SECOND ORDER MATHEMATICAL PROGRAMMING FORMULATIONS FOR DISCRIMINANT ANALYSIS

Antonio Pedro Duarte Silva
Dept. of Management Sciences and Information Technology,
The University of Georgia, Athens, GA, USA
and Curso de Administração e Gestão de Empresa,
Universidade Católica Portuguesa, Porto, Portugal

Antonie Stam
International Institute for Applied Systems Analysis
Laxenburg, Austria
and Dept. of Management Sciences and Information Technology,
Terry College of Business, The University of Georgia, Athens, GA, USA

RR-95-7
July 1995

Reprinted from *European Journal of Operational Research* 72(1994)4–22.

INTERNATIONAL INSTITUTE FOR APPLIED SYSTEMS ANALYSIS
Laxenburg, Austria

Reprinted with permission from *European Journal of Operational Research* 72(1994)4–22.

Printed by Novographic, Vienna, Austria.

Foreword

Classification analysis is one of the most widely-used statistical decision making tools. While traditional, normality-based parametric classification methods play an important role in classification, the distributional properties of many data sets in practice do not satisfy normality. Linear and quadratic parametric classification methods have shown to be fairly robust with respect to deviations from normality. However, it is well-known that their performance is sensitive to, for instance, outlier observations and high degrees of skewness. To address this issue, an important class of nonparametric classification methods has been recently developed that use the absolute error criterion in estimating the optimal linear classification function. Several studies have shown that these methods can yield effective and robust classifiers if the data contain outlier observations. Previous absolute error-based methods were limited to linear classifiers, which limited their application as in many situations nonlinear classification fucntions are better able to separate the groups. The current paper extends the framework of linear absolute error-based classification methods to the case of nonlinear (e.g., second order, quadratic, and polynomial) classifiers, and clearly shows through several simulation experiments that nonlinear classification rules based on the absolute error criterion can yield excellent results, thus providing a valuable contribution to nonparametric classification.

Jaap Wessels
Leader, Methodology of
Decision Analysis

European Journal of Operational Research 72 (1994) 4–22
North-Holland

Second order mathematical programming formulations for discriminant analysis

Antonio Pedro Duarte Silva

Department of Management Sciences and Information Technology, Terry College of Business, The University of Georgia, Athens, GA 30602, USA, and Curso de Administração e Gestão de Empresa, Universidade Católica Portuguesa, Centro Regional do Porto, Rua Diogo Botelho 1327, 4100 Porto, Portugal

Antonie Stam

Department of Management Sciences and Information Technology, Terry College of Business, The University of Georgia, Athens, GA 30602, USA

Received November 1992

Abstract: This paper introduces a nonparametric formulation based on mathematical programming (MP) for solving the classification problem in discriminant analysis, which differs from previously proposed MP-based models in that, even though the final discriminant function is linear in terms of the parameters to be estimated, the formulation is quadratic in terms of the predictor (attribute) variables. Including second order (i.e., quadratic and cross-product) terms of the attribute variables in the model is similar in concept to the usual treatment of multiple predictor variables in statistical methods such as Fisher's linear discriminant analysis, and allows an analysis of how including nonlinear terms and interaction effects affect the predictive ability of the estimated classification function. Using simulation experiments involving data conditions for which nonlinear classifiers are appropriate, the classificatory performance of this class of second order MP models is compared with that of existing statistical (linear and quadratic) and first order MP-based formulations. The results of these experiments show that the proposed formulation appears to be a very attractive alternative to previously introduced linear and quadratic statistical and linear MP-based classification methods.

Keywords: Linear programming; Nonparametric statistics; Linear statistical models; Discriminant analysis

1. Introduction

The classification problem in discriminant analysis is concerned with correctly classifying observations into well-defined groups or classes, when group membership of these observations is either known or unknown (Huberty, 1984). The usual procedure is to estimate classification rules which classify the

Correspondence to: Prof. A. Stam, Department of Management, Terry College of Business, The University of Georgia, Athens, GA 30602, USA.

SSDI 0377-2217(93)E0009-M

observations in the training sample for which the group membership is known, based on known characteristics of these observations. Once established, the appropriate classification rules are subsequently used to predict group membership of future observations for which the group membership is unknown. Discriminant analysis has been applied in various different disciplines, including medical disease diagnosis (Spiegelhalter and Knill-Jones, 1984), the social sciences, psychology (Huberty, 1984) and business disciplines including accounting, marketing, finance (Eisenbeis, 1977), bond rating, corporate bankruptcy, new product success, credit granting, and personnel management.

Existing parametric statistical methods include Fisher's linear discriminant function (LDF) (Fisher, 1936) and Smith's quadratic discriminant function (QDF) (Smith, 1947). The LDF and QDF are based on the assumption that for each group the attribute variables follow a multivariate normal distribution, with equal and unequal variance-covariances across groups, respectively. However, it has been shown that the LDF and QDF, albeit fairly robust with respect to mild violations of the normality assumption, may not accurately predict class membership if the normality assumption is violated to a considerable extent, for instance in the presence of outlier observations (Eisenbeis, 1977; Glorfeld and Kattan, 1989; Stam and Ragsdale, 1992).

Inspired in part by the apparent weaknesses of existing classification methods, a number of different nonparametric mathematical programming (MP) formulations for solving the classification problem in discriminant analysis have been proposed the recent years. In the presence of outliers, and for several non-normal data conditions, MP-based methods have proven to be viable alternatives to the LDF and QDF (Glorfeld and Olson, 1982; Stam and Joachimsthaler, 1989; Stam and Ragsdale, 1992). Several researchers have indicated that in business-related problems outlier-contaminated data conditions are not uncommon (Glorfeld and Kattan, 1989; Mahmood and Lawrence, 1987), and that as much as ten percent of typical financial data may consist of outliers (Hample et al., 1986). Therefore, it is not surprising that MP-based methods have attracted considerable attention from business-related research areas.

The most commonly used MP formulations are the minimize the sum of deviations (MSD) method (Freed and Glover, 1981b; Hand, 1981), the minimize the maximum deviation (MMD) method (Freed and Glover, 1981a), the minimize the number of misclassifications (MIP) method (Bajgier and Hill, 1982; Gehrlein, 1986; Rubin, 1990a), and Hybrid methods (Freed and Glover, 1986; Glover, Keene and Duea, 1988; Glover, 1990) which, among others, combine ideas of the MSD and MMD methods. A concise review of MP formulations is provided by Erenguc and Koehler (1990). A number of studies have compared these methods with each other and with statistical methods such as the LDF and QDF in terms of their classification performance, using either real or simulated data (Bajgier and Hill, 1982; Freed and Glover, 1986; Joachimsthaler and Stam, 1988; Koehler and Erenguc, 1990; Mahmood and Lawrence, 1987; Markowski and Markowski, 1987; Rubin, 1989, 1990a; Stam and Joachimsthaler, 1989, 1990; Stam and Jones, 1990).

