TWO-GROUP CLASSIFICATION IN BUSINESS: AN EVALUATION OF PARAMETRIC AND NON-PARAMETRIC APPROACHES

A. Pedro Duarte Silva[1]

Antonie Stam[2]

John Neter[3]

[1] Faculdade de Ciências Económicas e Empresariais
Universidade Católica Portuguesa - Centro Regional do Porto
4150 PORTO PORTUGAL

[2] Department of Management
Terry College of Business - The University of Georgia
ATHENS GA U.S.A.

[3] Faculty of Management Science
Terry College of Business - The University of Georgia
ATHENS GA U.S.A.

This study investigates the relative performance, for the two-group case, of a number of classification methods under various different data conditions that are common in business and economic applications. Often, business classification problems are characterized by skewed attribute distributions and unequal misclassification costs across groups (Altman, Haldeman and Narayanan 1977; Rudolph and Karson 1988). In particular financial, accounting and demographic variables tend to have distributions that are highly skewed to the right (Altman, Haldeman and Narayanan 1977; Altman, Avery, Eisenbeis and Sinkey 1980; Eisenbeis 1977; Johnson, Leitch and Neter 1981; Gibbons, Dianne, McDonald and Gunst 1987).

Whereas classification rules with optimal properties for discriminant problems with multivariate normally distributed attribute variables are well-known (Wald 1944, 1949; Smith 1947), alternative rules may be more appropriate if some of the attributes are skewed. Most of the studies that compared non-normal classification methods with normality-based methods for various different data conditions have assumed equal misclassification costs across groups. Hence, it is not clear to what extent the conclusions in these studies can be generalized to typical business problems with

distributions that are skewed to the right and with unequal misclassification costs across groups.

The purpose of the current study is to establish guidelines for choosing an appropriate classification method if the problem at hand is characterized by the non-normal data conditions described above. To achieve this objective, several Monte Carlo simulation experiments are conducted to compare the performance of a number of well-known traditional classification methods with several nontraditional methods designed specifically to handle problems with skewed distributions. This study is limited to the two-group classification problem. A generalization of the results to problems involving more than two groups is an issue that needs to be addressed in future research.

REVIEW OF CLASSIFICATION METHODS

Suppose that a set of observations belonging to one of two mutually exclusive groups is described by a set of *p*-dimensional attribute vectors. Denote the attribute vector of observation *i* by $\mathbf{x}_i$, membership in group *j* by G_j, the probability or probability density of $\mathbf{x}_i$ given membership in G_j by $p(\mathbf{x}_i|G_j)$, the prior probabilities of membership in G_j by π_j, and the cost of misclassifying an observation belonging to G_j by C_j (j=1,2). Then, the Bayes rule that minimizes the expected cost of misclassification assigns observation *i* to the group G_j for which $C_j*\pi_j*p(\mathbf{x}_i|G_j)$ is maximized (Wald 1939, 1949). Parametric classification methods assume that the $p(\mathbf{x}_i|G_j)$ follow known probability distributions that can be fully described by a small set of parameters, and estimate these parameters from a training sample. For example, if the $p(\mathbf{x}_i|G_j)$ are assumed to be multivariate normally distributed with different mean vectors but equal covariance matrices, the parametric approach implies a linear classification rule based on Fisher's Linear Discriminant Function (LDF) (Fisher 1936), in which a classification score is compared with a threshold value that depends on the π_j and the ratio of group-wise misclassification costs (Wald 1944). If the $p(\mathbf{x}_i|Gj)$ are assumed to be multivariate normally distributed with heterogeneous covariance matrices across groups, the parametric approach implies a similar rule that replaces the LDF by a quadratic function, Smith's Quadratic Discriminant Function (QDF) (Smith 1947).

If the attribute distributions are clearly non-normal, the usual approach is to apply data transformations that reduce the deviations from normality, or to use methods that do not make strong distributional assumptions. Although the use of data transformations in discriminant analysis has been criticized by some authors (*e.g.*, Eisenbeis 1977), because transformations may hide the interrelationships of the original attributes, their use may be considered legitimate, as long as the purpose of the analysis is strictly classification, rather than description (McLachlan 1992).

The most important alternatives to the LDF and QDF are logistic regression methods (Anderson 1972; McCullagh and Nelder 1989), and methods based on kernel (Breiman, Meisel and Purcell 1977; Hand 1982; Parzen 1962) and *k*-nearest neighbor (Agrawala 1977; Fix and Hodges 1951; McLachlan 1992) estimators of the $p(\mathbf{x}_i|G_j)$. Generally, *k*-Nearest neighbor methods are simpler than kernel methods, and have the advantage that they are adaptable to the amount of information available to estimate the $p(\mathbf{x}_i|G_j)$. Logistic regression methods estimate the posterior group membership probabilities

$p(G_j|\mathbf{x}_i)$ used in the Bayes rule directly, without the intermediate step of estimating the $p(\mathbf{x}_i|G_j)$.

A different approach to two-group classification is to estimate the boundaries of the region of the attribute domain for which $C_1*p(G_1|\mathbf{x}_i)>C_2*p(G_2|\mathbf{x}_i)$ directly, without making any assumptions about the attribute distributions. This approach assumes that these boundaries can be described by the equation $f(\mathbf{b},\mathbf{x}_i)=c$**,** where $\mathbf{b}=(b_1,b_2,...,b_t)^T$ is a vector of unknown parameters, c is a threshold value and the functional form of $f(\mathbf{b},\mathbf{x}_i)$ is known *a priori*. The parameters **b,** and in some cases c, are estimated such that some training sample accuracy criterion is optimized. This criterion typically seeks to establish a classification rule that is not affected disproportionately by, "extreme" training sample observations. As the optimization of models with these types of criteria can be done in a straightforward manner using mathematical programming (MP) techniques, this approach is often referred to as the MP approach to classification.

Most criteria proposed in the MP approach belong to one of the following two classes, 1) L_1-norm distance criteria, which are based on some function of misclassification cost and the absolute deviations of the training sample observations from the surface $f(\mathbf{b},\mathbf{x}_i)=c$ that separates the two groups, and 2) L_0-norm criteria, which are based on the number (proportion) of misclassified observations or total misclassification cost in the training sample. For a detailed discussion of criteria proposed in the MP approach to two-group classification, see Erenguc and Koehler (1990) and Joachimsthaler and Stam (1990).

Among the vast literature that evaluates methods for two-group classification with equal misclassification costs across groups, there is a consensus that the LDF and QDF tend to be the most accurate methods if the $p(\mathbf{x}_i|G_j)$ are approximately multivariate normally distributed (Efron 1975; Fatti, Hawkins and Raath 1982; Lachenbruch, Sneeringer and Revo 1973; Murphy and Moran 1986). Whereas the LDF tends to perform the best if the group-wise covariance matrices are similar, the case of unequal covariance matrices implies a trade-off between estimating a rule with the same functional form as the Bayes rule (QDF) and one for which the parameters can be estimated more efficiently (LDF). The choice of classification rule in this situation depends on the extent of the covariance heterogeneity across groups, on how many parameters are to be estimated in the QDF, and on how many observations are available to estimate these parameters (Marks and Dunn 1974).

Although robust with respect to slight or moderate deviations from normality, the performance of the LDF and QDF deteriorates substantially if the $p(\mathbf{x}_i|G_j)$ are highly skewed (Clarke, Lachenbruch and Broffit 1979: Fatti, Hawkins and Raath 1982; Lachenbruch, Sneeringer and Revo 1973; Rawlings, Faden, Graubard and Eckardt 1986). If the attributes are highly skewed, methods based on weaker assumptions, particularly logistic regression, have been found to yield better results than the LDF and QDF (Byth and McLachlan 1980; Press and Wilson 1978). While numerous studies have evaluated nonparametric classification methods for various data conditions, the conclusions reported in these studies are difficult to generalize, given the large number of such methods. For instance, there is strong evidence that certain nonparametric methods tend to outperform the LDF and QDF if the data is skewed and the training samples are large. However, the results for these same methods are not so clear and highly variable if the training samples are small, and depend strongly on the particular data conditions analyzed and the choice of estimator for $p(\mathbf{x}_i|G_j)$. Remme, Habbema and Hermans (1980) found

that for small samples and skewed distributions the nearest neighbor and kernel methods perform about equally well as the LDF and QDF. However, the conclusions for small training samples are mixed, even if the attributes are multivariate normally distributed (Gessaman and Gessaman 1972; Murphy and Moran 1986; Van Ness 1979).
A number of studies have found that, although inferior to the LDF and QDF if the attributes are multivariate normally distributed, MP-based methods fare much better if some of the attributes are highly skewed (Duarte Silva and Stam 1994; Glorfeld and Olson 1982; Joachimsthaler and Stam 1990; Rubin 1990; Srinivasan and Kim 1987; Stam and Joachimsthaler 1990). Koehler and Erenguc (1990) and Stam and Jones (1990) remark that MP-based methods that use an L_0-norm criterion, without a secondary criterion to resolve ties in the total training sample misclassification cost, appear to be very sensitive to the training sample size, with an often erratic behavior for problems with small training samples, but that their relative accuracy improves significantly as the training sample size increases.
Summarizing, the general conclusion in the literature is that methods based on normal theory (LDF and QDF) usually yield the best classification results if the attribute distributions are approximately multivariate normally distributed, but tend to be inferior to non-normal methods if the deviations from normality are substantial, *e.g.*, if the attribute distributions are highly skewed.

METHODOLOGY

Objectives of the Current Study

This study focuses on three questions that have not been addressed fully in the classification analysis literature, but are highly relevant in practice, particularly in the case of business and economics applications: 1) what is the relative classification accuracy of various nonparametric methods if the attribute distributions are skewed, and which factors should guide the choice of nonparametric classification method; 2) is it possible to generalize the conclusions drawn from previous studies that have assumed equal misclassification costs across groups to problems involving unequal misclassification costs; and 3) how do data transformations aimed at improving the classification accuracy of parametric methods affect the performance of nonparametric methods, and what are the implications of using data transformations in terms of selecting an appropriate classification method.

Classification Methods

Table 1 summarizes the eight classification methods included in this study, the LDF, QDF, linear logistic regression with first order term predictors (LGR), a nearest neighbor method (NN), and four L_0- and L_1-norm MP-based methods for two-group classification, each with a linear and a quadratic classification function (L0L, L0Q, L1L, L1Q).
The LDF, QDF and LGR are the most widely used two-group classification methods. The NN method is a representative nonparametric method for estimating $p(\mathbf{x}_i|G_j)$. Due to their conceptual similarity to the NN method, general kernel methods were not included in this study, because these methods require several subjective choices, such as the choice of kernel function and smoothing parameters, complicating the generalization of their classification performance in a given experiment. The NN distance norm used in this

study is defined by $d(\mathbf{x}_i,\mathbf{x}_j)= (\mathbf{x}_i-\mathbf{x}_j)^T S_p^{-1}(\mathbf{x}_i-\mathbf{x}_j)$, where S_p is the pooled training sample attribute covariance matrix. Following recommendations by Enas and Choi (1986), the number of neighbors in the NN is determined as a function of the training sample size N and the variance heterogeneity across groups. In the case of equal covariances across groups, the number of neighbors equals the odd integer closest to $N^{3/8}$, and otherwise the odd integer closest to $N^{2/8}$.

TABLE 1: CLASSIFICATION METHODS INCLUDED IN THE EXPERIMENTS (*CM* FACTOR)

Acronym	Method (*CM* Factor)
LDF	Fisher's Linear Discriminant Function.
QDF	Smith's Quadratic Discriminant Function.
LGR	Logistic regression, using first order term predictors.
NN	Nearest neighbor method, with the number of neighbors equal to either the odd integer closest to $N^{2/8}$ (unequal variance-covariances) or $N^{3/8}$ (equal variance-covariances), and a Mahalanobis norm based on the sample pooled variance-covariance matrix.
L0L	MP model which minimizes the training sample misclassification cost, using a linear classification rule.
L0Q	MP model which minimizes the training sample misclassification cost, using a quadratic classification rule.
L1L	MP model which minimizes an objective based on the sum of absolute deviations from the threshold value multiplied by the misclassification costs, for all misclassified observations combined, using a linear classification rule.
L1Q	MP model which minimizes an objective based on the sum of absolute deviations from the threshold value multiplied by the misclassification costs, for all misclassified observations combined, using a quadratic classification rule.

The four MP-based methods are included in our study because these methods are designed specifically to handle distributions with extreme values. Some details of the MP-based methods used in this study are reviewed in Appendix A. Following the decision-theoretic tradition (Wald 1949), the LDF, QDF, LGR and NN classification rules incorporate the group-wise different misclassification costs (C_1 and C_2) by adjusting the respective threshold values with $\ln(C_1/C_2)$. The MP-based rules incorporate these costs by including proportional weights in the criterion components.

Factors in the Experimental Design

This study uses several Monte Carlo simulation experiments.Through the experiments, the prior group membership probabilities are assumed to be equal ($\pi_1=\pi_2=0.5$), and all training samples in the experiments are balanced.

The *primary* experiment evaluates the relative performance of the eight different classification methods (*CM*) for three factors, level of skewness (*SK*), group-wise ratio of misclassification costs (*RC*) and ratio of attribute variances across groups (*RV*), and analyzes the interactions between *SK*, *RC* and *CM*. Skewness is included as a factor because, as noted above, many variables used in business and economic classification problems are skewed to the right. The group-wise ratio of misclassification costs is included, because because in economic and business studies these costs are usually different. The group-wise ratio of variances is included because this factor is widely recognized as playing a critical role in the relative performance of different classification methods (Marks and Dunn 1974; Clarke, Lachenbruch and Broffit 1979; Remme, Habbema and Hermans 1980; McLachlan 1992).

A series of *secondary* experiments analyzes several additional factors not included in the primary experiment: the use of a data transformation (*DT*), the number of attributes (*P*), the relative training sample size (*TS*), the degree of group-overlap (*OVLP*) and the correlation structure of the attributes (*CORR*). As explained in more detail below, in the secondary experiment the factors are varied one at the time, and the results are compared pairwise with a typical data condition included in the primary experiment. Tables 2 and 3 present the factors and factor levels considered in the primary and secondary experiments, respectively.

Performance Measure

As in this study the misclassification cost varies by data condition, an assessment of the relative accuracy of the classification methods requires a special performance measure. To this purpose the ratio $R_{i/B}=EC_i/EC_B$ is used, where EC_i and EC_B are the expected misclassification cost of method *i* and the Bayes rule, respectively. EC_B is a benchmark measure that reflects the performance of the "best" classification rule for that particular data condition, and $R_{i/B}$ measures the performance of each classification method relative to this benchmark. In order to assess the degree of difficulty of a given data condition, $R_{i/B}$ is also compared with $R_{NAIV/B}=EC_{NAIV}/EC_B$ where EC_{NAIV} is the expected cost of the "naïve" rule (NAIV) that assigns all entities to the group with the highest misclassification cost.

For each data condition, the Bayes rule is derived mathematically for each level of *RC*, after which the misclassification proportions e_{B1} and e_{B2} of this Bayes rule are determined for a balanced validation sample of 14,000 randomly generated observations. EC_B is then estimated as $EC_B=C_1*e_{B1}+C_2*e_{B2}$, where C_1 and C_2 are the relative classification costs, normalized such that $C_1+C_2=1$. The EC_i are estimated as follows. For each data condition, 50 balanced independent training samples are generated, and after estimating the relevant classification rule for each level of *RC*, the misclassification rates and cost are computed for the above validation sample. The average misclassification cost on the validation sample for all 50 replications serves as the estimate of EC_i for the data condition in question.

TABLE 2: FACTORS AND FACTOR LEVELS - PRIMARY EXPERIEMENT

Factor	Description
CM	*Classification method.* Eight factor levels: see Table 1.
SK	*Degree of skewness*. Two factor levels: moderate skewness (*M*) and high skewness (*H*).
RC	*Ratio of the group-wise misclassification costs* (C_1/C_2). Five factor levels: 0.1, 0.5, 1, 2 and 10.
RV	*Ratio of the group-wise attribute variances* s^2_2/s^2_1 . Three factor levels: 1, 4 and 64.

TABLE 3: FACTORS AND FACTOR LEVELS - SECONDARY EXPERIMENT

Factor	Description
DT	*Data transformation*. Two factor levels: no transformation (*ORG*); and positive square root transformation (*TRF*).
TS	*Relative training sample size*. Two factor levels: small samples (*S*), with the number of observations in each group equal to 5 times the number of attributes; and large samples (*L*), with the number of observations in each group equal to 10 times the number of attributes.
OVLP	*Group overlap*. Two factor levels: high overlap (*H*), with an expected misclassification rate for the optimal rule in the case of equal misclassification costs of 31.85 percent; and low overlap (*L*), with an expected misclassification rate for the optimal rule in the case of equal misclassification costs of 6.67 percent.
P	*Number of attributes*. Two factor levels: 3 attributes; and 10 attributes.
CORR	*Correlation structure of the attributes*. Two factor levels: independent attributes (*I*); and positively correlated attributes(*C*), with $\rho_{12}=0.8$, $\rho_{13}=\rho_{23}=0.4$.

