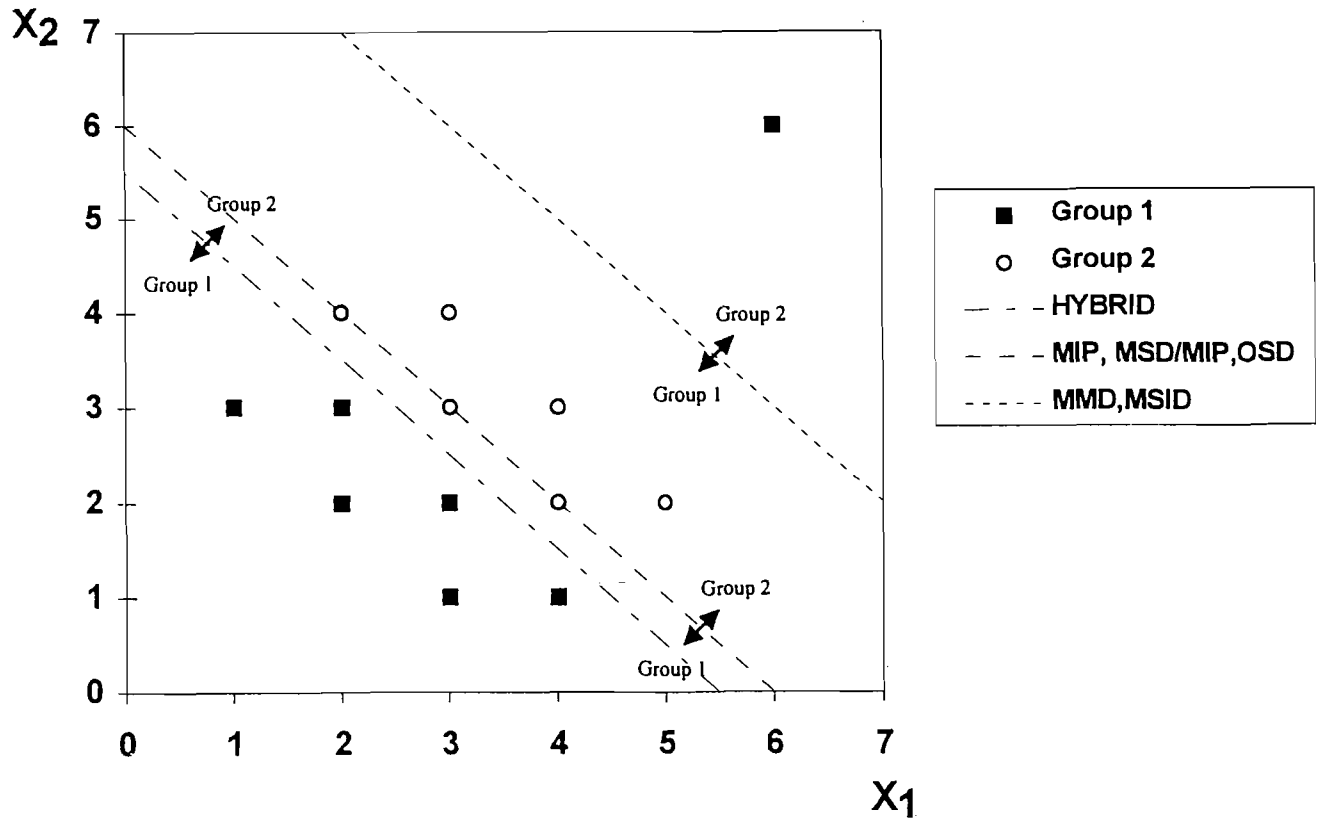

Figure 3: Classification Rules for MP Methods 7 Through 12, Example Problem

Figure 3 shows the classification rules fitted by Methods 7 through 12. Clearly, the MMD and MSID were influenced more by the presence of the outlier entity than the other methods, which is not surprising, as they are based – in full or in part – on the maximum external deviation, and thus are sensitive to outliers. In contrast to the remaining methods employing the same normalization scheme, the HYBRID and EMSD did not classify entities 9, 10 and 11 as undetermined (*i.e.*, the surface separating the groups did not pass through these entities).

6. Conclusions

In this paper, we review the MP approach to classification in two-group DA, and – in Appendix B – introduce BestClass, a software package that facilitates the use of recently developed MP-based classification methods. We also use a simulation experiment and an example to illustrate the approach and compare it with classical approaches.