Taking stock of the empirical evidence published to date, a recent survey article (Joachimsthaler and Stam, 1990) concludes that the classification performance of MP-based methods appears to rival that of the LDF and QDF, and note that these methods were in fact found to perform better under certain conditions. The evidence, however, is not uniformly supportive of the MP-based methods (see, e.g., Rubin, 1990b), and claims to the effect that these methods are clearly superior appear unwarranted given the empirical results. Nevertheless, there is a fair amount of support for the statement that the MSD and MIP methods have classified surprisingly well in the presence of outliers in training (and validation) samples, whereas the MMD results are very sensitive to the presence of outliers and tend to yield classification results which are inferior to the MSD and LDF (Freed and Glover, 1986; Markowski and Markowski, 1987; Stam and Joachimsthaler, 1989).

One drawback of the MP-based methods developed to date is that without exception the classification functions under consideration are linear. This choice of functional form is more likely due to limitations of MP packages and algorithms in terms of the types of functions which can be handled, than to a rigorous analysis of the nature of the data. In several recent comparative studies, *linear* classification functions were estimated (using MSD, MIP or MMD) and their classificatory performance evaluated,

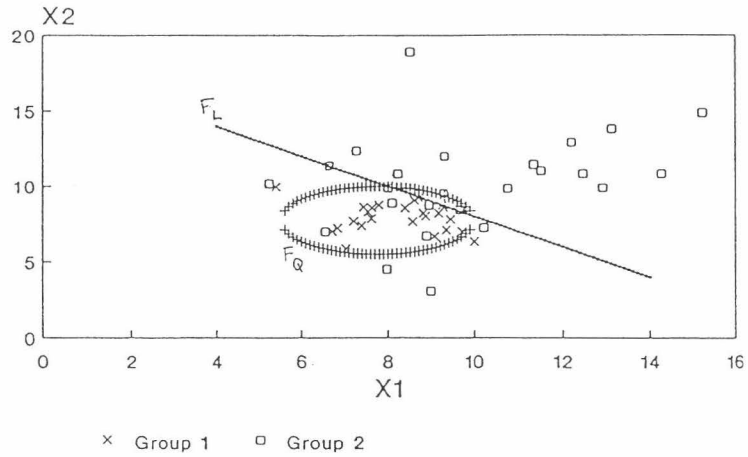

Figure 1. Two-group discriminant problem with two attribute variables for which a nonlinear classification function is clearly appropriate

even though a *nonlinear* (quadratic) function was clearly appropriate given the data conditions (Koehler and Erenguc, 1990; Stam and Joachimsthaler, 1990). For example, Figure 1 illustrates for the two-group discriminant problem with two attribute variables X_1 and X_2, that if the variance–covariances are strongly heterogeneous across groups, the linear function indicated by F_L is clearly incapable of correctly separating the two groups, while the quadratic function F_Q yields much better classification results.

A second drawback is that none of the MP-based methods proposed to date take interaction effects of the attributes into account explicitly, even though it has been demonstrated that these methods can provide efficient classifiers in the presence of correlation (Freed and Glover, 1986). As a result, only part of the information contained in the data set is used. Note that both the LDF and QDF methods do include the estimated variance-covariance structure of the predictor variables in their analysis (Anderson, 1984; Morrison, 1990). Therefore, while in situations where the null hypothesis of independently distributed variables cannot be rejected, it is reasonable to use previously proposed MP-based estimation methods, this may not always be the case if the variables are strongly correlated.

In the current paper, we propose to include quadratic and cross-product terms of the attribute variables (as well as the usual linear terms) in the classification rule, after which the analyst can proceed with solving the problem in modified form using the MP method of his/her choice. Within the framework of our model formulation it is possible to include other more general nonlinear terms, but in the current paper we limit ourselves to the quadratic case. Although our model is nonlinear in terms of the attributes, it still has a linear objective function and constraints which are linear in terms of the parameters to be estimated. In the remainder of the paper we will refer to our modified formulation as the second order MP model, in contrast with first order models which contain constraints which are linear in the attribute variables as well as the parameters to be estimated.

As alluded to above, the second order model addresses some of the weaknesses associated with previous MP formulations, while preserving the potential advantage of estimating robust classification rules using the appropriate MP criterion. In fact, the objective function in the second order formulation is the same as that in the corresponding MP model with a linear classification function. The second order formulation appears attractive in situations where a nonlinear classification function – in particular a quadratic function – is appropriate, rather than a linear classifier. We recommend that in general the decision of which terms (linear, quadratic or other) to include in the modified MP formulation be based on a preliminary analysis of the data at hand. Conducting a rigorous preliminary data analysis is common

practice in applied statistics, constitutes a natural and essential step in the process of data analysis, and should not present an unreasonable additional burden to the statistical analyst.

In general, depending on the nature of the data and the interrelationships of the attribute variables, a number of different types of functions and cross-products of the attribute variables can be included in the second order MP formulation. Our choice of analyzing models with second order terms, as opposed to including other functional forms, appears of particular interest, because it is consistent with the types of terms considered by parametric statistical procedures for analyzing the discriminant problem, such as the QDF. We view the contribution of our proposed formulation to be that it enables the analyst to analyze a larger class of data conditions in a meaningful way using existing MP procedures, without a significant increase in the modeling effort.

The remainder of our paper is organized as follows. In Section 2, the MSD and Hybrid formulations of the discriminant problem are reviewed. Section 3 introduces the corresponding second order model formulations. Simulation experiments were conducted to show how the inclusion of quadratic terms in the classification function can improve the classification accuracy of several previously introduced MP formulations. The design and results of these experiments are presented in Section 4. The paper concludes with final remarks in Section 5.

2. Existing MSD and Hybrid model formulations

We will focus our paper on two of the most successful MP formulations for analyzing the discriminant problem, the MSD (Freed and Glover, 1981b; Hand, 1981) and Hybrid (Freed and Glover, 1986; Glover, Keene and Duea, 1988; Glover, 1990) methods. The adaption of other MP formulations to the second order case is very similar. For reasons of simplicity we will restrict ourselves to the two-group case. The extension to more than two groups is conceptually straightforward, but notationally cumbersome (Freed and Glover, 1981b). Suppose the training samples consist of n_r observations in group r ($r = 1,2$), for a total of $n = n_1 + n_2$ observations in both groups combined. Let k be the number of predictor variables (attributes), and denote the attribute values for observation i by $\boldsymbol{A}_i = (a_{i1}, \ldots, a_{ik})^{\mathrm{T}}$. The MSD model may be expressed as Problem I.