Attribute Variable Generation

The attribute variables are generated from the family of log-normal distributions, with appropriate parameter values to control for the distributional characteristics of the experimental design. Let Y_{jk} represent the k^{th} attribute of observations from G_j, j=1, 2. In the primary experiment, Y_{jk} is defined by Y_{jk} I= exp(b_j*Z_{jk}II+c_j), where the Z_{jk} are independent standard normal random variables and b_j, c_j, are parameters (j=1, 2). The following equations are used to solve for b_1 and c_1:

$$\beta_1 = \left[\sqrt{\exp(b_1^2) - 1}\right](\exp(b_1^2) + 2)$$
$$\sigma_1^2 = [\exp(b_1^2 / 2 + c_1)]^2 (\exp(b_1^2) - 1) = 1$$

where β_1 and σ^2_1 represent the desired level of attribute skewness and variance for the observations in G_1. Once b_1 and c_1 have been determined, b_2 and c_2 are found by conducting a line search that ensures that (i) the misclassification rate of the Bayes rule (assuming equal costs) reflects the desired level of group overlap, (ii) the mode of the attributes of the observations belonging to G_2 is higher than that of the observations in G_1, and (iii) the attribute variance of the observations in G_2, $\sigma_2^2 = [\exp(b_2^2 / 2 - c_2)]^2 (\exp(b_2^2) - 1)$, yields the desired ratio of variances across groups σ^2_2/σ^2_1.

In one of the secondary experiments, in which the effect of the correlation structure (*CORR*) is analyzed, the assumption of independent attributes is relaxed, and several correlated attributes are generated. The details regarding the variable generation in that case will be presented within the discussion of the experiment.

PRIMARY SIMULATION EXPERIMENT

Experimental Design, Primary Experiment

The primary experiment uses a repeated measures design with *SK* and *RV* as between-subject factors and *CM* and *RC* as within-subject factors. The 50 training samples for each data condition are treated as "subjects." Moreover, each observation is described by *P*=3 independent identically distributed attributes, the size of each training sample equals 15 observations per group (*TS*=*S*), the group overlap corresponds to that of two multivariate normal populations with a common covariance matrix and a Mahalanobis distance of 3 (*OVLP=L*), all methods are applied to the original data (*DT=ORG*), and the attributes are independent (*CORR=I*). These last five factors (*DT*, *TS*, *OVLP*, *P* and *CORR*), are kept fixed at these levels in the primary experiment, but will be varied, one at the time, in the secondary experiments described in the following section.

The factor levels of *SK*, *RC* and *RV* are described next. The two levels of *SK* are "moderate" (*M*), defined by an attribute skewness coefficient b_1=1 for the observations in G_1, and "high" (*H*), with b_1=10 for the observations in G_1. In both cases, the attribute skewness of the observations belonging to G_2 is adjusted in order to achieve the desired levels of *OVLP* and *RV*.

The five levels of *RC* are *RC*=0.1, 0.5, 1, 2 and 10, covering the cases $C_1=C_2$, $C_i=2*C_j$ and $C_i=10*C_j$, $i, j=1, 2; i \neq j$. Preliminary comparisons show that these levels of *RC* differ in their impact on the performance of the classification methods, in particular whether the group where the attributes have the lower mode, has the lower or the higher cost. In the remainder of this paper, the group with the attribute distributions that have the lower mode will always be referred to as G_1. Hence, G_1 and G_2 may be described as "the group on the left" and "the group on the right," respectively.

The three levels of the group-wise variance ratio σ^2_2 /σ^2_1 are *RV* =1, 4, and 64. For skewed distributions with domain $[0,+\infty)$, higher modes usually correspond with higher

variances. For *RV*=1, the attribute distributions of G_1 are more skewed than those of G_2; if *RV*=4, the skewness levels are similar for both groups; and the distributions of G_2 are more skewed than those of G_1 if *RV*=64. The ratio *RV*=64 is included as a factor level to exemplify data conditions for which nonlinear rules will presumably yield the best results.

Table 4 summarizes the six data conditions considered in the primary experiment (data conditions 1-6). Table 5 lists the transformation parameters needed to achieve data conditions 1-6, as well as data conditions 9 and 10 of the secondary experiment, and Table 6 summarizes the corresponding distributional characteristics. This information is not included for data conditions 7 and 8 of the secondary experiments, because their distributional characteristics are the same as data condition 5. Data condition 11 of the secondary experiment involves correlated variables and is treated separately. In the remainder of this paper, the term "data condition" will be abbreviated by "dc." The estimated expected cost of the Bayes rule and the $R_{i/B}$ ratios for all methods considered and data conditions analyzed in both experiments are presented in Tables 7 and 8, respectively.

MANOVA (within-subject factors) and ANOVA (between-subject factors) analyses reveal that all main and interaction effects between *CM*, *SK*, *RV* and *RC* are significant at the .01 level. This result is not surprising, given the large number of replications (50 replications for each of the 8×2×3×5=240 different combinations of factor levels).

TABLE 4: DATA CONDITIONS - PRIMARY EXPERIMENT

	Factors Varied		Factors Fixed (in Primary Experiment)				
Condition	*SK*	*RV*	*DT*	*TS*	*OVLP*	*P*	*CORR*
1	*M*	1	*ORG*	*S*	*L*	3	*I*
2	*M*	4	*ORG*	*S*	*L*	3	*I*
3	*M*	64	*ORG*	*S*	*L*	3	*I*
4	*H*	1	*ORG*	*S*	*L*	3	*I*
5	*H*	4	*ORG*	*S*	*L*	3	*I*
6	*H*	64	*ORG*	*S*	*L*	3	*I*

TABLE 5: TRANSFORMATION PARAMETERS - PRIMARY AND SECONDARY EXPERIMENTS, INDEPENDENT ATTRIBUTES

Condition	Parameters			
	b_1	c_1	b_2	c_2
1	0.3143	1.0832	0.2114	1.5199
2	0.3143	1.0832	0.3488	1.6552
3	0.3143	1.0832	0.8035	1.8053
4	1.1651	-1.2087	0.6265	0.1700
5	1.1651	-1.2087	0.8005	0.4268
6	1.1651	-1.2087	1.1816	0.8256
9	1.1651	-1.2087	1.1724	-0.5353
10	1.1651	-1.2087	1.0410	-0.1841

TABLE 6: DISTRIBUTIONAL CHARACTERISTICS

Condition	Distributional Characteristics					
	Group 1			Group 2		
	Mode	S.D.	Skewness	Mode	S.D.	Skewness
	mo_1	s_1	s_1	mo_2	s_2	s_2
1	2.954	1.000	1.000	4.572	1.000	0.651
2	2.954	1.000	1.000	5.234	2.000	1.126
3	2.954	1.000	1.000	6.082	8.000	3.721
4	0.299	1.000	10.000	1.185	1.000	2.413
5	0.299	1.000	10.000	1.532	2.000	3.694
6	0.299	1.000	10.000	2.283	8.000	10.530
9	0.299	1.000	10.000	0.585	2.000	10.230
10	0.299	1.000	10.000	0.832	2.000	6.930

TABLE 7: ESTIMATED EXPECTED COST OF THE BAYES RULE

	RC				
Condition	0.1	0.5	1.0	2.0	10.0
1	0.0244	0.0573	0.0671	0.0670	0.0406
2	0.0349	0.0641	0.0668	0.0602	0.0306
3	0.0494	0.0680	0.0626	0.0506	0.0190
4	0.0222	0.0553	0.0676	0.0723	0.0468
5	0.0246	0.0582	0.0676	0.0665	0.0407
6	0.0338	0.0632	0.0667	0.0613	0.0324
7	0.0246	0.0582	0.0676	0.0665	0.0407
8	0.0246	0.0582	0.0676	0.0665	0.0407
9	0.0313	0.0651	0.0714	0.0670	0.0346
10	0.0910	0.2680	0.3138	0.2699	0.0914
11	0.0201	0.0523	0.0669	0.0712	0.0518

Skewness Effect (*SK*)

The results in Tables 7 and 8 for dc1 to dc6 show that each method tends to approximate the Bayes rule more accurately if the skewness level is moderate, as opposed to high. Figure 1 illustrates the effects on *SK* for the combination of *RV*=4 and *RC*=1. The pattern for the other combinations of *RV* and *RC* is similar. The degree of skewness tends to impact the parametric methods (LDF and QDF) most. For example, Figure 1 and Table 8 shows that the ratios $R_{LDF/B}$ and $R_{QDF/B}$ are almost twice as large for *SK=H* (dc5) as for *SK=M* (dc2). The corresponding results for the LGR and MP indicate that these methods are less sensitive to a high level of skewness.

TABLE 8: AVERAGE EXPECTED COST RATIOS OF THE EMPIRICAL AND BAYES RULES (PART 1: DATA CONDITIONS 1-6)

		Data		***RC***		
Condition	Method	0.1	0.5	1.0	2.0	10.0
1	NAIV	3.726	5.817	7.447	4.975	2.239
	LDF	1.538	1.510	1.515	1.559	1.718
	QDF	1.970	1.748	1.737	1.787	2.168
	LGR	2.523	1.751	1.686	1.754	2.448
	NN	1.824	1.996	1.677	1.893	1.984
	L0L	2.499	1.682	1.675	1.758	2.663
	L1L	3.748	2.065	2.025	2.229	4.026
	L0Q	3.912	2.184	2.169	2.481	4.820
	L1Q	6.489	3.342	3.192	3.550	6.669
2	NAIV	2.605	5.200	7.485	5.537	2.971
	LDF	1.422	1.354	1.367	1.378	1.322
	QDF	1.471	1.469	1.566	1.737	2.568
	LGR	2.056	1.482	1.445	1.521	2.150
	NN	1.764	1.693	1.521	1.903	1.672
	L0L	2.304	1.548	1.396	1.524	2.447
	L1L	2.691	1.718	1.786	2.156	4.447
	L0Q	2.651	2.015	2.049	2.548	4.011
	L1Q	5.220	3.223	3.349	3.974	8.626
3	NAIV	1.840	4.902	7.987	6.588	4.785
	LDF	1.971	2.123	2.474	2.712	2.904
	QDF	1.434	1.285	1.371	1.545	2.813
	LGR	2.450	2.007	1.943	2.001	2.797
	NN	2.287	1.977	2.088	2.531	2.038
	L0L	2.696	1.947	2.006	2.374	4.002
	L1L	3.021	2.198	2.310	2.715	6.587
	L0Q	1.891	1.781	2.205	3.005	6.774
	L1Q	3.620	2.809	3.104	3.966	11.468
	NAIV	4.095	6.028	7.396	4.610	1.943
	LDF	2.558	2.703	2.824	2.695	2.578
	QDF	4.019	3.334	3.173	3.052	3.827
	LGR	2.944	2.645	2.694	2.689	3.347
	NN	2.712	2.686	2.555	2.694	2.598
	L0L	3.911	2.626	2.586	2.703	3.787
	L1L	5.386	2.993	2.778	2.777	4.100
	L0Q	5.474	2.991	2.741	2.876	4.399
	L1Q	6.739	3.249	2.892	2.953	5.109
5	NAIV	3.695	5.727	7.396	5.013	2.234
	LDF	2.394	2.406	2.496	2.431	2.057
	QDF	3.727	2.843	2.699	2.786	3.991
	LGR	2.824	2.194	2.187	2.311	3.130
	NN	2.649	2.294	2.287	2.730	2.444
	L0L	3.391	2.228	2.095	2.285	3.436
	L1L	4.557	2.494	2.362	2.539	4.234
	L0Q	5.092	2.799	2.761	3.102	5.296
	L1Q	5.828	2.974	2.814	3.112	5.707
6	NAIV	2.690	5.274	7.496	5.438	2.806
	LDF	2.390	2.277	2.648	2.587	1.939
	QDF	2.686	1.942	1.985	2.229	4.009
	LGR	2.207	1.709	1.742	1.883	2.915
	NN	2.300	1.919	2.023	2.437	2.045
	L0L	2.581	1.763	1.705	1.884	3.287
	L1L	2.933	1.923	1.972	2.250	4.517
	L0Q	3.618	2.417	2.458	2.963	6.075
	L1Q	5.380	2.972	2.878	3.225	6.335

(Part 2: Data Conditions 7-11)

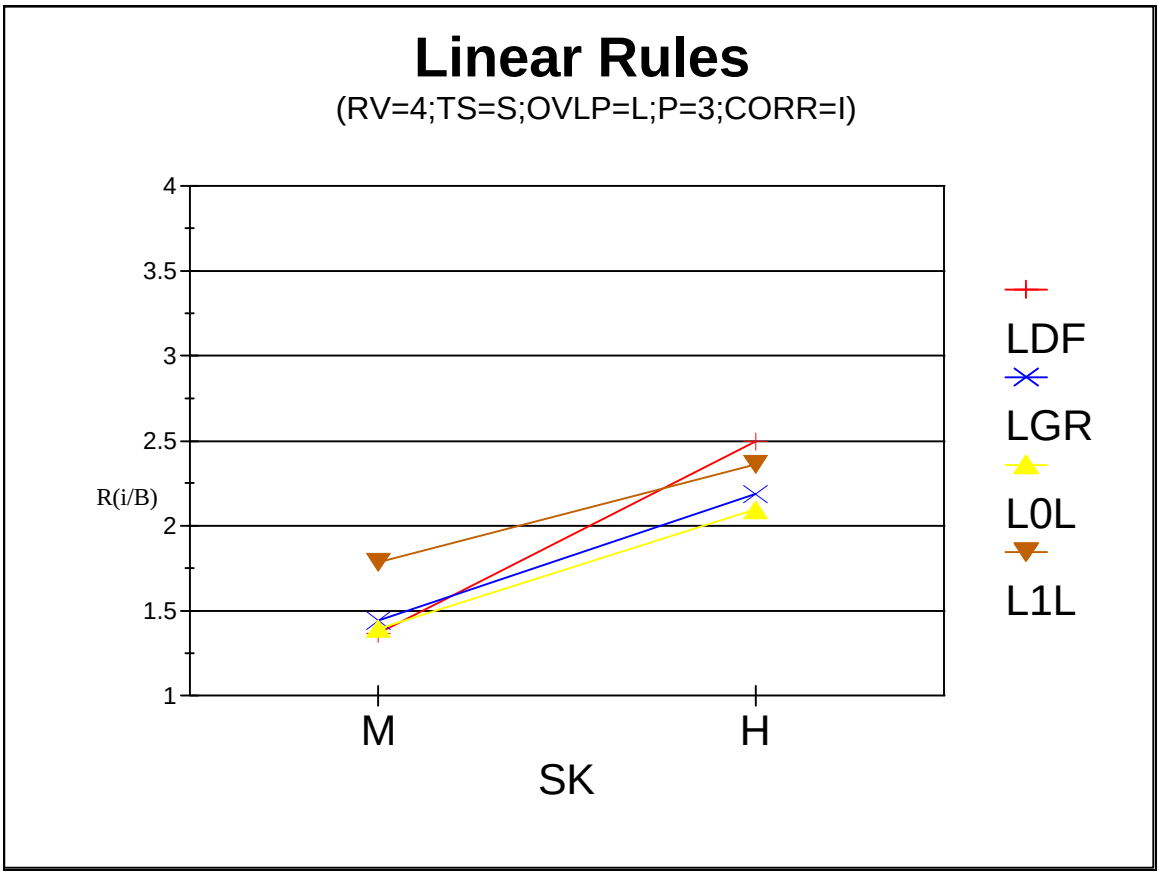
Data				***RC***		
Condition	Method	0.1	0.5	1.0	2.0	10.0
7	NAIV	3.695	5.727	7.396	5.013	2.234
	LDF	1.793	1.722	1.738	1.768	1.774
	QDF	2.271	2.104	2.126	2.257	2.977
	LGR	2.864	1.907	1.846	1.941	2.706
	NN	2.079	2.067	1.842	2.122	2.060
	L0L	3.177	1.884	1.781	1.906	2.911
	L1L	4.066	2.166	2.117	2.322	4.015
	L0Q	4.574	2.494	2.440	2.687	4.123
	L1Q	5.902	2.816	2.632	2.865	5.159
8	NAIV	3.695	5.727	7.396	5.013	2.234
	LDF	2.467	2.184	2.460	2.477	1.991
	QDF	3.753	2.924	2.678	2.602	3.174
	LGR	2.152	1.953	2.025	2.122	2.427
	NN	2.326	2.152	2.046	2.428	2.261
	L0L	2.658	2.065	2.064	2.182	2.877
	L1L	2.824	2.065	2.061	2.162	2.937
	L0Q	4.702	2.455	2.270	2.546	4.160
	L1Q	5.050	2.545	2.407	2.621	4.549
9	NAIV	0.999	1.244	1.593	1.235	0.995
	LDF	1.068	1.223	1.169	1.137	1.238
	QDF	2.881	1.530	1.244	1.224	2.258
	LGR	1.108	1.211	1.152	1.166	1.567
	NN	2.010	1.238	1.287	1.237	1.771
	L0L	2.394	1.254	1.181	1.286	2.542
	L1L	2.279	1.273	1.171	1.274	2.491
	L0Q	4.041	1.557	1.367	1.547	4.298
	L1Q	3.917	1.520	1.356	1.555	4.476
10	NAIV	2.904	5.120	7.003	4.975	2.627
	LDF	1.971	2.540	2.653	2.616	2.364
	QDF	6.804	3.778	3.430	3.500	5.719
	LGR	2.310	2.331	2.394	2.495	3.244
	NN	3.900	2.898	3.244	3.683	2.860
	L0L	4.360	2.665	2.498	2.716	4.631
	L1L	3.981	2.529	2.506	2.677	4.817
	L0Q	5.051	3.140	3.285	3.886	7.555
	L1Q	13.250	5.835	4.986	4.957	8.598
11	NAIV	4.523	6.373	7.474	4.682	1.755
	LDF	2.927	2.418	2.272	2.060	1.462
	QDF	3.274	2.444	2.140	2.066	2.508
	LGR	3.730	2.057	1.773	1.686	1.890
	NN	2.761	1.873	1.866	2.340	1.618
	L0L	4.586	2.108	1.654	1.630	1.893
	L1L	4.968	2.162	1.864	1.832	2.445
	L0Q	6.035	2.651	2.194	2.304	3.176
	L1Q	6.959	2.928	2.447	2.383	3.459