In contrast with several other recently developed stand-alone computer programs and software packages that implement MP-based classification methods (Banks and Abad 1991; Abad and Banks 1993; Lam and Choo 1991; Soltysik and Yarnold 1993; Stam and Ungar 1995), BestClass implements the most important MP-based classification methods within the SAS mainframe and workstation environment, thus facilitating an MP-based classification analysis by any statistical analyst with access to the SAS system, without requiring extensive knowledge of MP techniques and solvers. The current version of BestClass requires the mainframe or UNIX versions of SAS and SAS/OR and is available under the VMS-TSO, CMS and UNIX operating systems. BestClass can be used either in batch mode or interactively. The interactive mode is menu-driven and user friendly. The batch mode facilitates the automation of repeated classification analysis which can be easily embedded within larger SAS programs. The source files and documentation of BestClass are available from the authors upon request. It is the authors' hope that the BestClass package will encourage statistical analysts to explore the MP approach to two-group classification.

APPENDIX A: SUMMARY OF ACCURACY CRITERIA, NORMALIZATIONS, CLASSIFICATION RULES, AND MODELS

This appendix details the criteria, normalization options, weighted model formulations, quadratic model formulations and classification rules available through BestClass. The generic classification rule is to classify an entity i into group 1 if $f(\mathbf{b}, \mathbf{x}_i) < c$ and into group 2 if $f(\mathbf{b}, \mathbf{x}_i) > c$, whereas the classification is undetermined if $f(\mathbf{b}, \mathbf{x}_i) = c$. The models available within BestClass are: MSD, HYBRID, HYBRID2, MMD, MIP, MSID, OSD, MSD/MIP, and "Epsilon" methods, such as the EMSD.

For each basic model formulation in BestClass, $f(\mathbf{b}, \mathbf{x})$ can be linear, quadratic with cross-products, or quadratic without cross-products. Moreover, BestClass offers the option to weigh misclassifications differently across entities. For most formulations it is possible to include a secondary accuracy criterion. BestClass also facilitates several different normalization methods. We will present each model formulation for the case of a *linear* $f(\mathbf{x})$ and *equal weights* for each entity (*i.e.*, equal costs of misclassification) only, and will use one model, the MSD, to exemplify the modified formulations that incorporate quadratic components and differential weights.

Define the 0-1 binary variable δ_i such that $\delta_i = 1$ if and only if $d_i > 0$. Depending on the type of formulation, the maximum external deviation and global external deviation across all entities in the training sample are defined by d and d_o, respectively. The external deviational variables in the EMSD are denoted by d_i^*. Similarly, e_o is the global internal deviation across all entities. The deviational variables (d_i, d_i^*, d_{ij}, d, d_o, e_i, e_o) are restricted to nonnegative values. The b_j and c (if it is a variable) are unrestricted in sign.

A.1. Accuracy Criteria and Classification Equations for the BestClass Models

1. Minimize the Sum of Absolute Deviations (MSD) Model

Classification Criterion: Minimize $\Sigma_i d_i$,

Classification Equations: $\Sigma_j b_j x_{ij} - d_i \leq c$, for training sample entities i in group 1;

$\Sigma_j b_j x_{ij} + d_i \geq c$, for training sample entities i in group 2:

2. HYBRID Model

Classification Criterion: Minimize $\alpha_1 d_o - \alpha_2 e_o + \alpha_3 \Sigma_i d_i - \alpha_4 \Sigma_i e_i$,

Classification Equations: $\Sigma_j b_j x_{ij} + e_i + e_o - d_i - d_o = c$, for training sample entities i in group 1;

$\Sigma_j b_j x_{ij} - e_i - e_o + d_i + d_o = c$, for training sample entities i in group 2.

The α_1, ..., α_4 are scalars. Not all combinations of α-values guarantee a meaningful (finite, non-trivial) solution. See Glover *et al.* (1988) and Glover (1990) for details on how to restrict the α-s.

3. HYBRID2 Model

Classification Criterion: Minimize $\alpha_1 d_o + \alpha_2 \Sigma_i d_i - \alpha_3 \Sigma_i e_i$,

Classification Equations: $\Sigma_j b_j x_{ij} + e_i - d_i - d_o = c$, for training sample entities i in group 1;

$\Sigma_j b_j x_{ij} - e_i + d_i + d_o = c$, for training sample entities i in group 2.