(Problem I)

$$\text{Minimize} \quad z_1 = \sum_{i \in G_1} d_{i1}^{+} + \sum_{i \in G_2} d_{i2}^{-} \tag{1}$$

$$\text{subject to} \quad \sum_j a_{ij} x_j - d_{i1}^{+} + d_{i1}^{-} \leq c, \quad i = 1, \ldots, n_1, \tag{2}$$

$$\sum_j a_{ij} x_j - d_{i2}^{+} + d_{i2}^{-} > c, \quad i = 1, \ldots, n_2, \tag{3}$$

$$\sum_j x_j + c = 1, \tag{4}$$

$$d_{ir}^{+}, d_{ir}^{-} \geq 0, \quad i = 1, \ldots, n_r, \quad r = 1, 2, \tag{5}$$

$$x_j, c \text{ unrestricted}, \quad j = 1, \ldots, k \tag{6}$$

where the d_{ir}^{+} and d_{ir}^{-} ($i = 1, \ldots, n_r$; $r = 1,2$) are deviational variables, c is the cut-off value, and the x_j represent the attribute weights. Constraint (4) is a normalization constraint which serves to ensure that the trivial solution is not selected as the optimal classification rule. It is impossible to use linear programming techniques such as the simplex method to solve Problem I with the strict inequality in (3). In practice, some have relaxed this constraint to include the equality. Technically, doing so introduces some ambiguity into the classification scheme, as it is not clear into which group those observations with

a classification score of exactly c should be classified. Others have avoided this problem by using weak inequalities, while adding a small value ε to the right-hand side c in (3). This formulation, however, has a different drawback in that it requires additional analysis in order to determine the classification of those observations with classification scores between c and $c+\varepsilon$. Since constraint sets (2) and (3) in Problem I are expressed as inequalities, and the d_{i1}^- and d_{i2}^+ do not contribute to the objective function in (1), these terms can be omitted from the formulation. These last two remarks apply to Problem III as well.

Suppose the optimal solution to Problem I is given by $d_{ir}^{+}{}^*$, $d_{ir}^{-}{}^*$, x_j^*, c^*, so that the optimal classification rule for which z_1 is minimized is to classify an observation $\boldsymbol{A}_i=(a_{i1},\ldots,a_{ik})^{\mathrm{T}}$ into group 1 if $\sum_j a_{ij}x_j^* \le c^*$, and into group 2 otherwise. The hyperplane defined by $\sum_j a_{ij}x_j^* = c^*$ divides the $\Re^k$ space into two half-spaces, and is called the classification function or separating hyperplane.

The objective in the MSD formulation is to minimize the sum of the absolute values of all undesirable deviations from the separating hyperplane, i.e., the extent to which values from one group intrude into the half-space of the other. One rationale for using the absolute value objective in MSD, rather than a least squares criterion, is that the former is less sensitive to the presence of outlier observations. However, in spite of its appeal, Problem I suffers from several potential problems. One problem is that even though the normalization constraint in (4) guarantees that the trivial solution is not selected as the optimal solution, it also precludes any classification rule from passing through the origin. In our simulation experiment below, this never presented a problem due to the nature of the data generated. Another problem is that it may be necessary to solve Problem I twice, with the groups reversed. For a detailed discussion of the problems associated with various MSD formulations, the reader is referred to Koehler (1989, 1990).

The MSD formulation only considers undesirable deviations from the separating hyperplane of the misclassified observations. Such deviations are called *external* deviations, as opposed to desirable *internal* deviations, which are the deviations from the hyperplane for the correctly classified observations. Several researchers have proposed models which simultaneously consider both the external (to be minimized) and internal (to be maximized) deviations (Bajgier and Hill, 1982; Freed and Glover, 1986; Glover, Keene and Duea, 1988; Glover, 1990). One of these models, the Hybrid formulation, originally proposed by Freed and Glover (1986), has been found to be one of the most effective MP-based classifiers. The Hybrid model also considers maximum distances from the separating hyperplane (see d_0 and e_0 below) and can be represented as follows:

(Problem II)

$$\text{Minimize} \quad z_2 = w_0 d_0 + \sum_{i\in G_1} w_i d_{i1}^+ + \sum_{i\in G_2} w_i d_{i2}^- - k_0 e_0 - \sum_{i\in G_1} k_i d_{i1}^- - \sum_{i\in G_2} k_i d_{i2}^+ \tag{7}$$

$$\text{subject to} \quad \sum_j a_{ij}x_j - d_0 - d_{i1}^+ + e_0 + d_{i1}^- = c, \quad i=1,\ldots,n_1, \tag{8}$$

$$\sum_j a_{ij}x_j + d_0 - d_{i2}^+ - e_0 + d_{i2}^- = c, \quad i=1,\ldots,n_2, \tag{9}$$

$$\left(-n_2 \sum_{i\in G_1} \boldsymbol{A}_i^{\mathrm{T}} + n_1 \sum_{i\in G_2} \boldsymbol{A}_i^{\mathrm{T}}\right)\boldsymbol{x} = 1, \tag{10}$$

$$d_0, e_0, d_{ir}^+, d_{ir}^- \ge 0, \quad i=1,\ldots,n_r, \quad r=1,2, \tag{11}$$

$$x_j, c \text{ unrestricted}, \tag{12}$$

where d_{i1}^+, d_{i2}^- represent external deviational variables, and d_{i1}^-, d_{i2}^+ are internal deviations. In addition to the absolute values of the deviations, variables which represent the maximum external deviation (d_0) and the maximum internal deviation (e_0) across all observations if the other deviational variables equal zero are included in the Hybrid formulation. The variable c is the cut-off value. The coefficients w_0, w_i, k_0 and k_i in the objective (7) are nonnegative relative weights. Glover (1990) indicates that necessary

conditions for bounded optimality and (certain) nonzero coefficients require $w_m \geq k_m$ for all m, $\Sigma_{i \in G_1 \cup G_2} w_i > k_0$ and $\Sigma_{i \in G_1 \cup G_2} k_i < w_0$. Glover (1990) has shown that the originally proposed normalization schemes in Freed and Glover (1986) and Glover, Keene and Duea (1988) are flawed, and proposes (10) as a normalization constraint. In our simulation study below we have found this normalization scheme to work quite well. As before, $\boldsymbol{x} = (x_1, \ldots, x_k)^T$ is the vector of weights (to be estimated) associated with the k attributes, and $\boldsymbol{A}_i = (a_{i1}, \ldots, a_{ik})^T$ denotes the vector of attribute values for observation i. Glover, Keene and Duea (1988) and Glover (1990) have proven that the above Hybrid formulation is guaranteed to yield a bounded optimal solution minimizing (7), regardless of the nature of the data, and that this solution is invariant to linear transformations of the data.