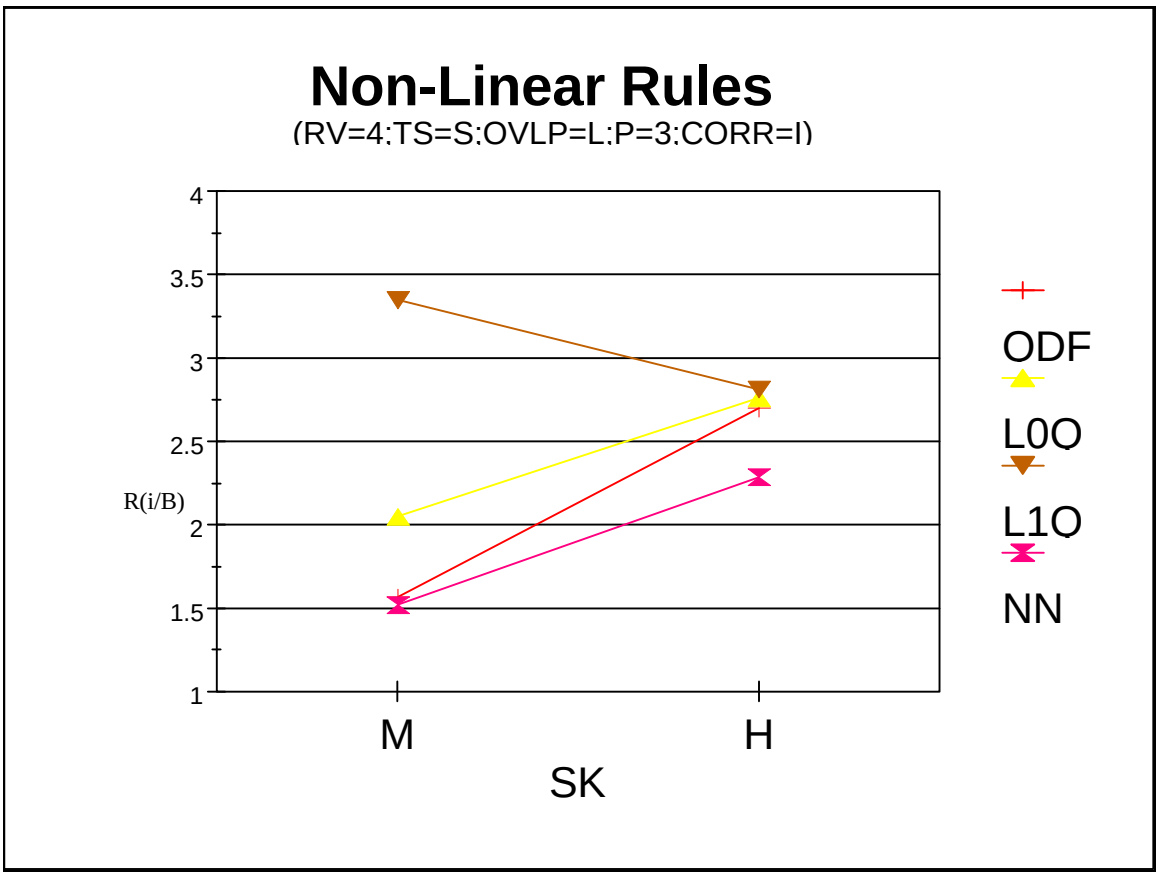
This is true in particular for the L_1-norm MP methods, as $R_{L1L/B}$ increases from 1.79 (dc2) to 2.36 (dc5) and $R_{L1Q/B}$ decreases from 3.35 (dc2) to 2.81 (dc5). The corresponding figures for $R_{LGR/B}$ are 1.45 (dc2) and 2.19 (dc5), and for $R_{L0L/B}$, 1.40 (dc2) and 2.10 (dc5).

Ratio of Group-Wise Misclassification Costs Effect (*RC*)

In general, each of the classification methods yields better results if the misclassification costs are similar across groups than if they are clearly different. Furthermore, *RC* has a strong impact on the relative performance of the individual classification methods.

Figure 1: Skewness Effect (SK)



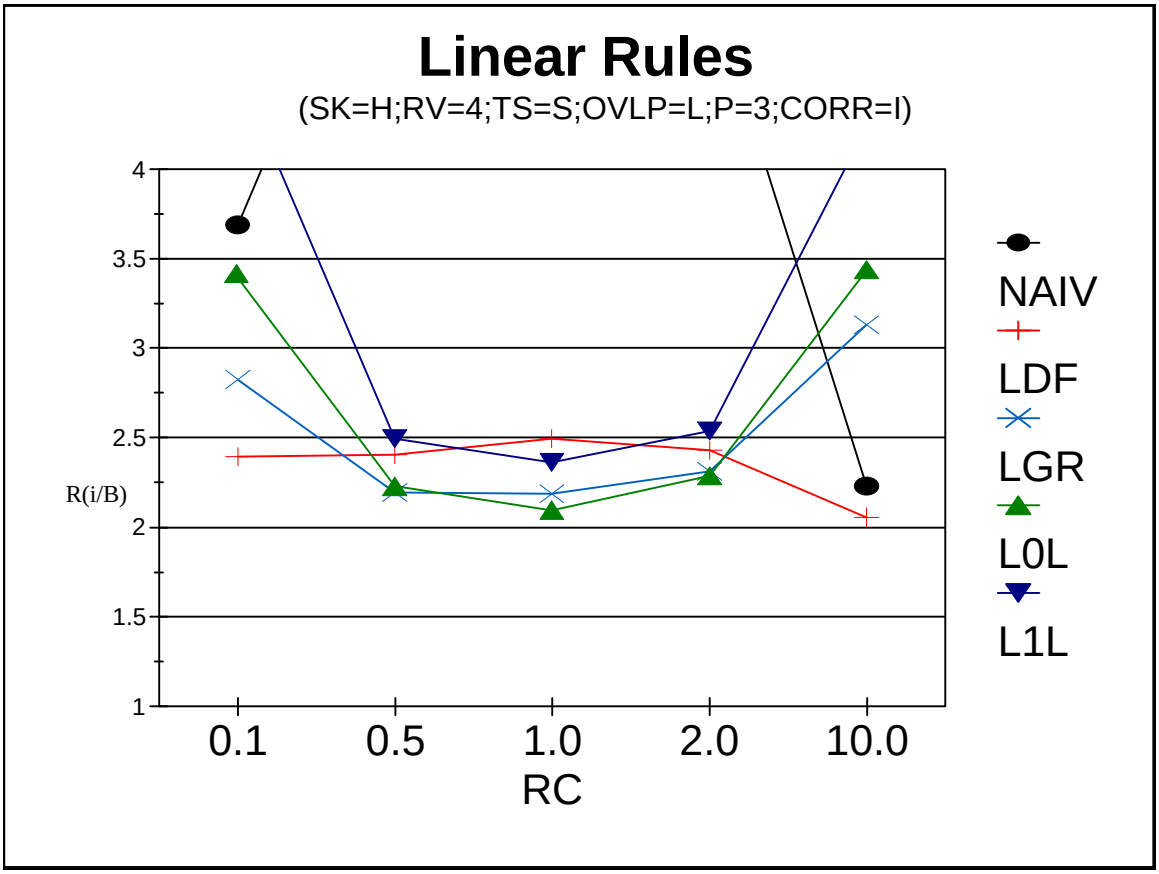


In particular, the LDF usually classifies more accurately, in relative terms, as the difference between misclassification costs increase. Notably, the LDF is more robust to different misclassification costs across groups than the LGR. The performance of each of the MP-based methods deteriorates quickly as *RC* is further apart from 1.
For instance, from Table 8 we see that, if *SK*=*H*, *RV*=4 (dc5) and *RC*=1, $R_{LDF/B}$=2.50 exceeds the $R_{i/B}$ ratios of the LGR, NN, L0L and L1L methods, indicating that the LDF is inferior to these other methods. However, if the group-wise ratio of misclassification costs equals 10 (RC = 0.1 or RC = 10.0), the LDF ranks first, with $R_{LDF/B}$ equal to 2.39 and 2.06, respectively, for *RC*=0.1 and *RC*=10.0. Under these conditions, NN is the second best method, with $R_{NN/B}$ equal to 2.65 (*RC*=0.1) and 2.44 (*RC*=10). The LGR and the MP methods are clearly inferior, with $R_{LGR/B}$, $R_{L0L/B}$ and $R_{L1L/B}$ equal to 2.82, 3.39, 4.56 (*RC*=0.1) and 3.13, 3.44, 4.23 (*RC*=10), respectively. Interestingly, for *RC*=10 the LDF is the only rule that beats the naïve rule ($R_{NAIV/B}$=2.23). The *RC* effect for dc5, with *SK*=*H* and *RV*=4, is illustrated in Figure 2. The results for other combinations of *SK* and *RV* are similar.

Ratio of Group-Wise Attribute Variances Effect (*RV*)
As expected, the relative performance of the QDF tends to improve, both in absolute and relative terms, as the variance heterogeneity across groups (*RV*) increases. For instance, from Table 8 it is seen that for the combination of *SK*=*M* and *RC*=1, $R_{QDF/B}$ decreases from 1.74 (dc1) to 1.37 (dc3) as *RV* increases from 1 to 64. In relative terms, the QDF moves from the fourth most accurate to the most accurate method as *RV* increases from 1 to 64. The second best method for dc3 is LGR, with $R_{LGR/B}$ =1.94.
Interestingly, the performance of the linear rules tends to approximates that of the optimal Bayes rule more closely for *RV*=4 than for *RV*=1. This behavior is contrary to known results for normal distributions, in which case linear rules give the best results when the covariances are equal across groups (*e.g.*, Marks and Dunn 1974). In the case of skewed distributions the performance of linear classifications rules can usually be improved by data transformations that reduce the skewness level. It is remarkable that the linear rules still yield the best performance for moderately different variances across groups, when applied to the original data.
The variance heterogeneity level of *RV*=64 has a stronger effect on the performance of the LDF than on that of the other linear classification methods. For example, for the combination of *SK*=*M* and *RC*=1, a change in *RV* from 4 (dc2) to 64 (dc3) results in an increase in the $R_{i/B}$ ratios from 1.37 to 2.47 (LDF), from 1.45 to 1.94 (LGR) and from 1.40 to 2.01 (L0L). This effect is illustrated in Figure 3.

Higher Order Interactions
All of the interactions between *CM* and the other factors are found to be significant at the 0.0001 level. The most important higher order interactions are those between *CM*, *SK* and *RC* (η^2=0.763) and between *CM*, *RV* and *RC* (η^2=0.756). Thus, the way in which *RC* affects the relative performance of the misclassification methods depends on the distributional characteristics of the different data conditions. In particular, the choice of which group has the highest misclassification cost can have a major impact on relative classification performance. This impact is directly related to the levels of *SK* and *RV*.
For instance, consider the following two observations (see Table 8):

Figure 2: Ratio of Misclassification Costs Effect (RC)