The HYBRID2 model is a simplified version of HYBRID, where e_o is excluded. Again, for meaningful solutions the weights of the different components of the HYBRID criterion, α_1, ..., α_3, are somewhat restricted (for details, see Glover *et al.* 1988; Glover 1990).

4. Minimize the Maximum Deviation (MMD) Model

Classification Criterion: Minimize d,

Classification Equations: $\Sigma_j b_j x_{ij} - d \leq c$, for training sample entities i in group 1;

$\Sigma_j b_j x_{ij} + d \geq c$, for training sample entities i in group 2.

In BestClass, the deviation d is restricted to be nonnegative. If for a particular data set the optimal value of d is zero, then the groups are perfectly linearly separable.

5. Minimize the Number of Misclassifications (MIP) Model

Classification Criterion: Minimize $\Sigma_i \delta_i$,

Classification Equations: $\Sigma_j b_j x_{ij} - M\delta_i \leq c$, for training sample entities i in group 1;

$\Sigma_j b_j x_{ij} + M\delta_i \geq c$, for training sample entities i in group 2.

In this model, M ("Big M") is a sufficiently large positive scalar which ensures that $\delta_i = 1$ if the corresponding deviation $d_i > 0$, and $\delta_i = 0$ otherwise. Hence, δ_i serves as a "counter" of misclassified entities. As the computational time required to solve this formulation depends heavily on the choice of M, it is important to select M large enough to ensure that $\delta_i = 1$ if and only if $d_i > 0$, but small enough that an optimal solution is reached within a reasonable time.

6. Maximize the Sum of Internal Deviations (MSID) Model

Classification Criterion: Minimize $\alpha_1 d_o - \alpha_2 \Sigma_i e_i$,

Classification Equations: $\Sigma_j b_j x_{ij} + e_i - d_o = c$, for training sample entities i in group 1;

$\Sigma_j b_j x_{ij} - e_i + d_o = c$, for training sample entities i in group 2.

The scalar parameters α_1 and α_2 reflect the relative importance of the two components of the MSID criterion. See Bajgier and Hill (1982) for some experimental classification results for different relative values of α_1 and α_2.

7. Optimize the Sum of Deviations (OSD) Model

Classification Criterion: Minimize $\alpha_1 \Sigma_i d_i - \alpha_2 \Sigma_i e_i$,

Classification Equations: $\Sigma_j b_j x_{ij} + e_i - d_i = c$, for training sample entities i in group 1;

$\Sigma_j b_j x_{ij} - e_i + d_i = c$, for training sample entities i in group 2.

In the OSD model, the relative weight of external and internal deviations is reflected by the scalar values α_1 and α_2.

8. Minimize the Sum of Deviations / Minimize the Number of Misclassifications (MSD/MIP) Model

Classification Criterion: Minimize $\alpha_1 \Sigma_i \delta_i + \alpha_2 \Sigma_i d_i$

Classification Equations: $\Sigma_j b_j x_{ij} - d_i \leq c$, for training sample entities i in group 1;

$\Sigma_j b_j x_{ij} + d_i \geq c$, for training sample entities i in group 2.

In this model, δ_i equals 1 if $d_i > 0$, and 0 if $d_i = 0$. The scalars α_1 and α_2 represent the relative weights assigned to the MIP and MSD components of the MSD/MIP criterion, respectively.

9. "Epsilon" Minimize the Sum of Deviations (EMSD) Model (available as an extension of the MSD through the Normalization Option only, not as a separate model option)

Classification Criterion: Minimize $\Sigma_i d_i^*$,

Classification Equations: $\Sigma_j b_j x_{ij} - d_i^* - c \leq 0$, for training sample entities i in group 1;

$\Sigma_j b_j x_{ij} + d_i^* - c \geq 1$, for training sample entities i in group 2.

In the "Epsilon" MSD (EMSD) formulation, which is due to Hand (1981) and has been analyzed by Ragsdale and Stam (1991), c is a variable to be estimated by the model. The EMSD formulation can be viewed as an implicit normalization scheme (see the section on normalizations below), and hence is included in BestClass as a *normalization option*, not as a separate formulation.

While the EMSD option of 0-1 right-hand sides can also be applied to model formulations other than the MSD, its mathematical properties have been established only for the EMSD model (Hand 1981; Ragsdale and Stam 1991). Ragsdale and Stam (1991) have shown that the EMSD formulation avoids any of the normalization problems that have plagued other MP formulations, and is invariant with respect to linear transformations of the attribute vector.