3. Second order formulations

We next introduce the modified formulations of the MSD and Hybrid models, including the second order terms. A similar modification to other existing MP models, such as the mixed-integer programming (MIP) approach which directly minimizes the number of misclassifications in the training sample and the minimize the maximum deviation (MMD) method, is possible as well, but will not be discussed in this paper. Denoting the vector of attribute values for observation i by $\boldsymbol{A}_i = (a_{i1}, \ldots, a_{ik})^T$ as before, the vector of squared attribute values is given by $\boldsymbol{A}_i^2 = (a_{i1}^2, \ldots, a_{ik}^2)^T$, and the cross product of the values for attributes h and m ($h \neq m$) is given by $a_{ih} a_{im}$. The second order MSD formulation is given in Problem III.

(Problem III)

$$\text{Minimize} \quad z_1 = \sum_{i \in G_1} d_{i1}^+ + \sum_{i \in G_2} d_{i2}^- \tag{1}$$

$$\text{subject to} \quad \sum_j a_{ij} x_{j\text{L}} + \sum_j a_{ij}^2 x_{j\text{Q}} + \sum_{h \neq m} a_{ih} a_{im} x_{hm} + d_{i1}^- - d_{i1}^+ \leq c, \quad i = 1, \ldots, n_1, \tag{13}$$

$$\sum_j a_{ij} x_{j\text{L}} + \sum_j a_{ij}^2 x_{j\text{Q}} + \sum_{h \neq m} a_{ih} a_{im} x_{hm} + d_{i2}^- - d_{i2}^+ > c, \quad i = 1, \ldots, n_2, \tag{14}$$

$$d_{ir}^+, d_{ir}^- \geq 0, \quad i = 1, \ldots, n_r, \quad r = 1,2, \tag{5}$$

$$\sum_{j; h \neq m} (x_{j\text{L}} + x_{j\text{Q}} + x_{hm}) + c = 1, \tag{15}$$

$$x_{j\text{L}}, x_{j\text{Q}}, x_{hm}, c \text{ unrestricted}, \quad h, j, m = 1, \ldots, k, \; h \neq m, \tag{16}$$

where the general terminology is the same as before. The attribute weights to be estimated are now defined by $x_{j\text{L}}$ and $x_{j\text{Q}}$ for the linear and quadratic terms of attribute j, respectively, and by x_{hm} for the cross-product term involving attributes h and m. Indicating the solution which minimizes z_1 in (1) of Problem III by a superscript star, the optimal classification rule associated with Problem III is to classify an observation i into group 1 if $\Sigma_j a_{ij} x_{j\text{L}}^* + \Sigma_j a_{ij}^2 x_{j\text{Q}}^* + \Sigma_{h \neq m} a_{ih} a_{im} x_{hm}^* \leq c^*$, and into group 2 otherwise. It is important to realize that, even though the second order MSD model in (III) is still *linear* in terms of the objective z_1 and still has *linear* constraints in terms of the parameters to be estimated ($x_{..}^*$), the constraints are *nonlinear* in terms of the attribute variables $a_{.j}$, $j = 1, \ldots, k$. We also remark once more that, while we will not analyze the second order MIP formulation in our paper, this formulation is straighforward, and would involve modifying merely equation (1) in Problem III, including zero-one binary variables which indicate either misclassification or correct classification of each observation, rather than distances from the estimated separating hyperplane.

Similarly, the second order Hybrid formulation is given in Problem IV.

(Problem IV)

$$\text{Minimize} \quad z_2 = w_0 d_0 + \sum_{i \in G_1} w_i d_{i1}^+ + \sum_{i \in G_2} w_i d_{i2}^- - k_0 e_0 - \sum_{i \in G_1} k_i d_{i1}^- - \sum_{i \in G_2} k_i d_{i2}^+ \tag{7}$$

$$\text{subject to} \quad \sum_j a_{ij} x_{j\mathrm{L}} + \sum_j a_{ij}^2 x_{j\mathrm{Q}} + \sum_{h \neq m} a_{ih} a_{im} x_{hm} - d_0 - d_{i1}^+ + e_0 + d_{i1}^- = c, \quad i = 1, \ldots, n_1, \tag{17}$$

$$\sum_j a_{ij} x_{j\mathrm{L}} + \sum_j a_{ij}^2 x_{j\mathrm{Q}} + \sum_{h \neq m} a_{ih} a_{im} x_{hm} + d_0 - d_{i2}^+ - e_0 + d_{i2}^- = c, \quad i = 1, \ldots, n_2, \tag{18}$$

$$\left(-n_2 \sum_{i \in G_1} \boldsymbol{B}_i^{\mathrm{T}} + n_1 \sum_{i \in G_2} \boldsymbol{B}_i^{\mathrm{T}} \right) \boldsymbol{y} = 1, \tag{19}$$

$$d_0, e_0, d_{ir}^+, d_{ir}^- \geq 0, \quad i = 1, \ldots, n_r, \quad r = 1,2, \tag{11}$$

$$x_{j\mathrm{L}}, x_{j\mathrm{Q}}, x_{hm}, c \text{ unrestricted}, \tag{20}$$

where $\boldsymbol{B}$ is the matrix of attribute values $\{a_{ij}\}$ augmented by the squared and cross-product terms, such that the i-th row of $\boldsymbol{B}$ is of the form $\boldsymbol{B}_i = (a_{i1}, \ldots, a_{ik}, a_{i1}^2, \ldots, a_{ik}^2, a_{i1}a_{i2}, \ldots, a_{i,k-1}a_{ik})$, and $\boldsymbol{y}^{\mathrm{T}} = (x_{1\mathrm{L}}, \ldots, x_{k\mathrm{L}}, x_{1\mathrm{Q}}, \ldots, x_{k\mathrm{Q}}, x_{12}, \ldots, x_{1k}, x_{23}, \ldots, x_{k-1,k})$ is the vector of all weights to be estimated. The restrictions on the objective function weights are the same as for Problem II. Note that the objective function (7) of Problem IV is the same as that of the first order Hybrid formulation of Problem II (z_2). The optimal classification rule for Problem IV has the same form as that of Problem III, and includes both linear, quadratic and cross-product terms of the attribute variables.

One obvious potential advantage of the second order formulation over previous linear MP methods and Fisher's linear classification method is that it facilitates the analysis of discriminant problems requiring a nonlinear classification function. An advantage of the second order method over quadratic statistical methods, such as Smith's quadratic method, is that the optimization criteria in Problems III and IV are still based on minimizing the sum of absolute deviations. Hence, in the second order formulations the robustness of the solution with respect to outlier observations, which is one of the attractive properties of previously proposed MP methods, is preserved. The following section reports on three simulation experiments designed to compare the classification performance of the second order formulations with other existing discriminant methods.