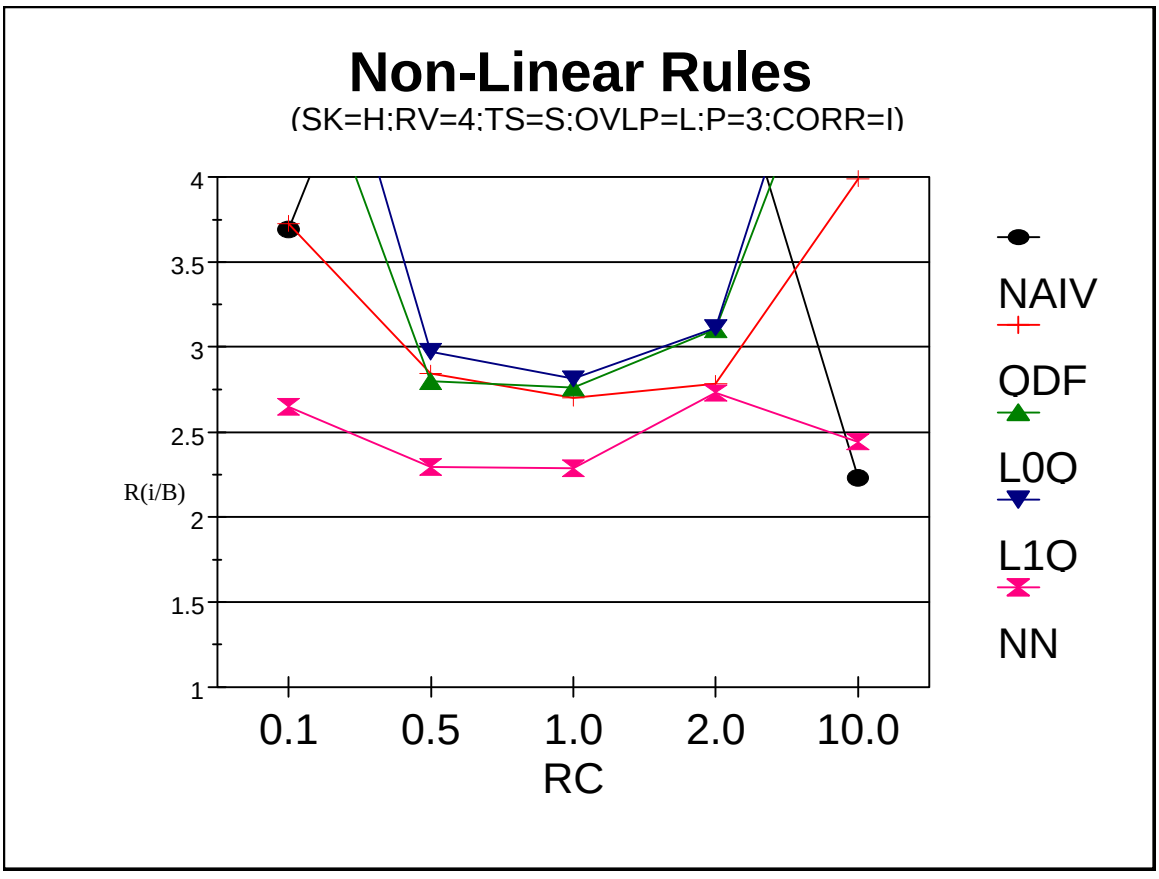
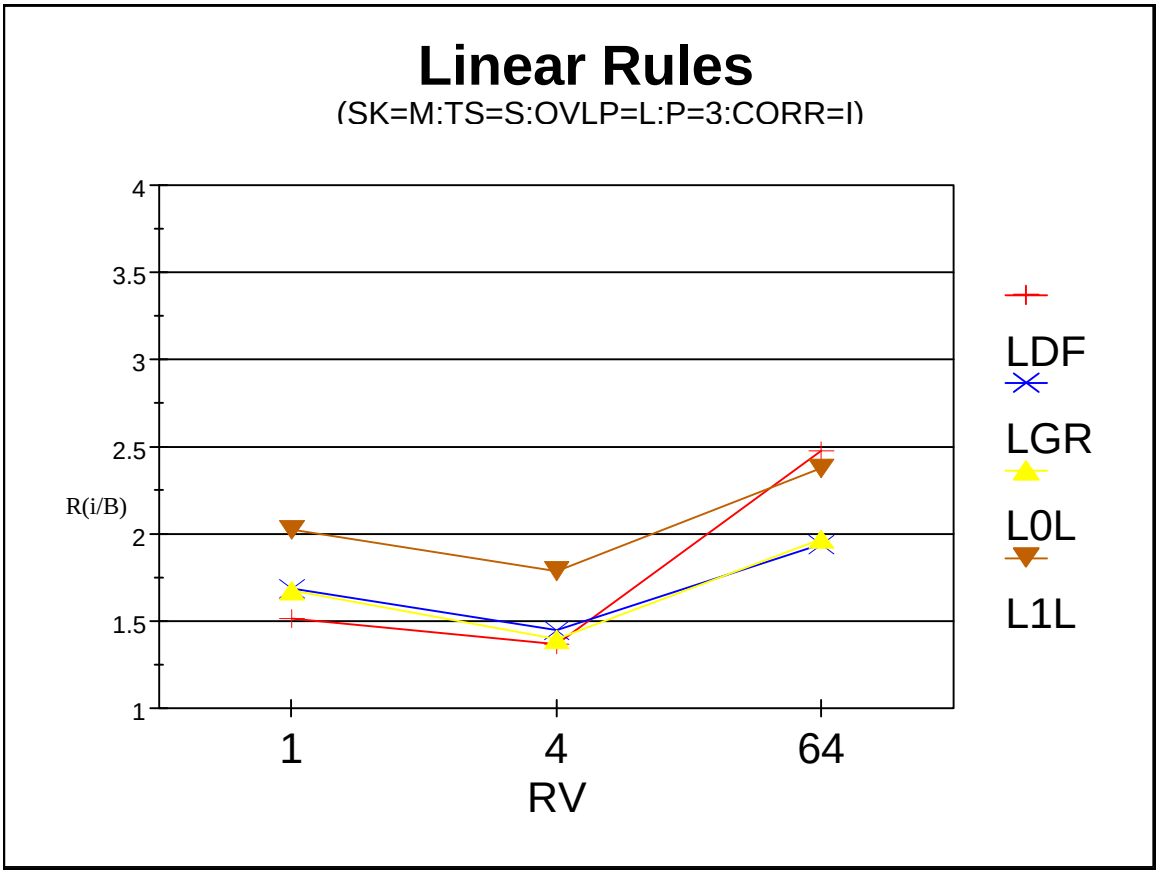
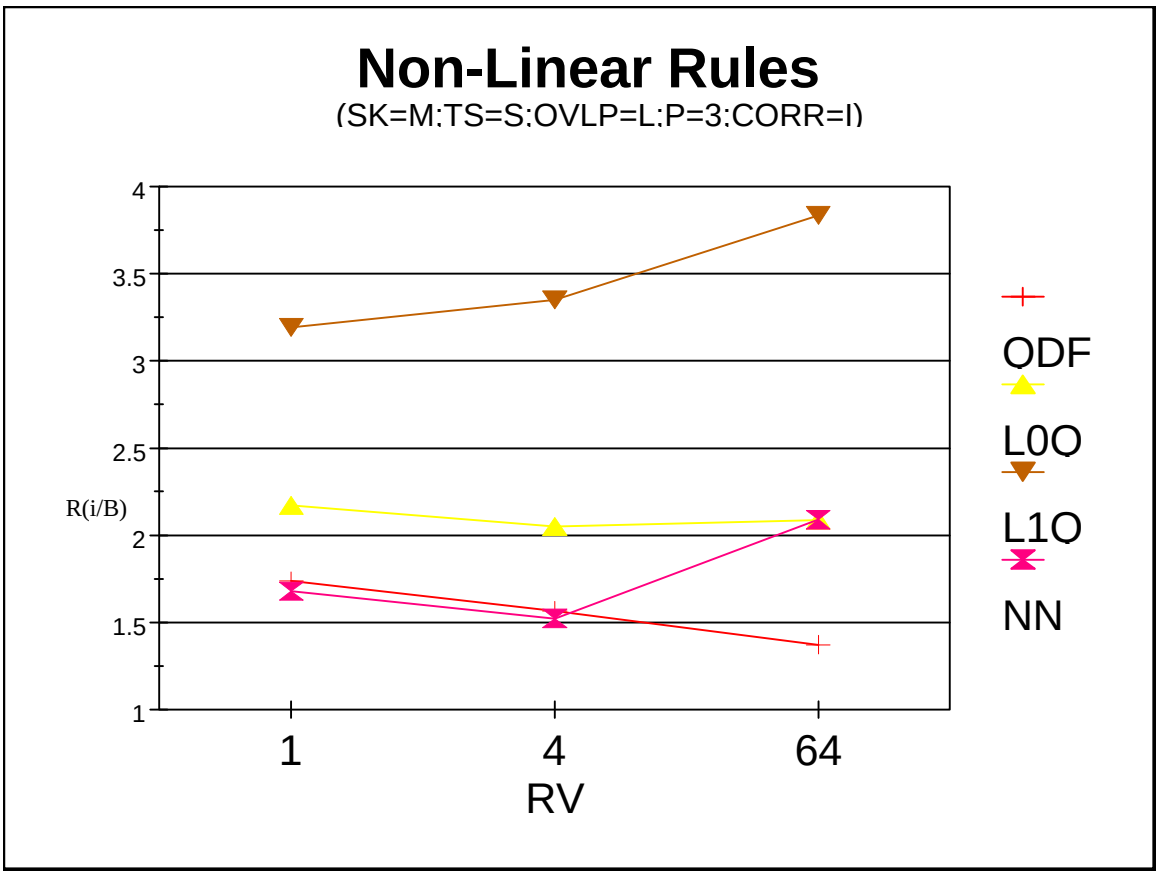


Figure 3: Ratio of Attribute Variances Effect (RV)





i) If *RV*=1, $R_{NAIV/B}$ tends to be closer to 1 when G_1 has the higher misclassification cost (*i.e.*, *RC*>1). This effect is stronger for *SK*=*H* (dc4) than for *SK*=*M* (dc1). For example, for the combination of *SK*=*H* and *RV*=1, $R_{NAIV/B}$ equals 4.10 when *RC*=0.1 and 1.94 when *RC*=10. The latter case can be described as a difficult data condition for classification purposes, since the best possible rule cannot improve the naïve rule by more than 50 percent. Remarkably, for this combination of factor levels (dc4, with *RC*=10) none of the empirically derived rules was able to beat the naïve rule on average, as $R_{i/B}>R_{NAIV/B}$, for all *i*. This effect is illustrated in Figure 4a).

ii) As *RV* increases, $R_{NAIV/B}$ decreases when G_2 has the higher misclassification cost (*i.e.*, *RC*<1), indicating that the corresponding data condition becomes more difficult in terms of accurate classification. For example, for dc3, with *SK*=*M* and *RV*=64, the naïve rule yields $R_{NAIV/B}$ =1.84 when *RC*=0.1, and the only method that beats the naïve rule is the QDF, with $R_{QDF/B}$ =1.43. However, for *RC*=10, the naïve rule perfroms poorly ($R_{NAIV/B}$=4.79), and the NN method ($R_{NN/B}$=2.04) handily beats all other methods, including the QDF ($R_{QDF/B}$=2.81). This effect is illustrated in Figure 4b.

The effect in i) may be explained as follows. If *RC*>1, *i.e.*, if the misclassification cost of G_1 is higher than that of G_2, the probability of misclassifying observations from G_1, p(2|1), is bound to be lower than if *RC*≤1. With attribute distributions that are skewed to the right, large reductions in p(2|1) can be achieved only by expanding the region of the attribute space assigned to G_1 well to the right and misclassifying almost all of the observations belonging to G_2. As a result, the Bayes rule becomes similar to the "naïve" rule, *i.e.*, $R_{NAIV/B}$ approaches 1. In contrast, if *RC*<1, *i.e.*, if the misclassification cost of G_2 is higher, significant reductions in p(1|2) can be achieved by smaller changes in the regions assigned to each group, and hence involve smaller increases in p(2|1). In this case, the performance of the Bayes rule can be significantly better than that of the "naïve" rule, so that $R_{NAIV/B}$ is larger. The higher the skewness of the attribute distributions, the stronger this effect is felt. The "difficult" data condition *RC*>1 has a similar effect for all of the classification methods included in this study.

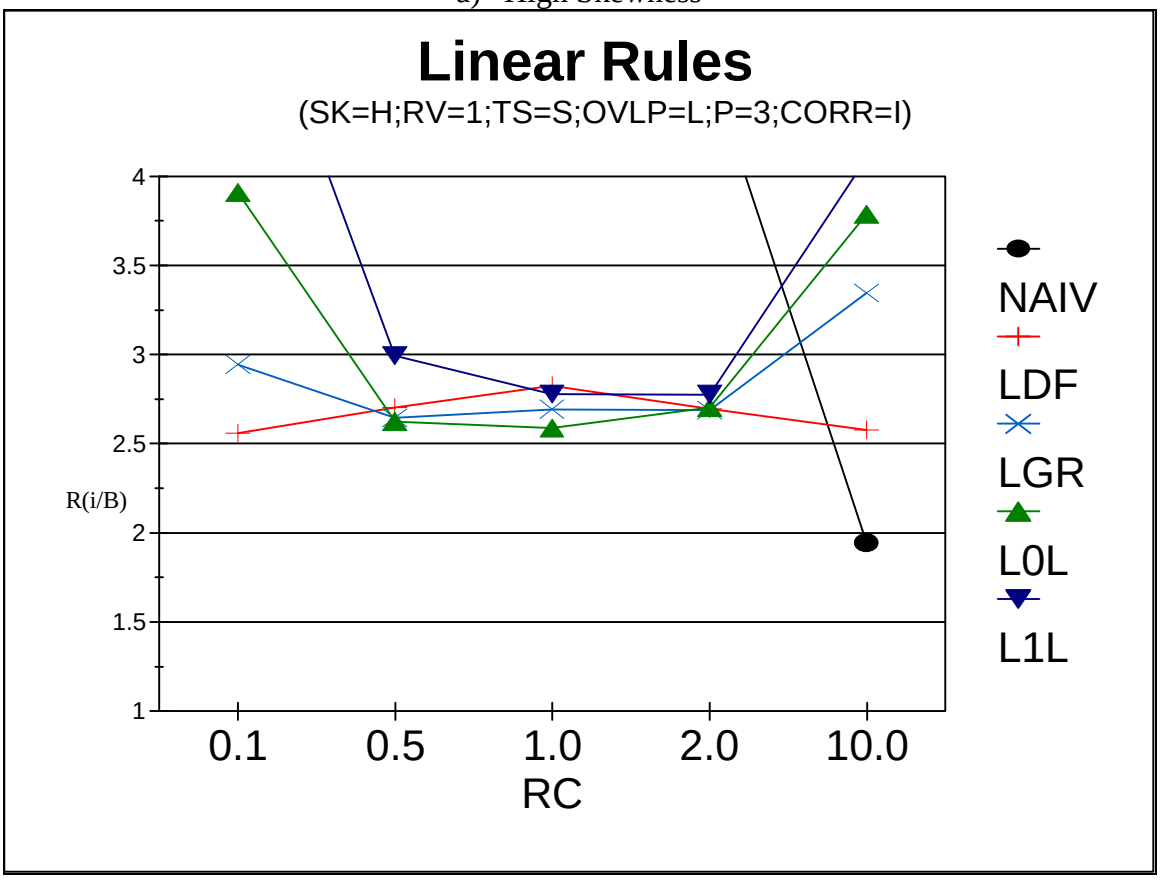
Effect ii) may be explained by the following argument. As *RV* (σ^2_2/σ^2_1) increases, the dispersion of G_2 increases, and a given expansion of the region in attribute space assigned to G_1 results in smaller increases of p(1|2). Thus, significant reductions in p(2|1) require larger expansions of the region assigned to G_2 and higher increases in p(1|2). As a result, $R_{NAIV/B}$ is bound to be lower ("the difficult condition") when the cost of the "group on the right" is higher than the cost of the "group on the left" (*i.e.*, *RC*<1). In this situation, the quadratic classification methods, notably the QDF, are less affected by the "difficult condition" than the other methods, and are better in relative terms when *RC* is smaller.

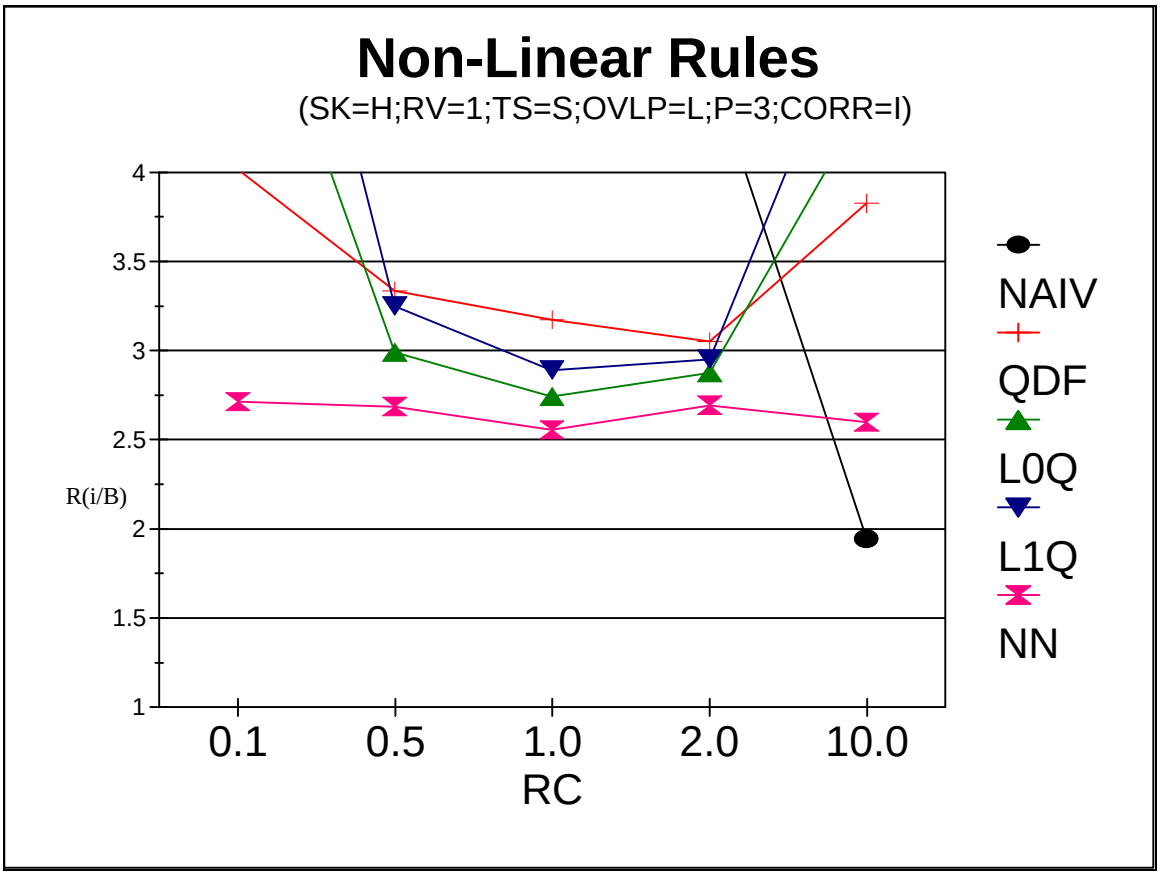
SECONDARY SIMULATION EXPERIMENTS

In this study, secondary experiments are conducted to analyze the impact on classification accuracy of the five factors (data transformation *DT*, training sample size *TS*, group overlap *OVLP*, number of attributes *P* and the correlation structure of the attributes *CORR*) that were kept at a fixed level in the primary experiment. These experiments are all based on dc5 of the primary experiment, because its factor levels *SK*=*H*, *RV*=4 reflect fairly typical conditions. The level *SK*=*H* is selected because the focus of this study is on skewed distributions. The level *RV*=4 is chosen because in the log-normal distribution,