The classification function obtained by solving the above EMSD is likely to yield some entities in the classification gap, with classification scores $f(\mathbf{b}, \mathbf{x}) - c$ *between* 0 (the right-hand side of group 1) and 1 (the right-hand side of group 2). Thus, the EMSD requires a value $\epsilon \in [0, 1]$, such that entity i in the gap is classified into group 1 if $f(\mathbf{b}, \mathbf{x}) < \epsilon$, into group 2 if $f(\mathbf{b}, \mathbf{x}) > \epsilon$, and remains unclassified if $f(\mathbf{b}, \mathbf{x})$.

Common choices for ϵ are either $\epsilon = 0.5$ (recommended if the prior probabilities of group

membership and the misclassification costs are equal), or the value of $\epsilon \in [0, 1]$ for which the number of misclassified training sample entities (or the misclassification cost) is minimized. Both of these options are available in BestClass. The value of ϵ can also be specified by the user.

As discussed above, it is recommended to use a secondary criterion to resolve ties in the primary criterion. In BestClass, a secondary criterion can be included by selecting the appropriate combination of formulation and weights. For instance, selecting the MSD/MIP formulation, with α_1 much larger than α_2, implies an MIP formulation, where a tie for the minimum number of misclassifications is resolved by the secondary criterion of mimimizing the sum of absolute external deviations.

We also note that for optimal classification, it may be necessary to solve the problem twice, with the group assignment (1 *vs.* 2) reversed. If a scalar value is used for c, this can be achieved by solving the problem twice – once with $c > 0$, and again with $c < 0$. If the "Standard" or "Coefficients" normalization scheme (see Appendix A.3) is used, this can be achieved by solving the problem twice, once with the right-hand side of the normalization constraint of 1, and again with a right-hand side of -1.

A.2. Normalization Functions

Several models require a normalization of the classification function coefficients to ensure that a meaning (non-trivial) classification rule will result from the analysis. The default value in BestClass for formulations with a *scalar-valued* intercept term c is $c = 1$, and in this case none of the normalizations below is applied. The following normalization options are available in BestClass, through the *Choose Normalization* option:

(1) "Coefficients" Normalization Constraint: $\Sigma_j b_j = r$.

In this normalization constraint, the sum of estimated attribute coefficients equals a non-zero scalar r.

(2) "Standard" Normalization Constraint: $\Sigma_j b_j + c = r$

This normalization constraint scales the sum of estimated attribute coefficients plus c to a non-zero scalar r.

(3) HYBRID Normalization Constraint: $\Sigma_i e_i + e_o = 1$

This normalization constraint, proposed by Glover, Keene and Duea (1988), applies exclusively to the HYBRID formulation. In a more recent paper, Glover (1990) shows that this normalization has some undesirable properties, and recommends using the Glover constraint introduced below in (5).

(4) HYBRID2 Normalization Constraint: $\Sigma_i e_i = 1$

(5) Glover Normalization Constraint: $(-n_2\Sigma_{i \in G_1}\mathbf{x}_i^T + n_1\Sigma_{i \in G_2}\mathbf{x}_i^T)\mathbf{b} = 1$

In this constraint, as before $\mathbf{x}_i$ is the vector of attribute values of entity i, n_1 and n_2 are the number of training sample entities in group 1 and 2, respectively, c in the problem constriants is a *variable* intercept term, and $\mathbf{b} = (b_1, ..., b_k)^T$ is the vector of estimated attribute coefficients. The b_j and c are unrestricted in sign. This normalization renders the MSD formulation invariant to group labeling and invariant with respect to linear transformations of the attribute vector.

(6) Epsilon-Constraint:

See the "Epsilon" method formulation above (as exemplified by the EMSD) for a discussion of this normalization constraint.

A.3. Weighted MSD Formulation

Minimize the Weighted Sum of Absolute Deviations (WMSD) Model

Classification Criterion: Minimize $\Sigma_i w_{i1} d_{i1} + \Sigma_i w_{i2} d_{i2}$,

Classification Equations: $\Sigma_j b_j x_{ij} - d_{i1} \leq c$, for training sample entities i in group 1;

$\Sigma_j b_j x_{ij} + d_{i2} \geq c$, for training sample entities i in group 2.