4. Simulation experiment

4.1. Design

We conduct three simulation experiments with different data conditions to verify the classification performance of the second order MP formulations. All three data conditions were based on linear transformations of exponentially distributed random variates. Our rationale for selecting this kind of data condition was that the data are skewed to the right, with the mean value greater than the median, representing a substantial violation of the multivariate normality assumption. Of course, if the data closely follow a multivariate normal distribution, either the LDF or QDF would be the appropriate discriminant technique.

It can be shown that for the first data condition selected (see Experiment A below) the optimal Bayesian classification rule is linear, so that the first order (linear) methods should perform well, while in the second and third experiments (B and C) the classification rule of choice is nonlinear, so that the second order MP formulations (and the QDF) should perform better than linear classifiers. In addition, while in the first and second data condition the attribute variables generated are independent, the variates used in the third experiment were highly correlated. In each experiment the training samples

Table 1
Distributions and parameter values for the simulation experiments

Experiment	Distribution [a]	Mean		Standard deviation	
		Group 1	Group 2	Group 1	Group 2
A	Exponential, $\lambda = 1$	1.0	2.0	1.0	2.0
B	Exponential, $\lambda = 1$	10.0	11.0	1.0	3.0
C	Exponential, $\lambda = 1$	10.0	11.0	1.0 [b]	3.0 [b]

[a] The data sets are linear transformations of variables following this distribution.
[b] The attributes for this experiment are correlated. The values reported here represent the conditional standard deviation of $a_{i,j} | a_{i,1}, \ldots, a_{i,j-1}$ $(j = 1, \ldots, k-1)$. The variance–covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are given by (21) and (22) in the text.

consist of $n_1 = 100$ observations in group 1 and $n_2 = 100$ in group 2, for a total of $n = 200$ observations in both groups combined. The validation sample size was 500 observations for each group. Forty replications were generated and analyzed for each data condition. In each experiment, observations were generated on five attributes. The simulation experiments were performed using the SAS statistical package (SAS, 1982) on an IBM 3090 mainframe computer. A summary of the data conditions in the three simulation experiments is provided in Table 1.

In Experiment A the values on the five attributes were generated as linear transformations of independent, identically exponentially distributed random variates. After applying a linear transformation $A_{ij} = Y_{ij}^i$ for $i \in G_1$ and $A_{ij} = 2Y_{ij}^2$ for $i \in G_2$ $(j = 1, \ldots, k)$, where Y_{ij}^1 and Y_{ij}^2 follow the exponential with $\lambda = 1$, the mean vectors were $\boldsymbol{\mu}_1 = (1.0, \ldots, 1.0)^T$ and $\boldsymbol{\mu}_2 = (2.0, \ldots, 2.0)^T$, and the variance–covariance matrices $\boldsymbol{\Sigma}_1 = 1.0\boldsymbol{I}$ and $\boldsymbol{\Sigma}_2 = 4.0\boldsymbol{I}$, where $\boldsymbol{I}$ is the 5×5 identity matrix. Using Bayes' rule, it can be shown that the optimal classification rule for this data condition is linear. A typical scatter diagram of two variates generated in Experiment A is shown in Figure 2.

The data in Experiment B were also generated from an exponential distribution with parameter $\lambda = 1$. As in Experiment A, the variates in Experiment B were independent and identically distributed. Let Y_{ij}^1 and Y_{ij}^2 be defined as in Experiment A. The value on attribute j of observation $i \in G_1$ was calculated as $A_{ij} = 9 + Y_{ij}^1$ $(j = 1, \ldots, k)$. Similarly, the observations in group 2 were determined by the transformation $A_{ij} = 8 + 3Y_{ij}^2$ $(j = 1, \ldots, k)$, yielding mean vectors for the attribute values in groups 1 and 2 of $\boldsymbol{\mu}_1 = (10.0, \ldots, 10.0)^T$ and $\boldsymbol{\mu}_2 = (11.0, \ldots, 11.0)^T$, respectively, with variance–covariance matrices equal to $\boldsymbol{\Sigma}_1 = 1.0\boldsymbol{I}$ and $\boldsymbol{\Sigma}_2 = 9.0\boldsymbol{I}$. In contrast to Experiment A, in Experiment B the domain of group 1 lies in the

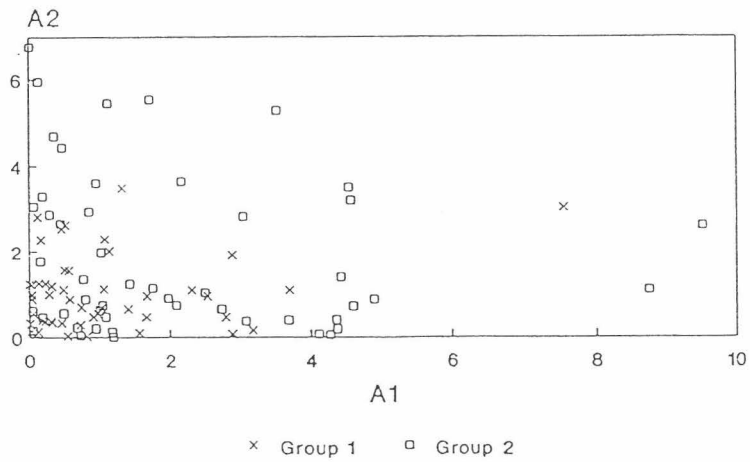

Figure 2. Typical scatter diagram for two out of five attributes of training sample data set for Experiment A: Linear transformation of exponential distribution ($\lambda = 1$), $\mu_1 = 1$, $\mu_2 = 2$, $\sigma_1 = 1$, $\sigma_2 = 2$, independent attributes

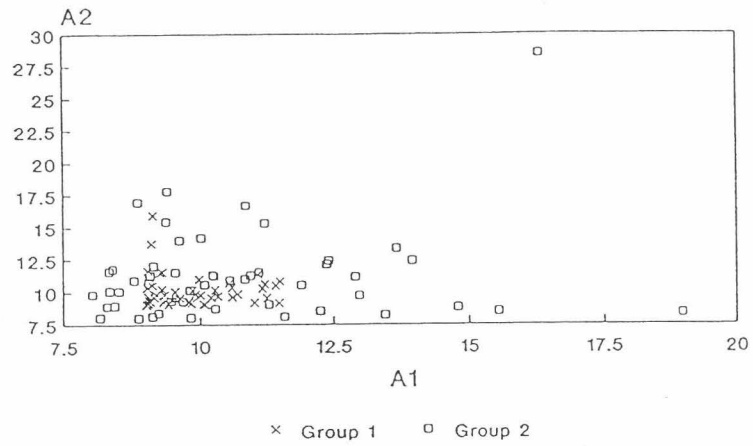

Figure 3. Typical scatter diagram for two out of five attributes of training sample data set for Experiment B: Linear transformation of exponential distribution ($\lambda = 1$), $\mu_1 = 10$, $\mu_2 = 11$, $\sigma_1 = 1$, $\sigma_2 = 3$, independent attributes

interior of the domain of group 2. The optimal classification rule for this data condition can be shown to be nonlinear. Figure 3 shows a representative plot of two attributes A_1 and A_2 from the data condition analyzed in Experiment B.