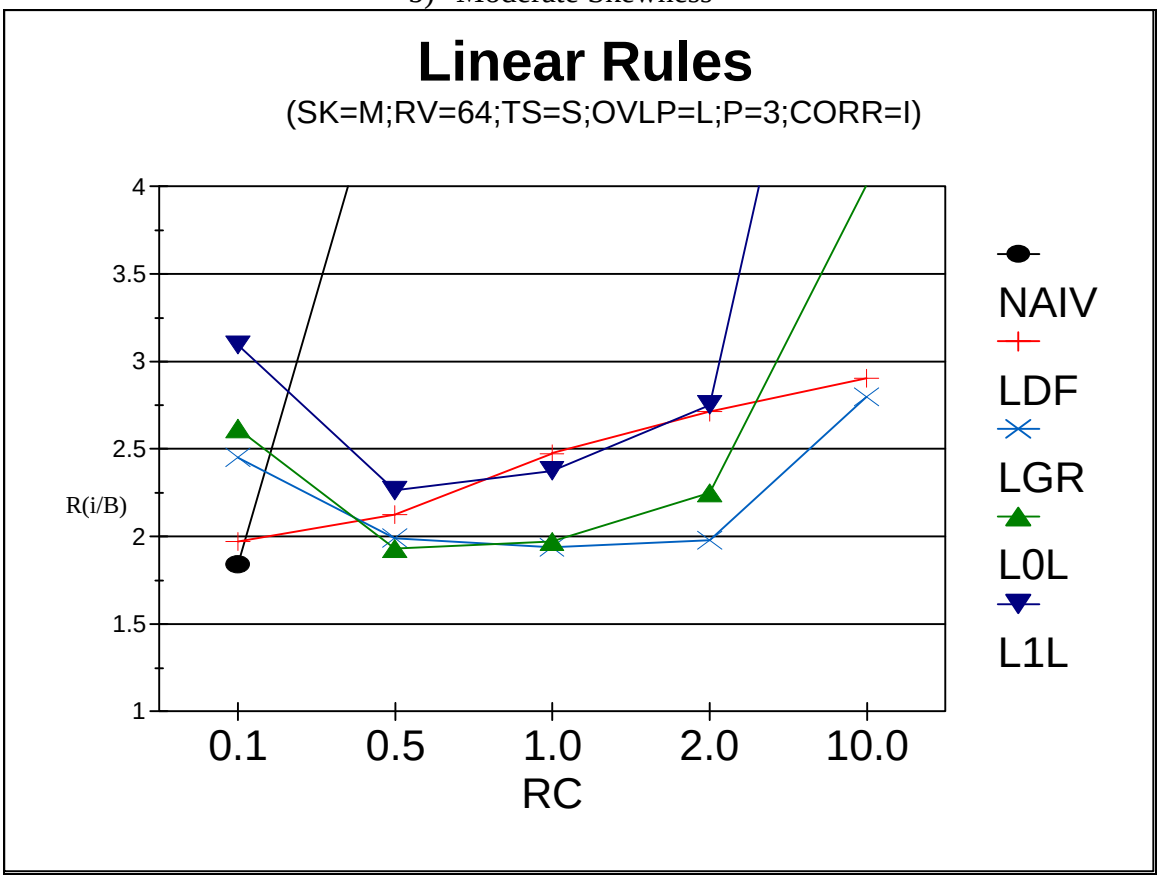
Figure 4: Higher Order Interactions

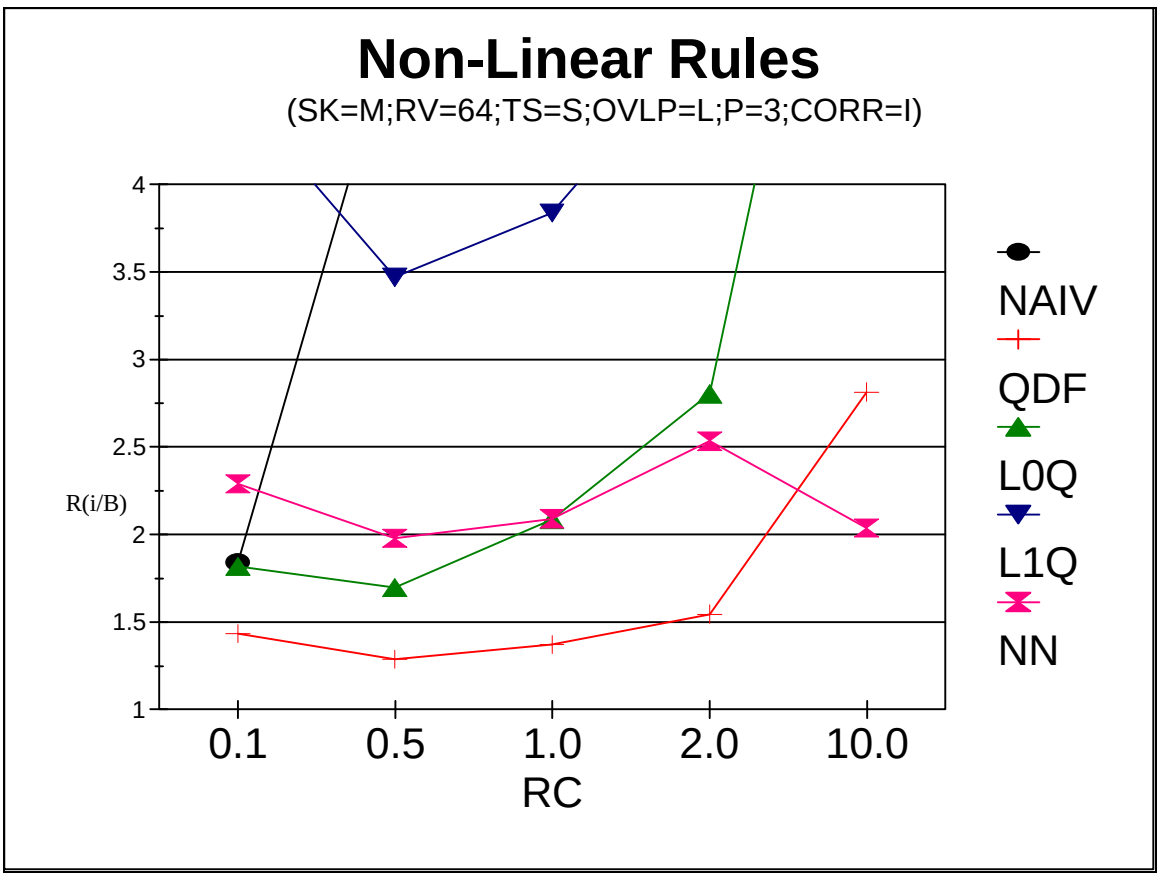
a) High Skewness





b) Moderate Skewness





as well as in other distributions with domain $\Re^+$, a "shift to the right" usually implies a moderate increase in dispersion. The classification methods and the levels of *RC* in the secondary experiments are the same as in the primary experiment (see Tables 1 and 2). In each of the secondary experiments, the classification performance of dc5 is compared pairwise with a variant of dc5, with *one* factor level modified.

The factor levels and the data conditions in the secondary experiments are described in Tables 3 and 9, respectively. The experimental design, the motivation for selecting the particular factor levels, and the relative classification performance for each of these pairwise evaluations in the secondary experiment are discussed next. The significance of the factor effects in each of the secondary experiments is again analyzed using MANOVA (within-subject factors) and ANOVA (between-subject factors) models, in which each of the 50 training samples generated is treated as a "subject." The estimated expected cost of the Bayes rules and the $R_{i/B}$ ratios for each data condition are presented in Tables 7 and 8, respectively.

TABLE 9: DATA CONDITIONS, PAIRWISE SECONDARY EXPERIMENTS

Factor Analyzed		Factors Fixed in the Secondary Experiments			Factors Varied in the Secondary Experiments			
	Data Condition	*SK*	*RV*	*DT*	*TS*	*OVLP*	*P*	*CORR*
Data Transformations	5	*H*	4	*ORG*	*S*	*L*	3	*I*
	7	*H*	4	*TRF*	*S*	*L*	3	*I*
Training Sample Size	5	*H*	4	*ORG*	*S*	*L*	3	*I*
	8	*H*	4	*ORG*	*L*	*L*	3	*I*
Group Overlap	5	*H*	4	*ORG*	*S*	*L*	3	*I*
	9	*H*	4	*ORG*	*S*	*H*	3	*I*
Number of Attributes	5	*H*	4	*ORG*	*S*	*L*	3	*I*
	10	*H*	4	*ORG*	*S*	*L*	10	*I*
Correlation Structure	5	*H*	4	*ORG*	*S*	*L*	3	*I*
	11	*H*	4	*ORG*	*S*	*L*	3	*C*

Transformation Effect (*DT*)

The first secondary experiment considers the effect of a positive square root transformation of the data (*DT*) on the relative performance of the classification methods. The experiment uses a repeated measures design with *CM*, *RC*, and *DT* as within-subject factors. The two levels of *DT* consist of applying the classification methods to either the original data (*ORG*, dc5) or the data after the transformation (*TRF*, dc7). Other than the data transformation, dc5 and 7 are identical (see Table 9). Thus, in this experiment the relative performance of dc5 is compared pairwise with that of dc7.

The purpose of this experiment is to evaluate the extent to which the performance of each method can be improved by transformations that are aimed at reducing the deviations from normality. Note that, due to the way in which the attribute distributions were generated in this study, the transformed attribute variables would be exactly normally distributed if a logarithmic transformation were used, implying that the assumptions underlying the parametric methods (LDF, QDF) would be satisfied perfectly. However,

the conclusions to be drawn from the results would be biased and less than interesting, because in practice it is not possible to find transformations that achieve perfect normality. Hence the choice in this paper to use the positive square root transformation, which does not yield exactly normally distributed attributes.

A MANOVA analysis reveals that all main and interaction effects between *CM*, *RC* and *DT* are significant at the .0001 level. From Tables 7 and 8, it is seen that the data transformation tends to reduce the expected cost for all classification methods, except under some very unfavorable circumstances. This reduction tends to be the strongest for the parametric methods (LDF and QDF). The primary beneficiaries from the data transformation are the LDF, in the case of similar misclassification costs, and the QDF, in the case of different classification costs. For instance, Table 8 shows that, assuming equal costs (*RC*=1), the performance of the LDF applied to the original data ($R_{LDF/B}$=2.5) ranks fifth, behind the LOL, LGR, NN and L1L (see dc5). Applied to the data after transformation (dc7), the LDF performs the best ($R_{LDF/B}$=1.74), followed closely by the L0L ($R_{L0L/B}$=1.78), NN ($R_{NN/B}$=1.84) and LGR ($R_{LGR/B}$= 1.85). If the misclassification costs differ strongly across groups (*RC*=0.1 or *RC*=10), the LDF and NN perform the best (in this order), regardless of whether the data have been transformed or not. The QDF performs poorly when applied to the original data, but its performance improves dramatically if the data have been transformed. For instance, when *RC*=0.1 with the original data, the $R_{i/B}$ ratios for the LDF and NN equal 2.39 and 2.65, respectively, while the QDF ($R_{QDF/B}$=3.73) is beaten even by the naïve rule ($R_{NAIV/B}$ =3.70). After the data transformation, the LDF and NN are still better ($R_{LDF/B}$=1.79, $R_{NN/B}$=2.08), but the QDF is a close third ($R_{QDF/B}$=2.27). These effects are illustrated in Figure 5.

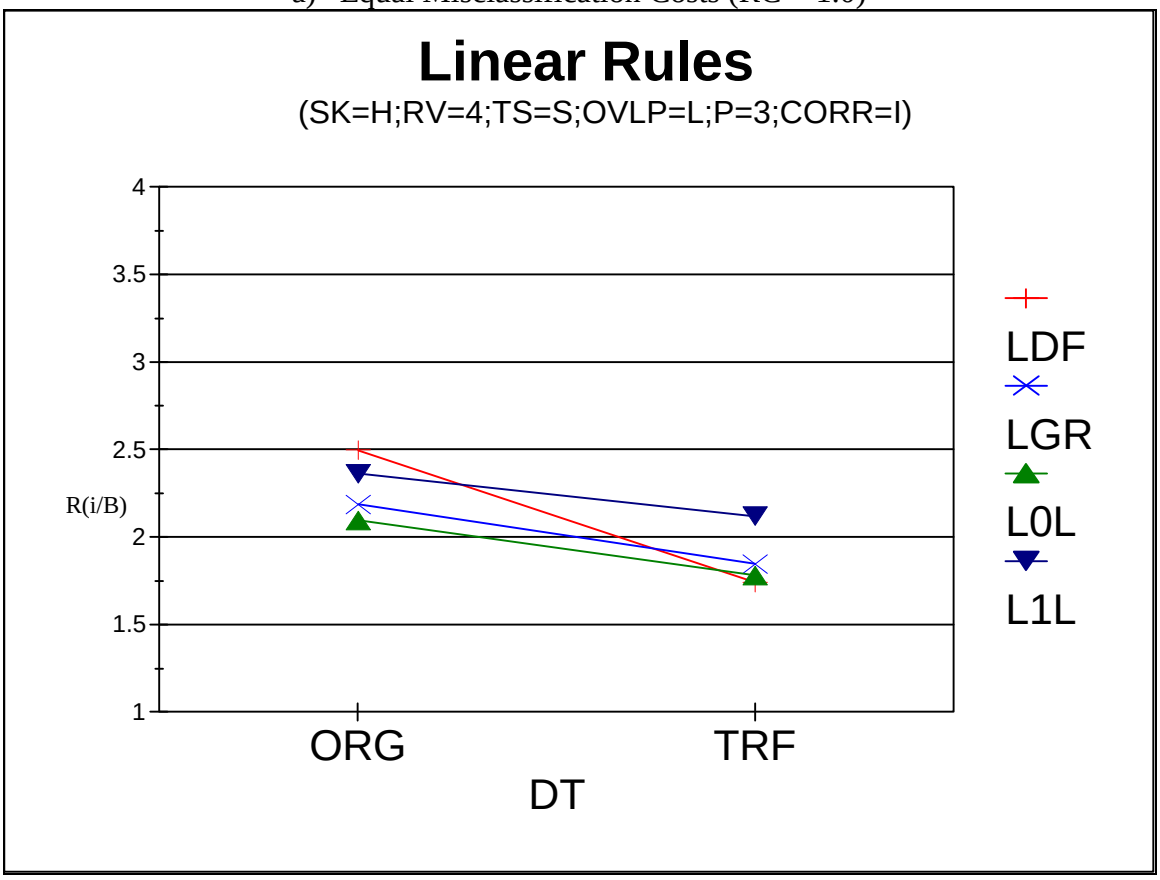
Training Sample Size Effect (*TR*)

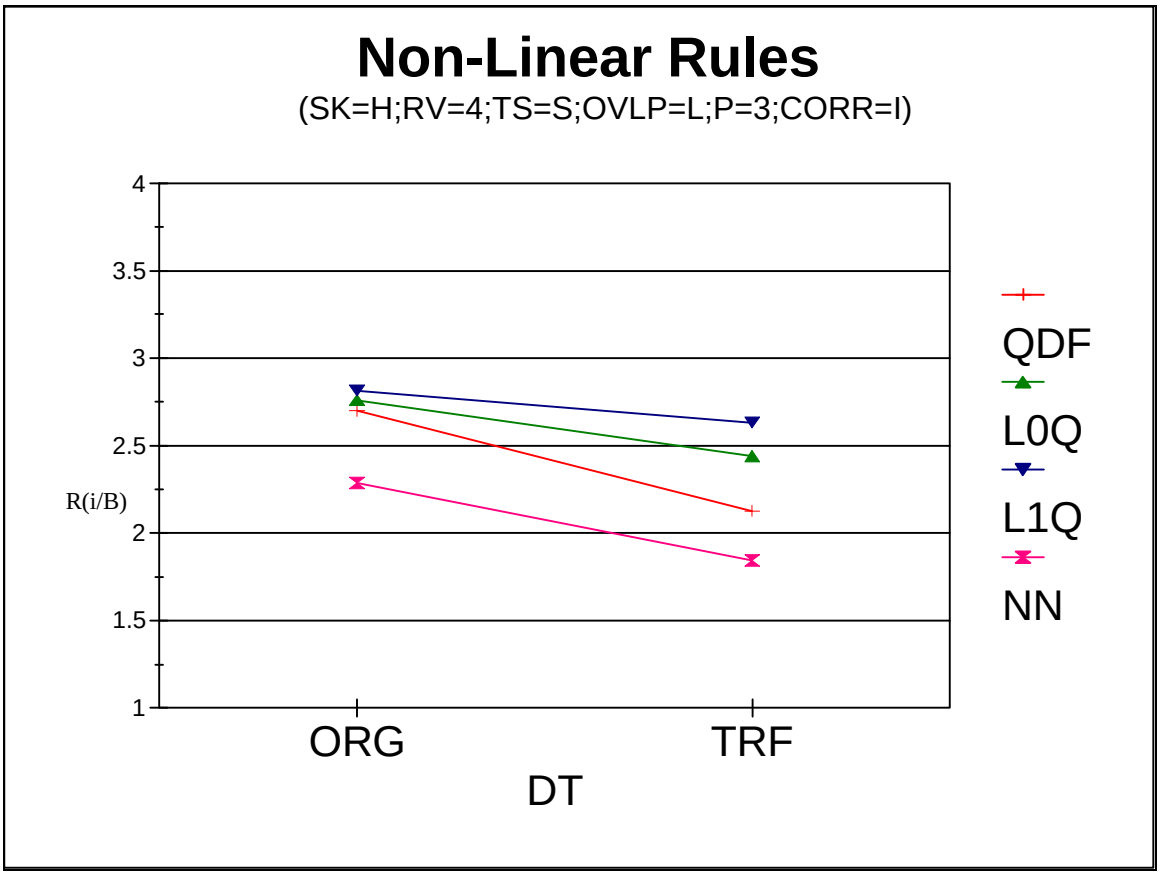
The second secondary experiment considers the effects of the size of the training sample (*TS*) and uses a repeated measures design with *TS* as between-subject factors and *CM*, *RC* as within-subject factors. The levels of *TS* reflect a small training sample size *TS=S*, with a ratio of the number of training sample observations per group (*n*) and the number of attributes (*p*) of *n/p*=5 (dc5), and a large training sample size *TS=L*, with *n/p*=10 (dc8) (see Table 9). Training samples with less than 5*p* observations per group usually do not provide enough information to fit useful classification functions. Training samples with at least 10*p* observations per group are typically considered large.