In this weighted MSD formulation, the criterion weight of deviation d_{ij} associated with entity i from group j equals $w_{ij} \geq 0$. The extension of including individual entity-wise criterion weights in other non-weighted formulations is similar to the extension of the MSD to WMSD.

A.4. Quadratic MSD Formulation (Without Cross-Products)

Quadratic formulations have been shown to lead to improved classification results for certain data conditions. The classification equations are nonlinear in the original attributes (x_{ij}), and imply a nonlinear separating surface.

Minimize the Sum of Absolute Deviations (MSDQ1) Model

Classification Criterion: Minimize $\Sigma_i d_i$,

Classification Equations: $\Sigma_j b_{jL} x_{ij} + \Sigma_j b_{jQ} (x_{ij})^2 - d_i \leq c$, for training sample entities i in group 1;

$\Sigma_j b_{jL} x_{ij} + \Sigma_j b_{jQ} (x_{ij})^2 + d_i \geq c$, for training sample entities i in group 2.

The generalization, for other formulations, from the linear to the quadratic formulation without cross-products is analogous to that for the MSD.

A.5. Quadratic MSD Formulation (With Cross-Products) Minimize the Sum of Absolute Deviations (MSDQ2) Model

Classification Criterion: Minimize $\Sigma_i d_i$,

Classification Equations:

$\Sigma_j b_{j\mathrm{L}} x_{ij} + \Sigma_j b_{j\mathrm{Q}} (x_{ij})^2 + \Sigma_{h \neq m} b_{hm} x_{ih} x_{im} - d_i \leq c$, for training sample entities i in group 1;

$\Sigma_j b_{j\mathrm{L}} x_{ij} + \Sigma_j b_{j\mathrm{Q}} (x_{ij})^2 + \Sigma_{h \neq m} b_{hm} x_{ih} x_{im} + d_i \geq c$, for training sample entities i in group 2:

For other formulations, the generalization from the linear to the quadratic formulation with cross-products is analogous to that for the MSD.

A.6. Classification Functions

BestClass Classification Functions for all Formulations, Except the "Epsilon" Formulation

For each formulation except "Epsilon" formulations, the linear classification function is of the form $f(\mathbf{b}, \mathbf{x}) = \Sigma_j b_j x_{ij}$, while the quadratic classification function with cross-products is $f(\mathbf{b}, \mathbf{x}) = \Sigma_j b_{j\mathrm{L}} x_{ij} + \Sigma_j b_{j\mathrm{Q}} (x_{ij})^2 + \Sigma_{h \neq m} b_{hm} x_{ih} x_{im}$, and that without cross-products $f(\mathbf{b}, \mathbf{x}) = \Sigma_j b_{j\mathrm{L}} x_{ij} + \Sigma_j b_{j\mathrm{Q}} (x_{ij})^2$. The classification rule is to classify entity i into group 1 if $f(\mathbf{b}, \mathbf{x}) < c$, into group 2 if $f(\mathbf{b}, \mathbf{x}) > c$, while i is unclassified if $f(\mathbf{b}, \mathbf{x}) = c$.

BestClass Classification Function for the "Epsilon" Formulation

The classification rule for the "Epsilon" formulation is to classify entity i into group 1 if $\Sigma_j b_j x_{ij} < c + \epsilon$, into group 2 if $\Sigma_j b_j x_{ij} > c + \epsilon$, while the classification of i is undetermined if $\Sigma_j b_j x_{ij} = c + \epsilon$.

Appendix B: BestClass Description

B.1. Overview

BestClass is a software package that implements the most widely two-group classification methods based on the MP approach in the SAS System (SAS Institute 1989a, 1989b, 1989c, 1990). BestClass is almost entirely written in the SAS macro language (SAS 1990) and uses the SAS/OR System (1989a) to solve the relevant MP models. The few files not written in the SAS macro language are system-dependent, and provide the interface with several host operating system. At the time of this writing, interfaces are available for the VMS-TSO, CMS and UNIX operating systems. BestClass can also be extended to work with PC-SAS, but the interface between BestClass and operating systems for PC's (DOS, Windows and OS2) have not been developed yet.

BestClass provides the following major features:

- BestClass can be used in two operating modes, interactive and batch.
- BestClass facilitates an analysis based on three types of classification function: linear, quadratic, and quadratic without cross-products.
- BestClass offers a choice between several nonparametric accuracy criteria.
- BestClass allows for assigning individual costs to each entities.
- BestClass can accept input and direct output through either text files or the SAS data set format.
- BestClass allows the retrieval of previously fitted classification functions, which can subsequently used to classify new data sets.
- BestClass offers an easy interface with other SAS programs, and can easily be embedded in larger SAS programs.