Finally, in Experiment C the attribute values within each group were generated to be correlated, in the following way. Let Y_{i1}^1 and Y_{i1}^2 again be exponentially distributed with parameter $\lambda = 1$. The values on the first attribute were taken equal to $A_{i1} = 9 + Y_{i1}^1$ for all $i \in G_1$, and $A_{i1} = 8 + 3Y_{i1}^2$ for all $i \in G_2$. The values of the remaining $j = 2, \ldots, 5$ attributes were given by $A_{ij} = A_{i,j-1} - 1 + Y_{ij}^1$ for all observations $i \in G_1$ and $A_{i,j} = A_{i,j-1} - 3 + 3Y_{ij}^2$ for all $i \in G_2$, so that the respective mean vectors are $\boldsymbol{\mu}_1 = (10.0, \ldots, 10.0)^T$ and $\boldsymbol{\mu}_2 = (11.0, \ldots, 11.0)^T$. The variance–covariance matrices $\boldsymbol{\Sigma}_1$ for group 1 and $\boldsymbol{\Sigma}_2$ for group 2 are defined by (21) and (22).

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1.0 & 1.0 & 1.0 & 1.0 & 1.0 \\ & 2.0 & 2.0 & 2.0 & 2.0 \\ & & 3.0 & 3.0 & 3.0 \\ & & & 4.0 & 4.0 \\ & & & & 5.0 \end{bmatrix}, \tag{21}$$

$$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 9.0 & 9.0 & 9.0 & 9.0 & 9.0 \\ & 18.0 & 18.0 & 18.0 & 18.0 \\ & & 27.0 & 27.0 & 27.0 \\ & & & 36.0 & 36.0 \\ & & & & 45.0 \end{bmatrix}. \tag{22}$$

The within-group correlation structure of the attributes in Experiment C is the same for both groups, and is given by matrix $\boldsymbol{R}$ in (23), indicating that the five attributes within each group are highly correlated, with correlation coefficients between 0.447 and 0.894.

$$\boldsymbol{R} = \begin{bmatrix} 1.000 & 0.707 & 0.577 & 0.500 & 0.447 \\ & 1.000 & 0.816 & 0.707 & 0.632 \\ & & 1.000 & 0.866 & 0.775 \\ & & & 1.000 & 0.894 \\ & & & & 1.000 \end{bmatrix}. \tag{23}$$

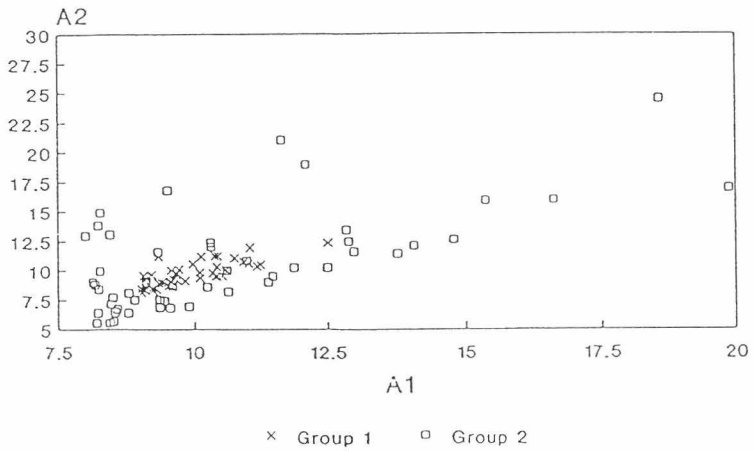

Figure 4. Typical scatter diagram for two out of five attributes of training sample data set for Experiment C: Linear transformation of exponential distribution ($\lambda = 1$), $\mu_1 = 10$, $\mu_2 = 11$, correlated attributes

The data condition in Experiment C with correlated attribute variables is of special interest, because it can be used to test the relevance of the cross-product terms in the second order MP formulations. A typical scatter plot of two variables from the data condition in Experiment C is given in Figure 4.

Ten different discriminant methods were compared based on their classification performance, as measured by the percentage of misclassified observations. The linear classification methods under evaluation included Fisher's LDF, the MSD and the Hybrid formulation. Smith's quadratic discriminant method was included in the analysis, since the optimal Bayesian classification rule is nonlinear for two out of the three simulation experiments. Following an approach often used in statistics, Smith's quadratic method was applied not only to the original data (QDF), but also to the natural logarithms (QDFL) and the square roots (QDFS) of the original data. The second order methods evaluated in this study consist of the MSD formulation with second order terms *including* cross-products (MSDQ1) and *excluding* cross-products (MSDQ2), as well as the second order Hybrid method *with* (HYBQ1) and *without* (HYBQ2) cross-products. Including cross-product terms in the second order formulation is appropriate if the attributes are correlated. Analogous to the case of parametric statistical methods, for reasons of parsimony it may be preferable to omit the cross-product terms from the model if the attributes are uncorrelated. A complete list of the methods analyzed in our study is provided in Table 2.

As mentioned before when discussing the Hybrid formulations of Problems II and IV, there are several restrictions on the coefficients used in the objective function (6). The optimal classification rule will depend in part on the relative magnitude of these coefficients. In our simulation experiments, $w_0 = 1.5n = 300$, $w_i = 2$ (for all $i \in G_1 \cup G_2$), $k_0 = n = 200$, $k_i = 0.5$ (for all $i \in G_1 \cup G_2$) were selected. This choice of coefficient values is consistent with Glover's recommendations (Glover, 1990).

4.2. Results

The nature of the differences between the misclassification rates across the various methods is assessed by (a) pairwise T-tests of the mean difference between all combinations of methods, (b) an analysis of variance (ANOVA) to test for the significance of two factors, the classification method and the training sample used to fit the classification rule, and (c) Tukey's studentized range test, which simultaneously tests for differences between the misclassification rates across methods.