The MANOVA and ANOVA results reveal that all main and interaction effects between *CM*, *RC* and *TS* are significant at the .0001 level. As expected, the results in Table 8 show that all methods perform better if *TS=L* (dc8). The L1L, L1Q and LGR are particularly sensitive to the training sample size, whereas the LDF and L0L are affected the least by the training sample size. For instance, for *RC*=1 a reduction of the training sample size from 30 to 15 observations results in an increase in $R_{LGR/B}$ from 2.03 to 2.19. The corresponding increases for the LDF and L0L are much smaller, from 2.46 to 2.50 and from 2.06 to 2.10, respectively. Under the same conditions, $R_{L1L/B}$ increased from 2.06 to 2.36. The good performance of the L0L with small training samples is particularly surprising because this result contradicts previous studies (Koehler and Erenguc 1990; Stam and Jones 1990;), which found this method to be very sensitive to the training sample size. One possible explanation for this discrepancy is that, in contrast with Koehler and Erenguc (1990) and Stam and Jones (1990), the implementation of the

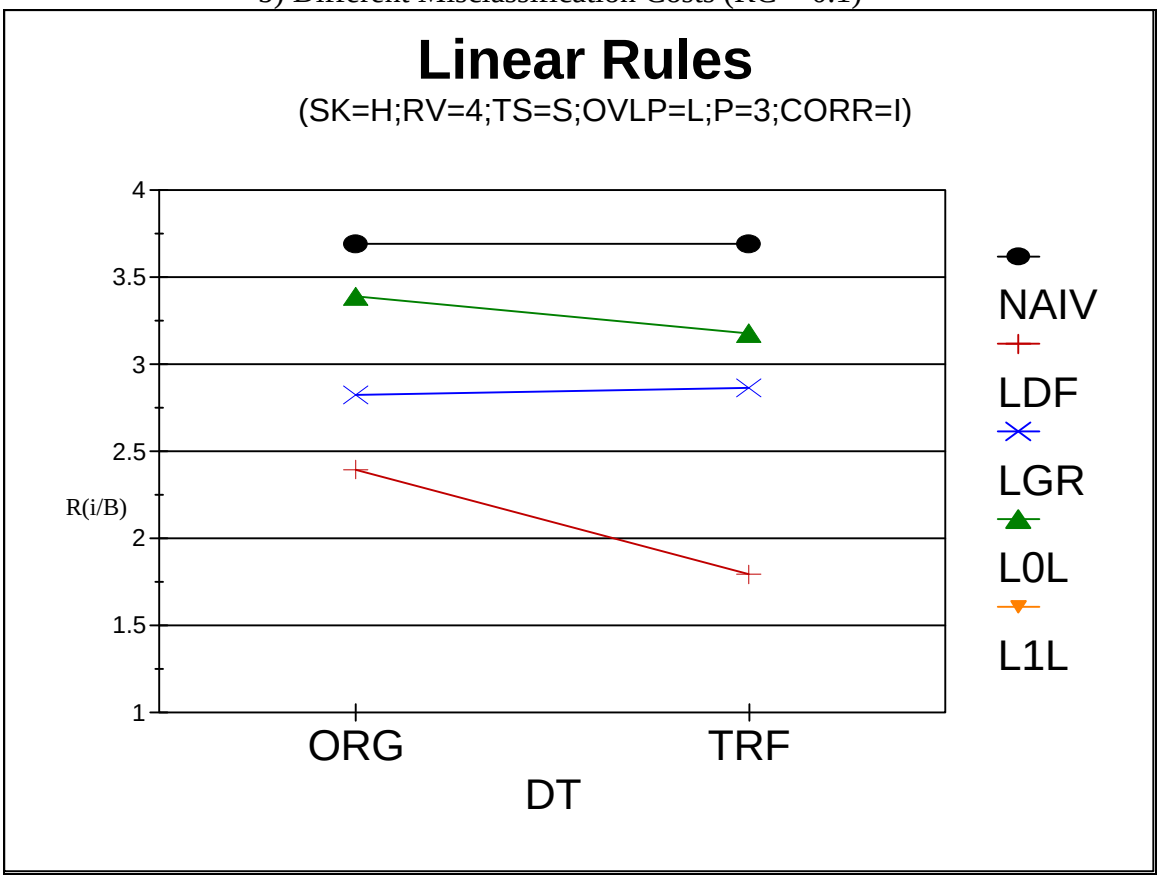
Figure 5: Transformation Effect (DT)

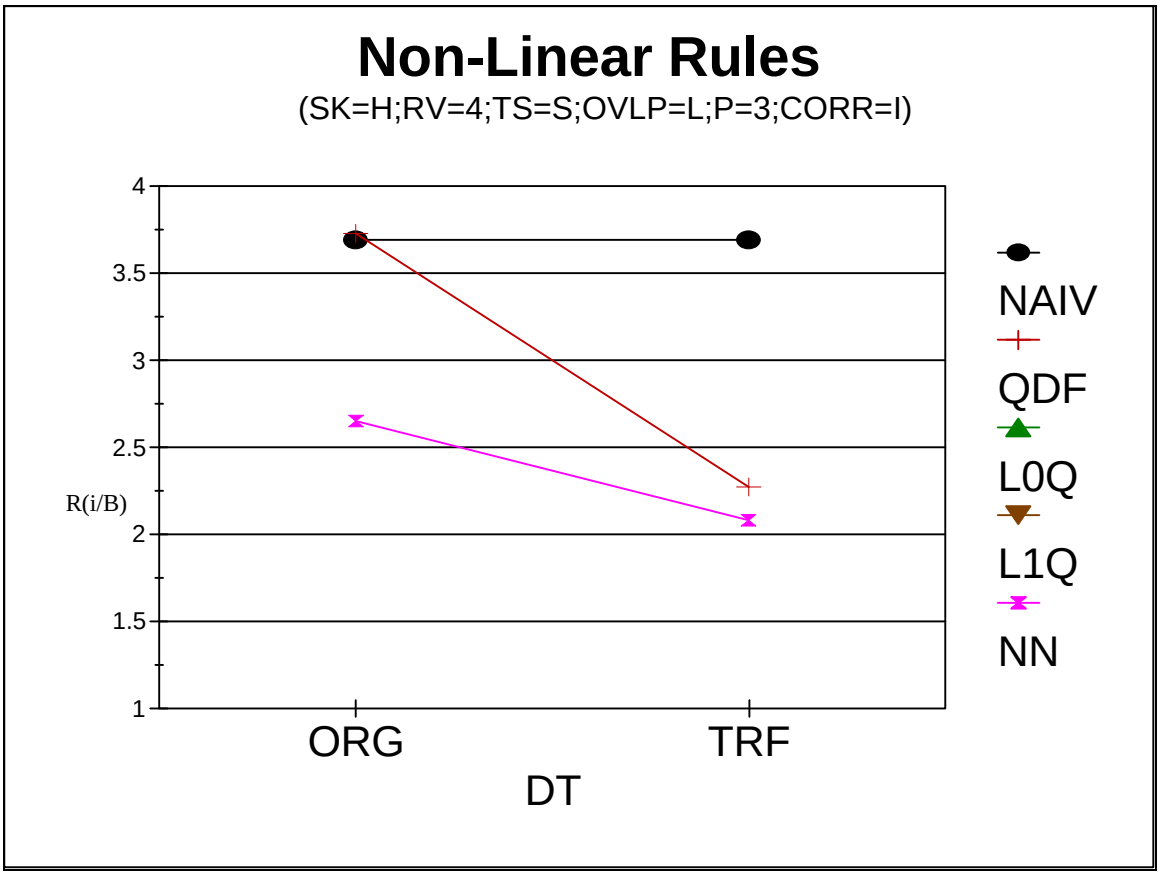
a) Equal Misclassification Costs (RC = 1.0)





b) Different Misclassification Costs (RC = 0.1)





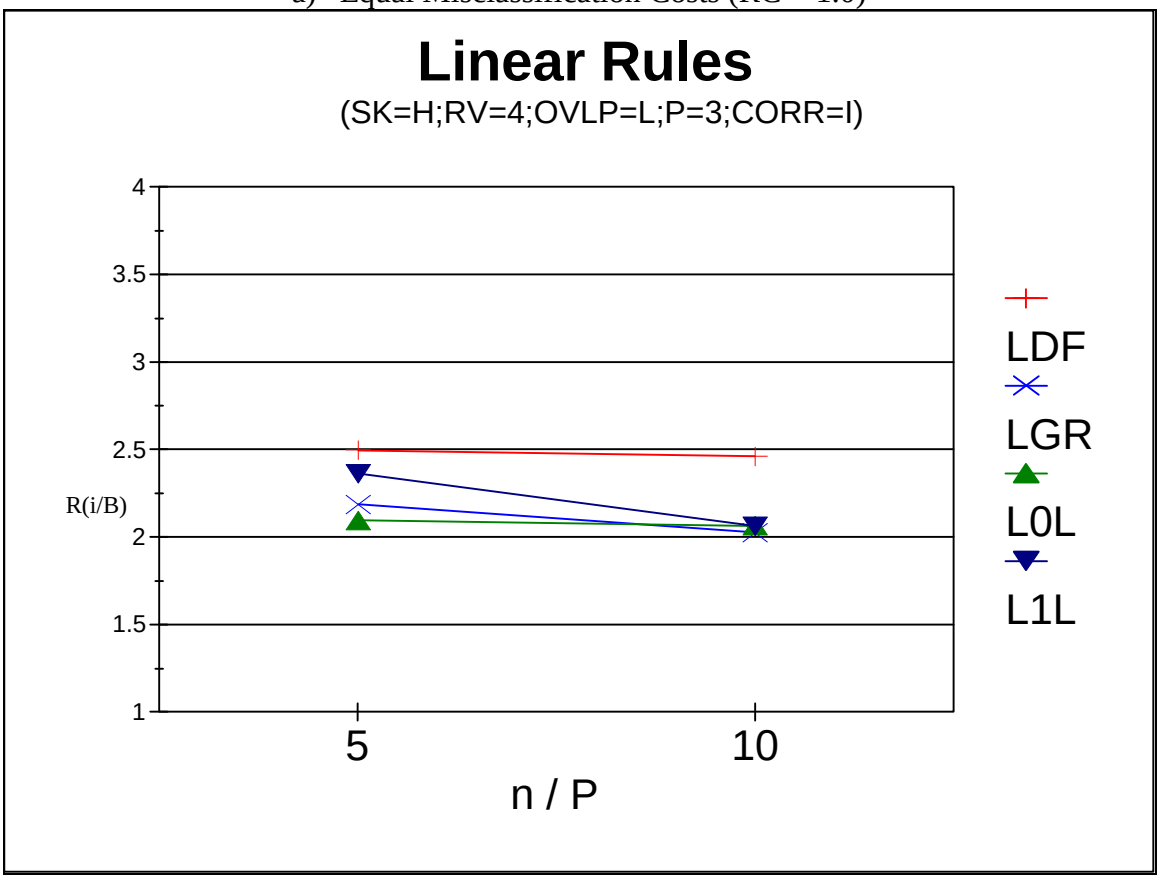
L_0-norm used in this study includes a secondary objective to resolve ties among alternative classification rules with the same training sample misclassification cost. The inclusion of this secondary objective improves the stability of the method, yielding better accuracy on validation samples, especially if the training samples are small.
While the LGR, L0L and L1L appear to be affected most by the training sample size if the misclassification costs differ across groups, this is not typically the case for the other methods, in particular the LDF. For instance, for *RC*=10.0, the changes in $R_{i/B}$ due to reducing the training sample are from 2.43 to 3.13 (LGR), from 1.99 to 2.06 (LDF), from 2.88 to 3.44 (L0L) and from 2.94 to 4.23 (L1L). The effects described in this section are illustrated in Figure 6.

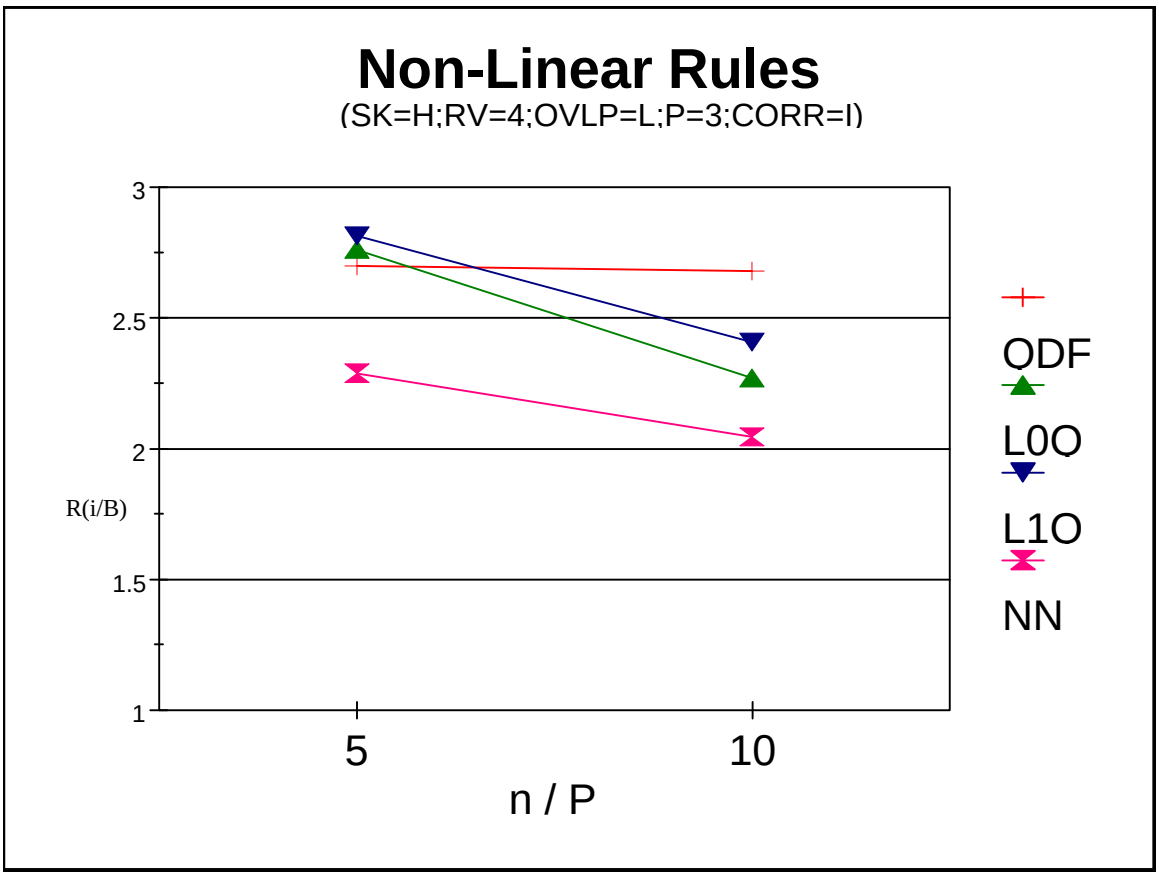
Group Overlap Effect (*OVLP*)
The next secondary experiment uses a repeated measures design with group overlap (*OVLP*) as between-subject factors and *CM* and *RC* as within-subject factors to evaluate the effect of *OVLP*. Group overlap, measured by the expected misclassification rate of the Bayes rule in the case of equal costs, has two levels, *OVLP=H* (dc9) and *OVLP=L* (dc5), corresponding to an optimal misclassification rate of 31.85 and 6.67 percent, respectively (see Table 3). These misclassification rates correspond to the optimal misclassification rates of multivariate normal populations with a common covariance matrix and a Mahalanobis distance of 1 and 3, respectively. These two cases represent standard problems that have been used previously in the literature to illustrate "high" and "low" group overlap (Lachenbruch, Sneeringer and Revo 1973, Konishi and Honda 1990). Dc5 and dc9 are summarized in Table 9. The transformation parameters and distributional characteristics for these data conditions are shown in Tables 5 and 6. The significance of the factor effects, analyzed using MANOVA and ANOVA models, reveals that all main and interaction effects between *CM*, *RC* and *OVLP* are significant at the .0001 level.
Table 8 shows that for all methods, $R_{i/B}$ is lower for *OVLP=H* than for *OVLP=L*, indicating that the estimated expected cost is closer to the expected cost of the Bayes rule for *OVLP=H*. The difference in performance between the various methods tends to decrease as the group overlap increases, especially in the case where the misclassification costs across groups are similar. For instance, whereas for *RC*=1 and *OVLP=L* the $R_{i/B}$ ratios vary between 2.10 ($R_{L0L/B}$) and 2.81 ($R_{L1Q/B}$), these ratios are between 1.17 ($R_{LDF/B}$) and 1.37 ($R_{L0Q/B}$) for *RC*=1 and *OVLP=H*. Other than this observation, there is no clear pattern of how the degree of group overlap favors any particular method. The effect of *OVLP* is illustrated in Figure 7.