BestClass is implemented as a series of SAS macros. The %control macro controls the flow of the program under interactive mode, and *%bestc* controls the flow under batch mode. BestClass requires at least one input data set, which is either the "training sample" used to estimate (fit) a new classification function, or a "current sample," for which the (either known or unknown) group membership of the entities can be predicted using a previously saved classification function. In Appendix B.2 we described the interactive mode, followed by a brief discussion of the batch mode in Appendix B.3. An extensive description of the batch mode features of BestClass is contained in Duarte Silva and Stam (1994b).

B.2. Interactive Mode

In interactive mode, BestClass is activated by calling an executable file (*i.e.*, a CLIST under VMS-TSO, a REXX programming file under CMS, or a shell file under UNIX), which first allocates all necessary source files and then invokes the SAS System and the macro *%control*. BestClass prompts the user for the name and location of the input data set (either in text file or SAS data set format), and for the number of attributes per entity. Once this information has been provided, the main menu of BestClass will be displayed. This menu offers the following options:

D: Define data sets.

This option allows the user to define a SAS data set of one of the following types: (1) the training sample data set, which contains a sample of entities that will be used to fit the classification function, (2) the current sample data set, consisting of entities that are to be classified according to some already existing classification function, and (3) the parameter data set, which contains several parameter values that define the BestClass environment.

E: Create new data sets from **E**xternal files.

Using this option, the user can convert text files to SAS data set format.

F: Choose classification function **F**orm.

This option enables the user to choose between three types of classification function: linear, quadratic, and quadratic without cross-products.

I: View the **I**ndividual classification results.

This option presents the individual classification results for the most recently classified data set.

L: **L**ist external file.

This option allows the user to view text files without leaving BestClass.

M: Choose **M**odel

This option allows the user to choose between a number of MP-based models, each implementing a different accuracy criterion. The models available through BestClass are presented in Appendix A.1.

N: Choose **N**ormalization option.

The estimated classification functions are unique up to a scaling constant. A unique function can be achieved by normalization. The *Choose Normalization* option allows the user to choose between several of the most important normalization schemes, as described in Appendix A.2.

O: Switch to **O**perating system shell.

This option enables the user to switch to the operating system shell without leaving BestClass.

P: Change BestClass Parameters.

This option enables the user to change several of the parameters of BestClass that, among others, control the number of attributes describing each entity, the precision level used for displaying the results of the analysis, the relative importance of each subcriteria in models that combine several criteria, the value of the threshold of c when c is a scalar.

R: Run model.

This option invokes the PROC lp of the SAS/OR System to solve the MP formulation selected using the *Choose Model* option.

S: Save output.

This option saves the results of the BestClass analysis.

AC: Apply Classification function to current data set.

MS: Manage/list SAS data sets and libraries.

This option enables the user to access basic SAS library management utilities, such as utility programs to create, allocate, delete, merge and list SAS libraries and data sets, without leaving BestClass.

RF: Retrieve classification Function from SAS data set.

This option retrieves a classification function previously saved as a SAS data set, and makes it available to classify the current data set.

X: EXit BestClass.

This option terminates BestClass, and exits the user to the operating system.

B.3. Batch Mode

In batch mode, BestClass can be activated by calling the macro *%bestc* from a SAS program. When invoking *%bestc*, one can optionally specify parameters that, among others, control the origin and form of the input, the amount and destination of the output, the accuracy measure to be used, and the form of the classification function. Except for those options that are relevant in interactive mode only (such as the listing of external files, switching to the operating system, and the interactive management of SAS libraries), all options available in the interactive mode of BestClass can also be accessed in batch mode, by selecting an appropriate combination of parameter values.

External calls of BestClass to or from a non-SAS System environment are best accommodated through text files. However, SAS data sets are more efficient if BestClass communicates with other programs within the SAS System environment.

A complete list of %bestc options, their use and default values is given in the documentation of BestClass (Duarte Silva and Stam, 1994b) and will be made available together with the source files of BestClass upon request to the authors.

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.

Albert, A. and Anderson, J. A., "Probit and Logistic Discriminant Functions," *Communications in Statistics – Theory and Methods*, **A10** (1981), 641–657.