Table 2
Abbreviations and description of methods evaluated in the simulation experiments

Method	Description
LDF	Fisher's Linear Discriminant Function applied to original data
QDF	Smith's Quadratic Discriminant Function applied to original data
QDFL	Smith's Quadratic Discriminant Function applied after natural logarithmic transformation of the data
QDFS	Smith's Quadratic Discriminant Function applied after square root transformation of the data
MSD	MSD formulation with linear (first order) discriminant function applied to the original data
MSDQ1	MSD formulation with quadratic (second order) discriminant function (including cross-products) applied to the original data
MSDQ2	MSD formulation with quadratic (second order) discriminant function (excluding cross-products) applied to the original data
HYB	Hybrid formulation with linear (first order) discriminant function applied to the original data
HYBQ1	Hybrid formulation with quadratic (second order) discriminant function (including cross-products) applied to the original data
HYBQ2	Hybrid formulation with quadratic (second order) discriminant function (excluding cross-products) applied to the original data

It is well-known that, when performing several independent statistical tests at the same time, such as multiple paired T-tests, the probability of making a type I error α, i.e., erroneously rejecting the null hypothesis, may be different from (and is usually higher than) the specified significance level. Tukey's multiple comparison test overcomes this problem by testing several hypotheses simultaneously, assuring that the probability that none of the null hypotheses is rejected erroneously is at least $1-\alpha$. Hence, using Tukey's test one can safely infer several conclusions at the same time, providing a relevant means of verifying the results obtained using the paired T-tests. However, Tukey's method tends to be very conservative, and may not detect significant differences even though these are present in the data. Therefore, it is useful to report the results of both the paired T-test analysis and Tukey's studentized range tests for differences between the various classification methods.

The two-factor ANOVA model without interactions is given in (24). Note that since we have only one observation per cell, a model with interaction effects would result in zero degrees of freedom. Represent the misclassification proportion for method j and training sample i by P_{ij}. In order to stabilize the variances, we apply the following transformation to the original dependent variable P_{ij}: $P'_{ij} = 2 \arcsin(\sqrt{P_{ij}})$ (see Joachimsthaler and Stam, 1988; Neter, Wasserman and Kutner, 1985).

$$P'_{ij} = U + T_i + M_j + e_{ij}, \tag{24}$$

where U is the overall mean of P'_{ij}, T_i is the effect of training sample i, M_j is the effect of classification method j, and the e_{ij} are independent normally distributed error terms with a constant variance.

Experiment A. The mean misclassification rates for Experiment A are presented in Table 3. The paired T-values for this experiment are given in Tables 4–5. A positive t-value in position (i, j) of the table indicates that the mean misclassification rate of method i is higher than that of method j, while a negative t-value implies that the reverse is true. Tukey's multiple comparison test results for Experiment A are reported in Table 6, while the two-factor ANOVA results for all three experiments are summarized in Table 7. Since the optimal classification rule for this experiment as determined using Bayes' rule is linear, it is not surprising that the mean misclassification results for Experiment A in Table 3 indicate that the first order MSD method (24.9%) and the first order hybrid method HYB (25.2%) have the lowest misclassification rate in the validation sample analysis. The t-values in Table 5 suggest that the performance of MSD in the validation sample analysis is significantly better than that of HYB and QDFS (significant at the one percent level), but Tukey's multiple comparison test groups these three methods together (see Table 6).

In the training sample analysis, the three MSD formulations clearly give the best classification results, MSDQ1 yielding a slightly better misclassification rate (18.4%) than the MSD and MSDQ2 methods

Table 3
Classification results, simulation Experiment A

Method	Percentage of misclassified observations			
	Training sample		Validation sample	
	Mean	STD	Mean	STD
LDF	27.4	4.2	31.1	2.5
QDF	22.8	3.3	26.6	1.2
QDFL	25.3	3.6	29.4	1.6
QDFS	20.8	3.3	25.4	1.0
MSD	20.1	2.6	24.9	0.8
MSDQ1	18.4	3.1	28.9	1.4
MSDQ2	20.1	3.2	26.2	1.1
HYB	21.3	5.5	25.2	1.1
HYBQ1	22.0	8.6	30.3	3.0
HYBQ2	23.2	5.4	28.6	2.6

Table 4
Pairwise T-tests [a] of mean difference in classification performance, training sample, Experiment A

Method	Method								
	QDF	QDFL	QDFS	MSD	MSDQ1	MSDQ2	HYB	HYBQ1	HYBQ2
LDF	8.94	3.72	13.69	15.99	15.90	17.50	13.14	6.94	7.58
QDF		−3.91	7.32	7.09	9.39	5.52	3.90	1.24	−0.71
QDFL			8.97	11.99	13.16	11.99	9.08	4.58	3.58
QDFS				2.67	5.70	1.93	−1.47	−1.75	−5.48
MSD					4.77	−0.15	−5.33	−3.05	−7.51
MSDQ1						−4.25	−7.14	−5.70	−8.86
MSDQ2							−3.56	−3.02	−6.89
HYB								−1.04	−4.61
HYBQ1									−1.75

[a] A positive (negative) t-value in position (i, j) of the table indicates that the mean misclassification rate of method i is higher (lower) than that of method j. The critical values for $\alpha = 5\%$ and $\alpha = 1\%$ are 2.023 and 2.709, respectively.

(20.1%). The t-values associated with pairwise comparing MSDQ1 with each of the other methods in Table 4 are all significant at the one percent level. Tukey's test confirms that in this case the MSDQ1 method classifies significantly better than any of the other methods.

Table 5
Pairwise T-tests [a] of mean difference in classification performance, validation sample, Experiment A

Method	Method								
	QDF	QDFL	QDFS	MSD	MSDQ1	MSDQ2	HYB	HYBQ1	HYBQ2
LDF	11.68	4.63	15.25	16.62	4.71	12.63	14.52	1.82	6.45
QDF		−8.48	6.21	8.66	−8.36	1.50	6.57	−13.55	−6.73
QDFL			14.24	15.96	1.29	10.56	13.43	−2.41	1.86
QDFS				3.64	−13.99	−3.78	1.28	−15.58	−10.41
MSD					−17.06	−8.80	−3.31	−15.84	−11.49
MSDQ1						12.56	15.41	−3.83	0.93
MSDQ2							5.19	−11.68	−6.87
HYB								−15.23	−10.58
HYBQ1									5.33

[a] A positive (negative) t-value in position (i, j) of the table indicates that the mean misclassification rate of method i is higher (lower) than that of method j. The critical values for $\alpha = 5\%$ and $\alpha = 1\%$ are 2.023 and 2.709, respectively.

Table 6
Tukey's studentized range test for significant difference between methods, Experiment A

Training sample			Validation sample		
Method	Mean	Tukey grouping [a]	Method	Mean	Tukey grouping [a]
LDF	1.1000	A	LDF	1.1832	A
QDFL	1.0531	B	HYBQ1	1.1657	A B
HYBQ2	1.0018	C	QDFL	1.1447	C B
QDF	0.9944	C D	MSDQ1	1.1357	C
HYBQ1	0.9693	C D E	HYBQ2	1.1272	C
HYB	0.9567	F D E	QDF	1.0825	D
QDFS	0.9461	F E	MSDQ2	1.0748	D E
MSDQ2	0.9272	F	QDFS	1.0571	F E
MSD	0.9270	F	HYB	1.0525	F
MSDQ1	0.8831	G	MSD	1.0435	F

[a] Determined based on a significance level of $\alpha = 5\%$; the minimum significant difference in mean is 0.0386 for the training sample and 0.0220 for the validation sample.