Number of Attributes Effect (*P*)
Another secondary experiment considers the effect of varying the number of attributes *P* from 3 (dc5) to 10 (dc10) (see Table 9). The corresponding transformation parameters and distributional characteristics are shown in Tables 5 and 6. The range from 3 to 10 attribute variables is representative of most prior experiments and real life business applications. The secondary experiment use a repeated measures design with *P* as between-subject factors and *CM* and *RC* as within-subject factors. MANOVA and ANOVA analyses reveal that all main and interaction effects between *CM*, *RC* and *P* are significant at the .0001 level.

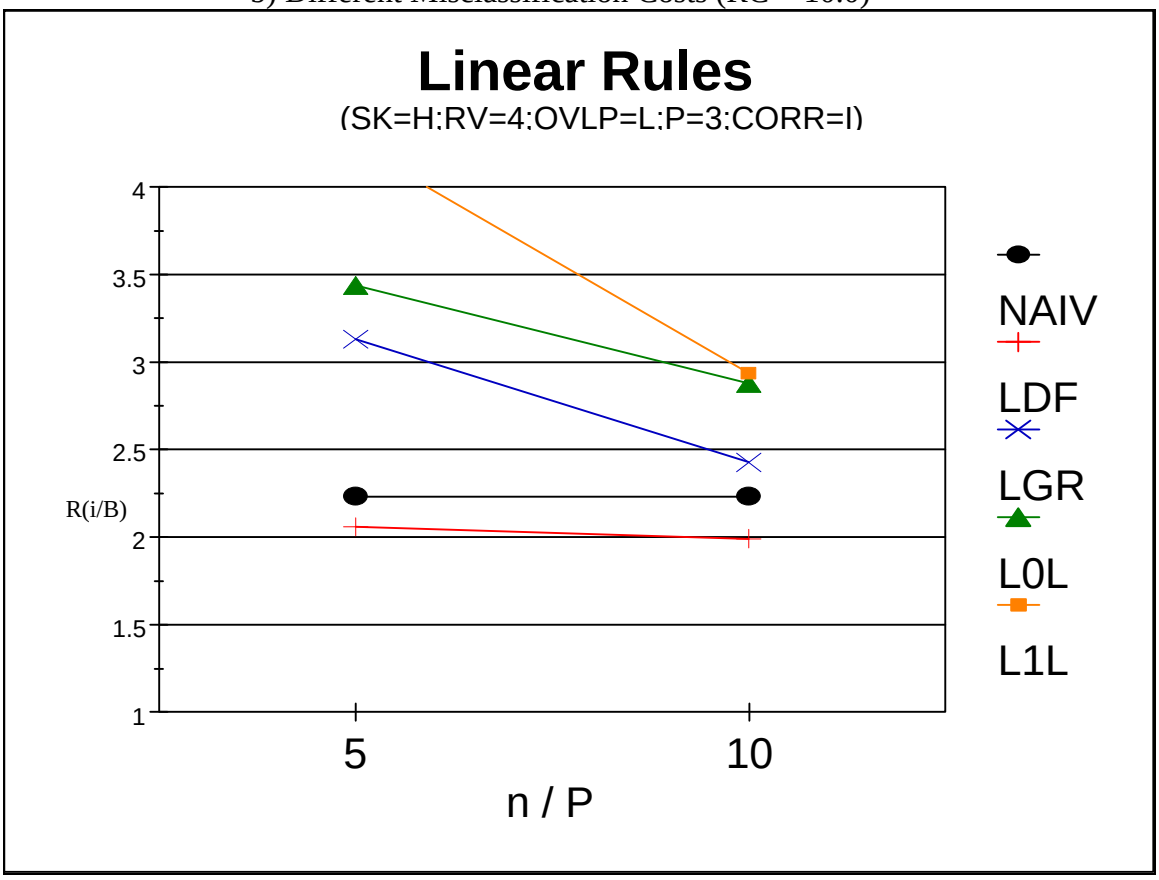
Figure 6: Training Sample Size Effect (TS)

a) Equal Misclassification Costs (RC = 1.0)





b) Different Misclassification Costs (RC = 10.0)



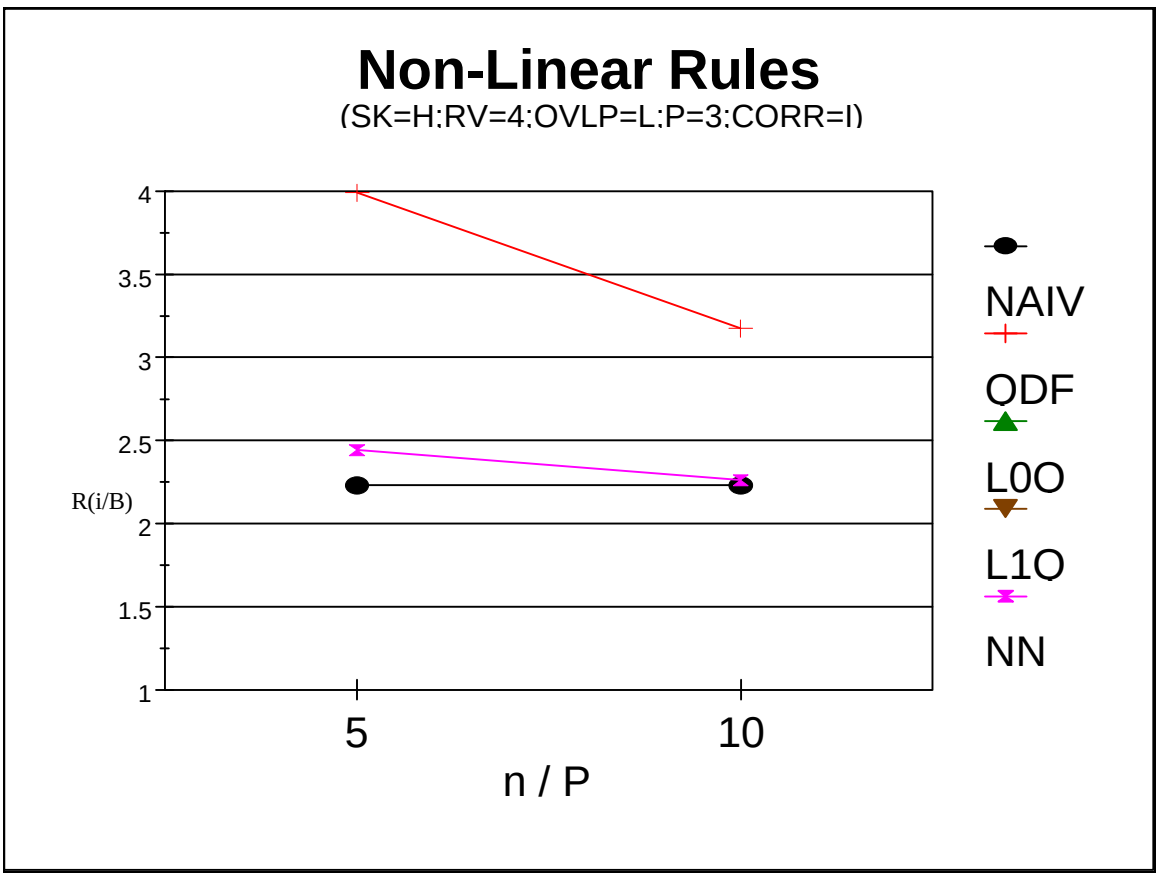
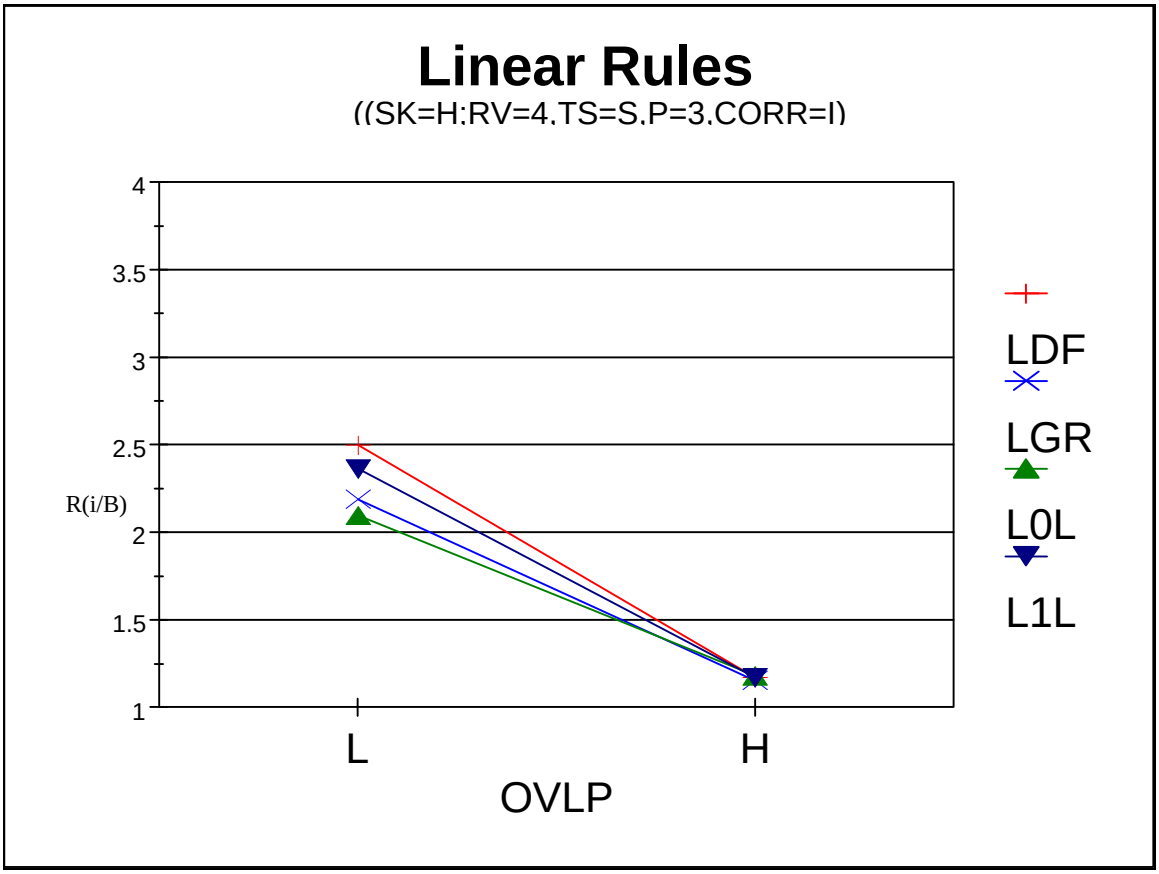


Figure 7: Group Overlap Effect (OVLP)



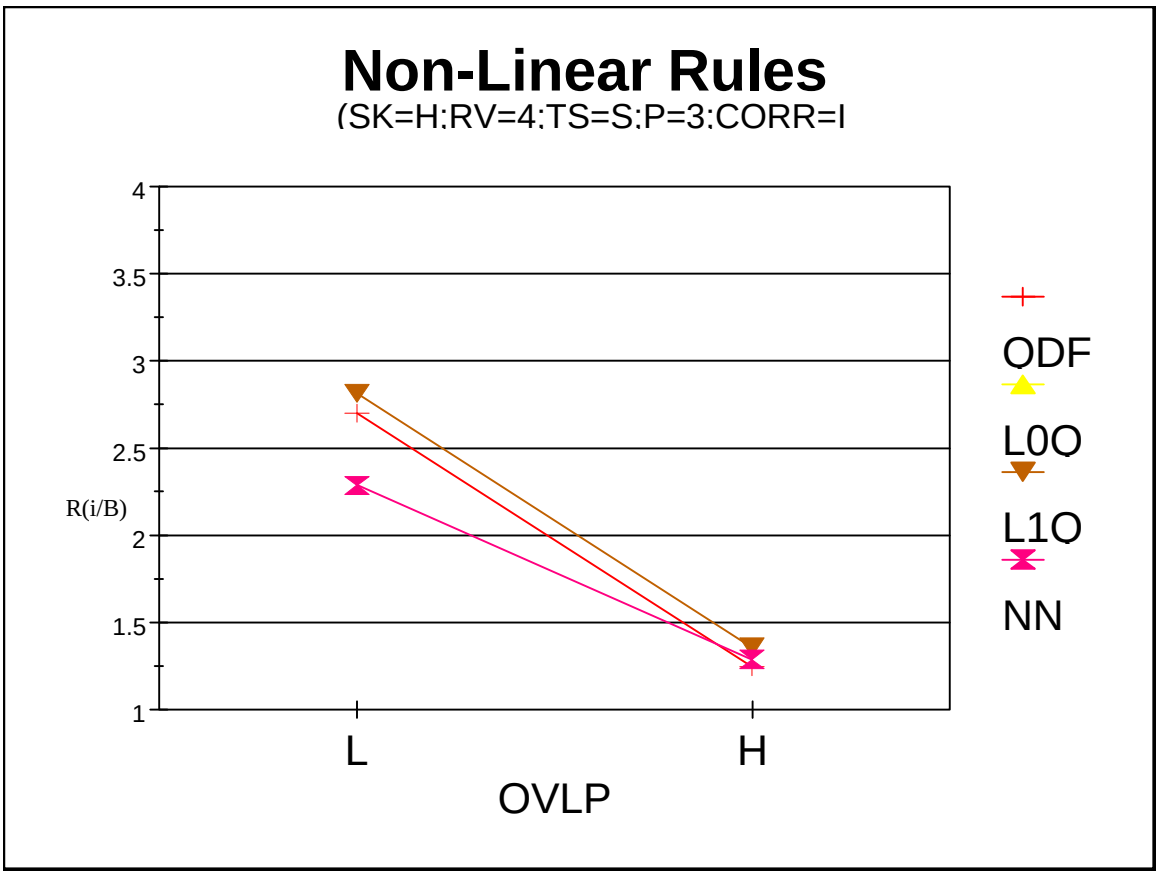
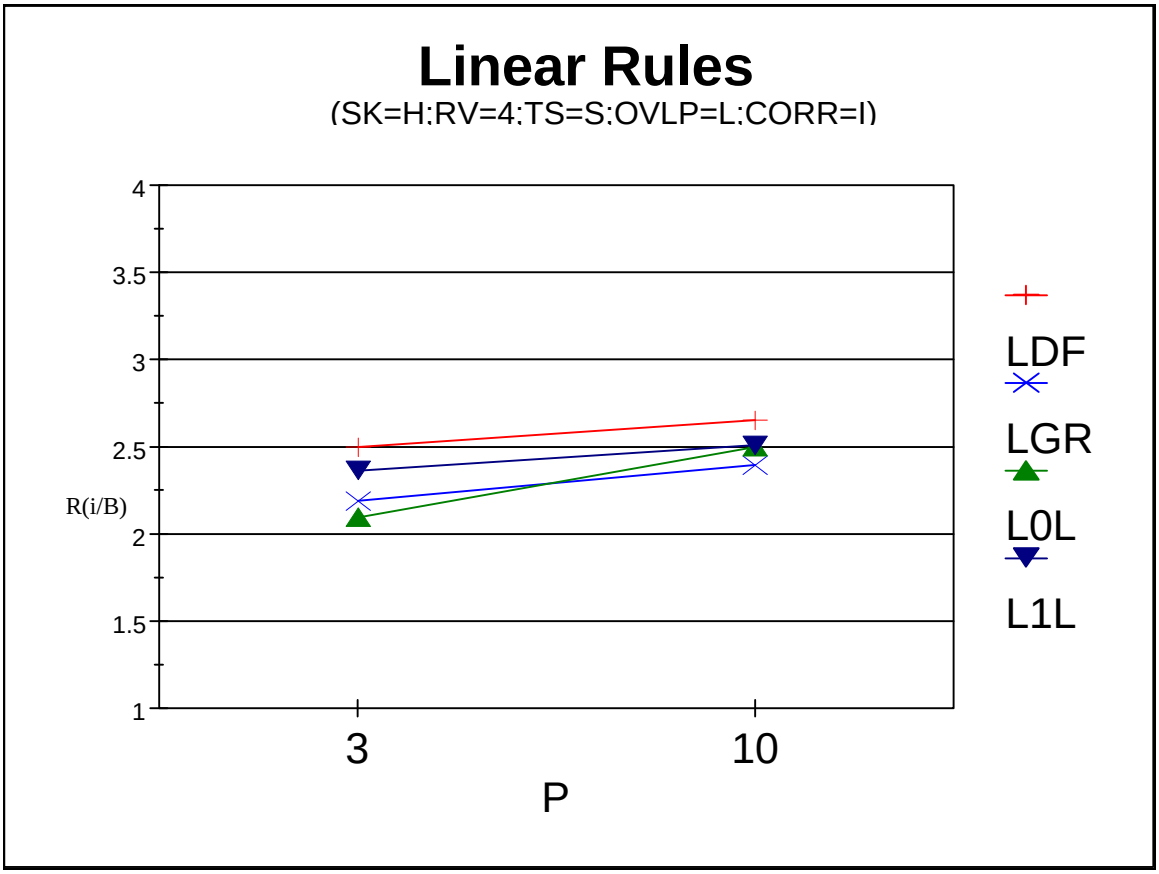
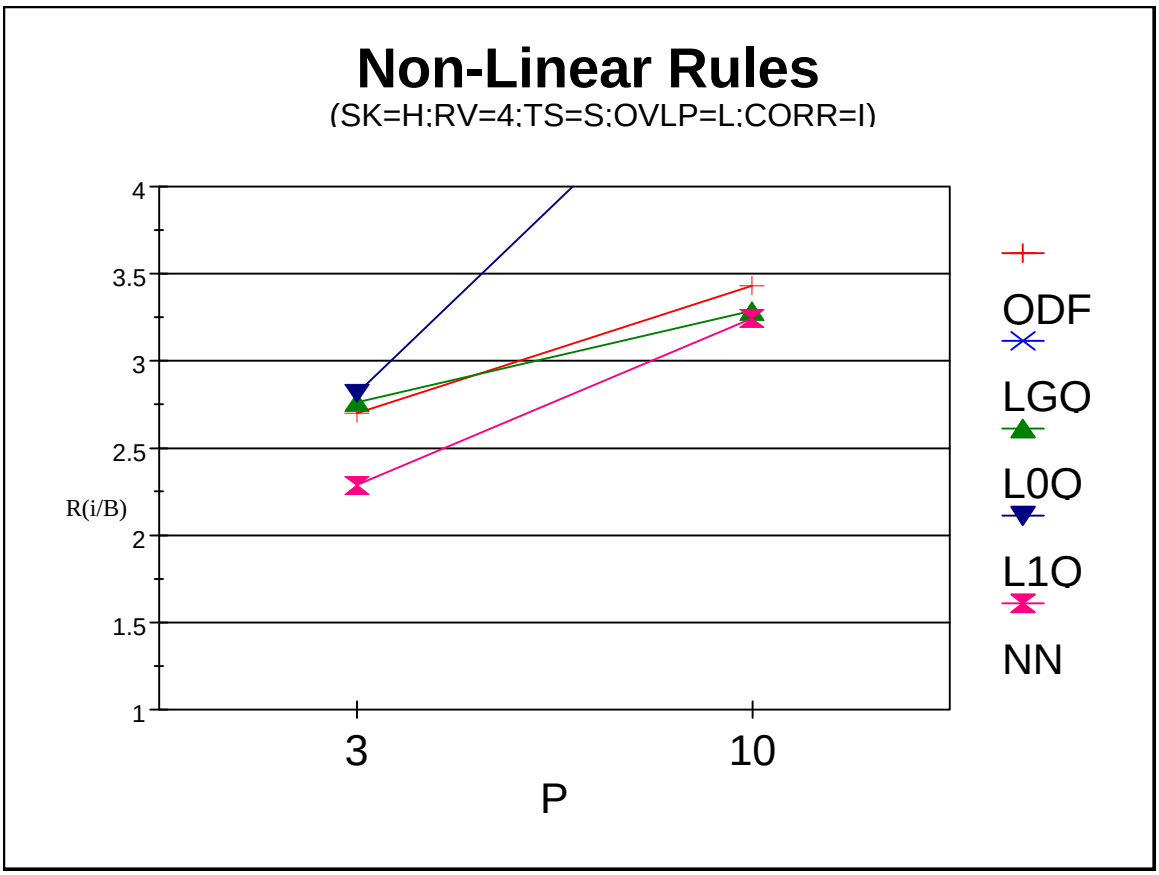