Anderson, J. A., "Separate Sample Logistic Discrimination," *Biometrika*, **59** (1972), 19–35.

Anderson, J. A., "Quadratic Logistic Discrimination," *Biometrika*, **62** (1975), 149–154.

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

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

Banks, W. J., *New Solution Algorithms for the Classification Problem*, Unpublished Doctoral Thesis, McMaster University, Hamilton, Ontario, Canada, 1991.

Banks, W. J. and Abad, P. L., "An Efficient Optimal Solution Algorithm for the Classification Problem," *Decison 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.

Bayne, C. K., Beauchamp, J. J., Kane, V. E. and McCabe, G. P., "Assessment of Fisher and Logistic Linear and Quadratic Discrimination Models," *Computational Statistics and Data Analysis*, **1** (1983), 257–273.

Byth, K. and McLachlan, G. J., "Logistic Regression Compared to Normal Discrimination for Non-normal Populations," *Australian Journal of Statistics*, **22** (1980), 188–196.

Crawley, D. R., "Logistic Discrimination as an Alternative to Fisher's Linear Discriminant Function," *New Zealand Statistician*, **14** (1979), 21–25.

Delgado, M. A. and Stengos, T., "N-Kernel: a Review," *Journal of Applied Econometrics*, **5** (1990), 299–304.

Duarte Silva, A. P., *Minimizing the Misclassification Costs in Two-Group Classification Analysis*, Unpublished Doctoral Dissertation, Terry College of Business, The University of Georgia, Athens, GA, 1995.

Duarte Silva, A. P. and Stam, A., "Second Order Mathematical Programming Formulations for Classification in Two-Group Discriminant Analysis," *European Journal of Operational Research*, **72** (1994a), 4–22.

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 30602, 1994b.

Efron, B., "The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis," *Journal of the American Statistical Association*, **70** (1975), 892–898.

Epanechnikov, V. A., "Nonparametric Estimation of a Multivariate Probability Density," *Theory Prob., Appld. USSR*, **14** (1969), 153–158.

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), 215–225.

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, pp. 1–71.

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** (1986), 151–162.

Gessaman, M. P. and Gessaman, P. H., "A Comparison of Some Multivariate Discrimination Procedures," *Journal of the American Statistical Association*, **67** (1972), 468–472.

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

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.

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

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

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.

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

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

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.

Lachenbruch, P. A. and Mickey, M. R., "Estimation of Error Rates in Discriminant Analysis," *Technometrics*, **10** (1968), 1–11.

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, 1991.

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

Mahmood, M. O. 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.

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

Marks, S. and Dunn, O. J., "Discriminant Functions when Covariance Matrices are Unequal," *Journal of the American Statistical Association*, **69** (1974), 555–559.

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

Murphy, B. J. and Moran, M. A., "Parametric and Kernel Density Methods in Discriminant Analysis: Another Comparison," *Computers and Mathematics with Applications*, **12A** (1986), 197–207.

Ng, P. T. and Sickles, R. C., "Xplore-ing the World of Nonparametric Analysis," *Journal of Applied Econometrics*, **5** (1990), 293–298.

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

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.

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.

SAS Institute Inc., *SAS/OR® User's Guide, Version 6, First Edition*, SAS Institute Inc., Cary, NC, 1989a.

SAS Institute Inc., *SAS/STAT® User's Guide, Version 6, First Edition, Volume 1*, SAS Institute Inc., Cary, NC, 1989b.

SAS Institute Inc., *SAS/STAT® User's Guide, Version 6, First Edition, Volume 1*, SAS Institute Inc., Cary, NC, 1989c.

SAS Institute Inc., *SAS® Language: Reference, 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*, **17** (1968), 367–372.

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

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

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., "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), 113–122.

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** (1990), 243–253.

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

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

Van Ness, J. W., "On the Effects of Dimension in Discriminant Analysis for Unequal Covariance Populations," *Technometrics*, **21** (1979), 119–127.

Wahl, P. W. and Kronmal, R. A., "Discriminant Functions when Covariances are Unequal and Sample Sizes are Moderate," *Biometrics*, **33** (1977), 479–484.

Wald, A., "On a Statistical Problem Arising in the Classification of an Individual Into One of Two Groups," *Annals of Mathematical Statistics*, **15** (1944), 145–162.

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*, **C22** (1973), 1065–1075.

Welch, B. L., "Note on Discriminant Functions," *Biometrika*, **31** (1939), 218–220.