Similarly, HYB classifies slightly better than the second order variants (HYBQ1 and HYBQ2). Compared to the other methods, the second order Hybrid methods do not perform very well on the data condition in Experiment A, suggesting that these methods in general may not be well-suited for this particular data condition. The LDF method also fares poorly, both in the training sample (27.4% misclassified, on average) and validation sample analysis (31.1% misclassified), which may be explained by the asymmetric nature of the exponential distribution used to generate the data for this experiment.

Not surprisingly, given the fact that the attributes are uncorrelated and the optimal classification rule is linear in Experiment A, the second order methods generally do not perform as well as the first order methods, especially on the validation samples. As indicated by the t-test scores and Tukey's studentized range values (Table 4 and 6), the second order MSD (MSDQ1 and MSDQ2) methods classify significantly better than their counterpart second order Hybrid formulations (HYBQ1 and HYBQ2). The performance of the QDF methods is very similar to that of the second order MSD and Hybrid methods.

Experiment B. From Tables 8–11 we see that, as anticipated, in Experiment B the linear methods LDF, MSD and HYB classify considerably worse than most of the statistical quadratic and second order methods on both the training and validation samples, with average misclassification rates of close to or over 30%. This finding is not surprising, given that a nonlinear classification function is appropriate for this data condition. It is interesting, nevertheless, to note the significant difference in classification performance between the linear and nonlinear classifiers in Experiment B (and below in Experiment C),

Table 7
Analysis of variance results for Experiments A, B and C

Experiment	Overall significance of the model			Training sample effect			Classification method effect		
	F-value	DF [a]	P-value	F-value	DF [a]	P-value	F-value	DF [a]	P-value
Training sample:									
A	24.17	48, 351	0.0000	16.85	39, 351	0.0001	55.92	9, 351	0.0001
B	103.17	48, 351	0.0000	10.37	39, 351	0.0001	505.30	9, 351	0.0000
C	145.60	46, 273	0.0000	1.58	39, 273	0.0206	948.02	7, 351	0.0000
Validation sample:									
A	23.65	48, 351	0.0000	4.14	39, 351	0.0001	108.18	9, 351	0.0000
B	175.33	48, 351	0.0000	1.74	39, 351	0.0054	927.55	9, 351	0.0000
C	295.32	46, 273	0.0000	1.10	39, 273	0.3297	1934.54	7, 273	0.0000

[a] DF = degrees of freedom (numerator, denominator).

Table 8
Classification results, simulation Experiment B

Method	Percentage of misclassified observations			
	Training sample		Validation sample	
	Mean	STD	Mean	STD
LDF	30.9	4.4	32.2	3.5
QDF	17.5	3.1	17.1	0.8
QDFL	14.8	3.0	14.8	0.9
QDFS	16.3	3.2	15.9	0.9
MSD	28.2	4.4	28.8	1.6
MSDQ1	15.7	4.4	24.6	2.7
MSDQ2	36.3	5.7	39.9	3.9
HYB	30.0	4.8	30.1	2.3
HYBQ1	8.2	4.2	14.2	3.2
HYBQ2	9.7	2.6	12.3	0.9

Table 9
Pairwise T-tests [a] of mean difference in classification performance, training sample, Experiment B

Method	Method								
	QDF	QDFL	QDFS	MSD	MSDQ1	MSDQ2	HYB	HYBQ1	HYBQ2
LDF	20.34	22.91	20.90	5.17	19.38	−7.97	1.69	29.31	30.32
QDF		9.89	7.37	−18.11	3.73	−23.32	−20.14	21.96	19.99
QDFL			−8.29	−21.59	−1.50	−25.31	−23.05	15.77	13.84
QDFS				−19.58	1.16	−24.64	−21.64	18.78	16.81
MSD					16.16	−18.02	−8.05	27.06	27.19
MSDQ1						−21.80	−18.17	14.40	11.20
MSDQ2							13.10	28.68	28.90
HYB								29.67	30.00
HYBQ1									−4.58

[a] A positive (negative) t-value in position (i, j) of the table indicates that the mean misclassification rate of method i is higher (lower) than that of method j. The critical values for $\alpha = 5\%$ and $\alpha = 1\%$ are 2.023 and 2.709, respectively.

Table 10
Pairwise T-test [a] of mean difference in classification performance, validation sample, Experiment B

Method	Method								
	QDF	QDFL	QDFS	MSD	MSDQ1	MSDQ2	HYB	HYBQ1	HYBQ2
LDF	25.80	30.35	28.31	5.74	11.67	−11.77	3.72	30.58	33.32
QDF		18.53	12.78	−38.26	−19.74	−33.58	−37.84	11.74	29.32
QDFL			−14.87	−47.16	−23.74	−36.72	−42.70	2.14	12.98
QDFS				−44.02	−21.24	−34.91	−40.34	6.23	19.11
MSD					8.70	−18.28	−5.58	36.39	51.76
MSDQ1						−21.12	−10.86	25.40	28.96
MSDQ2							16.00	38.90	44.43
HYB								35.86	48.57
HYBQ1									7.28

[a] A positive (negative) t-value in position (i, j) of the table indicates that the mean misclassification rate of method i is higher (lower) than that of method j. The critical values for $\alpha = 5\%$ and $\alpha = 1\%$ are 2.023 and 2.709, respectively.

because in several previous studies linear methods have been used as a benchmark for the performance of other linear classifiers, even under data conditions which clearly favor nonlinear functions. From our results it is clear that using linear classifiers for nonlinear data conditions similar to our study can lead to

SAS (1982), "*SAS user's-guide: Statistics*", SAS Institute, Carey, NC.

Smith, C.A.B. (1947), "Some examples of discrimination", *Annals of Eugenics* 13, 272–282.

Spiegelhalter, D.J., and Knill-Jones, R.P. (1984), "Statistical and knowledge-based approaches to clinical decision-support systems, with an application to gastroenterology", *Journal of the Royal Statistical Society Series A* 147, Part 1, 35–77.

Stam, A., and Joachimsthaler, E.A. (1989), "Solving the classification problem in discriminant analysis via Linear and Nonlinear Programming methods", *Decision Sciences* 20, 285–293.

Stam, A., and Joachimsthaler, E.A. (1990), "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, 113–122.

Stam, A., and Jones, D.G. (1990), "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. (1992), "On the classification gap in MP-based approaches to the discriminant problem", *Naval Research Logistics* 39, 545–559.