Figure 8: Number of Attributes Effect (P)





In general, the performance of all of the classification methods is closer to that of the optimal rule for problems with 3 attributes than for problems with 10 attributes. The nonlinear methods as well as the L0L tend to perform worse in the case of 10 attributes than for 3 attributes. This effect is illustrated in Figure 8 for the case of *RC*=1.
As shown in Figure 8, the L0L has the lowest $R_{i/B}$ ratio for *P*=3 (2.10, compared with 2.19 for the LGR), but for *P*=10 this ratio increases sharply to 2.50 and the L0L is outperformed by the LGR ($R_{LGR/B}$ = 2.39). The LDF, with an increase from 2.50 to 2.65, and the L1L, with an increase from 2.36 to 2.51 also improved in relative terms. The nonlinear methods deteriorate considerably for *P*=10. For example, $R_{NN/B}$ increases from 2.29 to 3.24 and $R_{QDF/B}$ from 2.70 to 3.43. This effect appears to be the most pronounced when the misclassification costs vary widely across groups.

Correlation Structure Effect (*CORR*)
The effects of the correlation structure is studied next. Again, the basic data condition dc5 is modified, in this case by introducing correlation between the attribute variables. The two levels types of correlation structures (*CORR*) considered are independent attributes (*I*) (dc5) and positively correlated attributes (*C*) (dc11) (see Table 9). *CORR=I* is included because it is the simplest structure. It should be remarked that this condition is by no means common in practice, and in most business and economics problems the attributes tend to be correlated. However, it is useful to understand how the classification methods perform in the simplest case, before attempting to establish general results for more realistic conditions with complex correlation structures. The condition *CORR=C* is defined as follows. Two of the three attributes (X_1 and X_2) are strongly correlated (ρ_{12}=0.8), and the third attribute (X_3) is moderately correlated with the X_1 and X_2 (ρ_{13}=ρ_{23}=0.4). Although many other correlation structures could have been selected, positive correlations are common in business problems. An exhaustive analysis of how different correlation structures affect classification performance reaches beyond the scope and objectives of this study. This experiment uses a repeated measures design with *CORR* as between-subject factors and *CM* and *RC* as within-subject factors.
The generation of positive correlated attributes is based on the multivariate log-normal distribution (Johnson and Kotz 1972). The variables corresponding to the attributes in G_j, Y_{j1}, Y_{j2}, Y_{j3}, are generated as follows,

$$Y_{11} = \exp(b_1 + c_{1,11} Z_{11})$$
$$Y_{12} = \exp(b_1 + c_{1,21} Z_{11} + c_{1,22} Z_{12})$$
$$Y_{13} = \exp(b_1 + c_{1,31} Z_{11} + c_{1,32} Z_{12} + c_{1,33} Z_{13})$$
$$Y_{21} = \exp(b_2 + c_{2,11} Z_{21})$$
$$Y_{22} = \exp(b_2 + c_{2,21} Z_{21} + c_{2,22} Z_{22})$$
$$Y_{23} = \exp(b_2 + c_{2,31} Z_{21} + c_{2,32} Z_{22} + c_{2,33} Z_{23})$$

where the Z_{ij} IIIare independent standard normal random variables and the parameters b_j and $c_{i,jk}$ are selected in order to achieve the desired distributional characteristics. The parameters used in this study to generate dc11 are presented in Table 10.

TABLE 10: TRANSFORMATION PARAMETERS, SECONDARY EXPERIMENT WITH CORRELATED ATTRIBUTES

Group j	b_j	$c_{j,11}$	$c_{j,12}$	$c_{j,13}$	$c_{j,22}$	$c_{j,23}$	$c_{j,33}$
1	-1.2087	1.1651	1.0271	0.5500	0.6588	0.1653	0.9466
2	0.9046	0.6103	0.5051	0.3426	0.2719	0.0835	0.5400

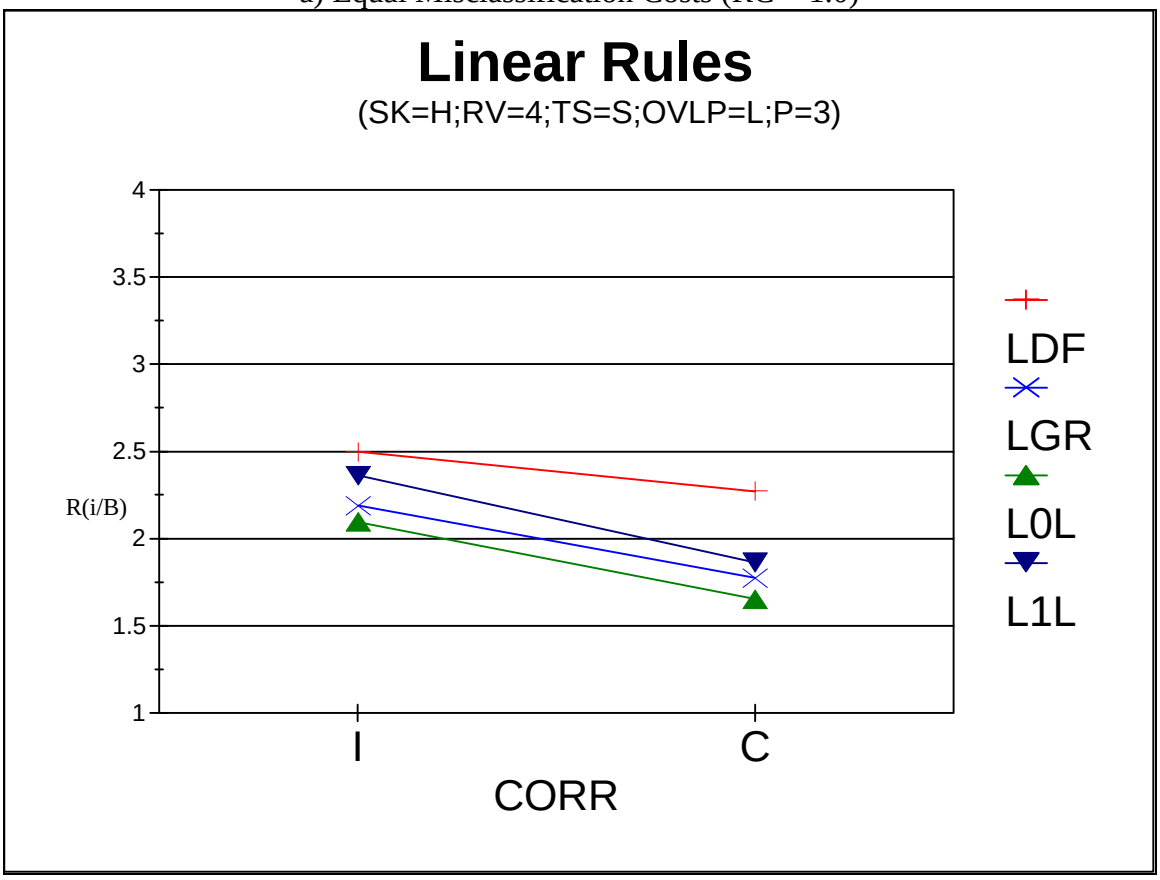
The MANOVA and ANOVA tests reveal that all main and interaction effects between *CM*, *RC* and *CORR* were significant at the .0001 level. Generally, as long as the misclassification cost ratio is moderate (*RC*=0.5, 1 or 2), each method tends to approximate the Bayes rule more closely if the attributes are correlated than if they are independent. However, the reverse holds if *RC*=0.1, *i.e.*, if the misclassification cost of G_2 is ten times higher than that of G_1, as shown by the high $R_{i/B}$ ratios in the left-most column of Table 8. The LDF benefits the least with the presence of positive correlations. For instance, for RC=1 the L0L, LGR and L1L were only slightly better than the LDF in the condition with independent attributes (dc5) (RL0L/B=2.10, RLGR/B=2.19, RL1L/B=2.36, RLDF/B=2.50). However in the condition with correlated attributes (dc11) the LDF was clearly inferior (RL0L/B=1.65, RLGR/B=1.77, RL1L/B=1.86, RLDF/B=2.27). On the other hand, the nonlinear classification methods yield substantially improved results for dc11. For example, again comparing the results for dc5 with dc11 in the case of *RC*=1, introducing positive correlation between the attributes reduces the $R_{i/B}$ ratios from 2.70 to 2.14 (QDF) and from 2.29 to 1.87 (NN). The correlation effect is illustrated in Figure 9 for RC=1 and RC=0.1.
As seen from Figure 9 and Table 8, in the case of RC=1, the L0L and the LGR gives the best results for both *CORR=I* and *CORR=C*. Comparing the parametric methods, the LDF performs better than the QDF for *CORR=I*, but the reverse is true for *CORR=C*. When *RC*=0.1, the LDF gives the best results for *CORR=I*, but for *CORR=C* the NN method ranks first. Interestingly, while for *RC*=0.1 most of the $R_{i/B}$ ratios increase as a result of introducing correlation, $R_{QDF/B}$ decreases from 3.73 to 3.27.

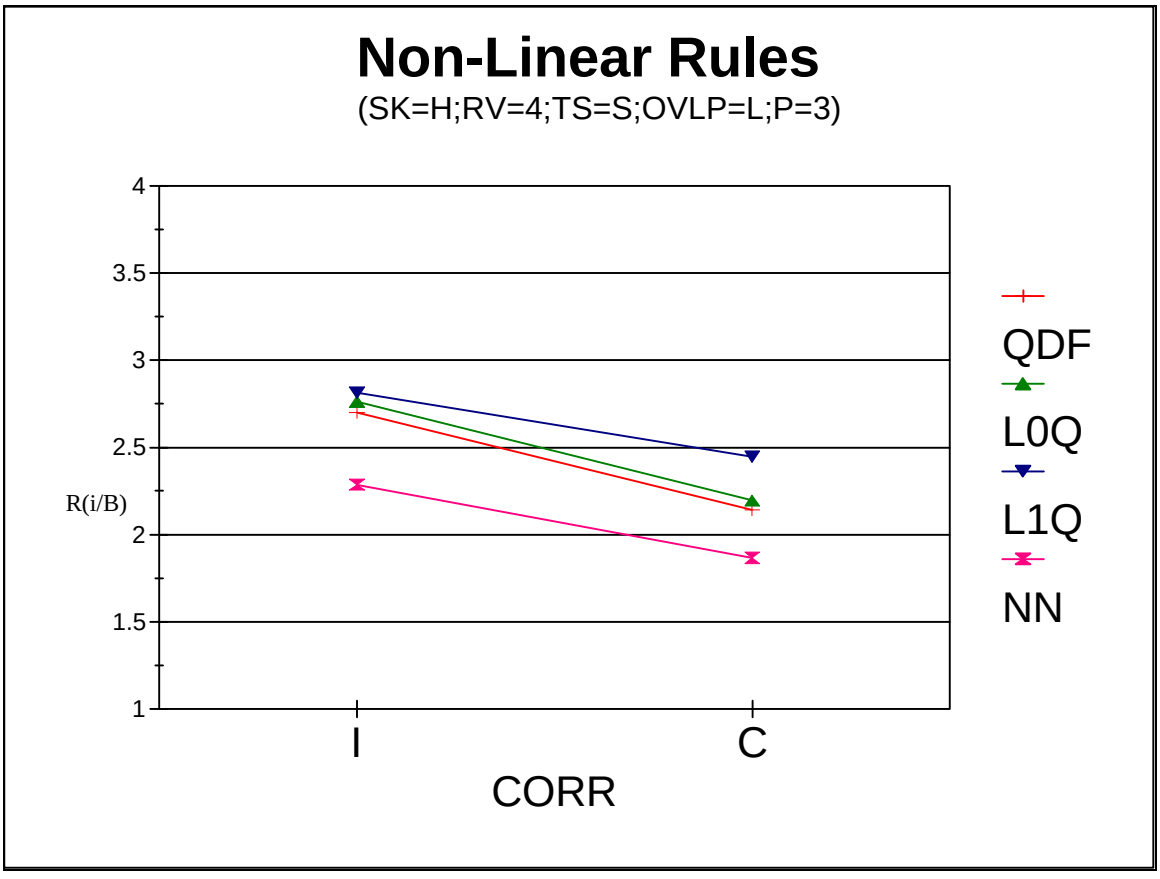
CONCLUSIONS

This study reveals that both the ratio of misclassification costs across groups and the attribute skewness have a clear impact on the relative performance of different misclassification methods. If the misclassifications costs differ widely across groups, the LDF is favored over the other methods considered, notably over logistic regression methods. For this data condition, the MP-based methods with objective function weights proportional to the misclassification costs perform poorly. An analysis of the group-specific rates in the simulation experiment reveals that, in the case of unequal costs across groups, the MP-based methods consistently yield higher misclassification rates for the group with the higher costs than the optimal rule, and lowest misclassification rates for the group associated with the lower cost. This finding suggests that weights that are larger than proportional to the ratio of misclassification costs may lead to better results for the MP-based methods. This issue should be addressed in future studies.

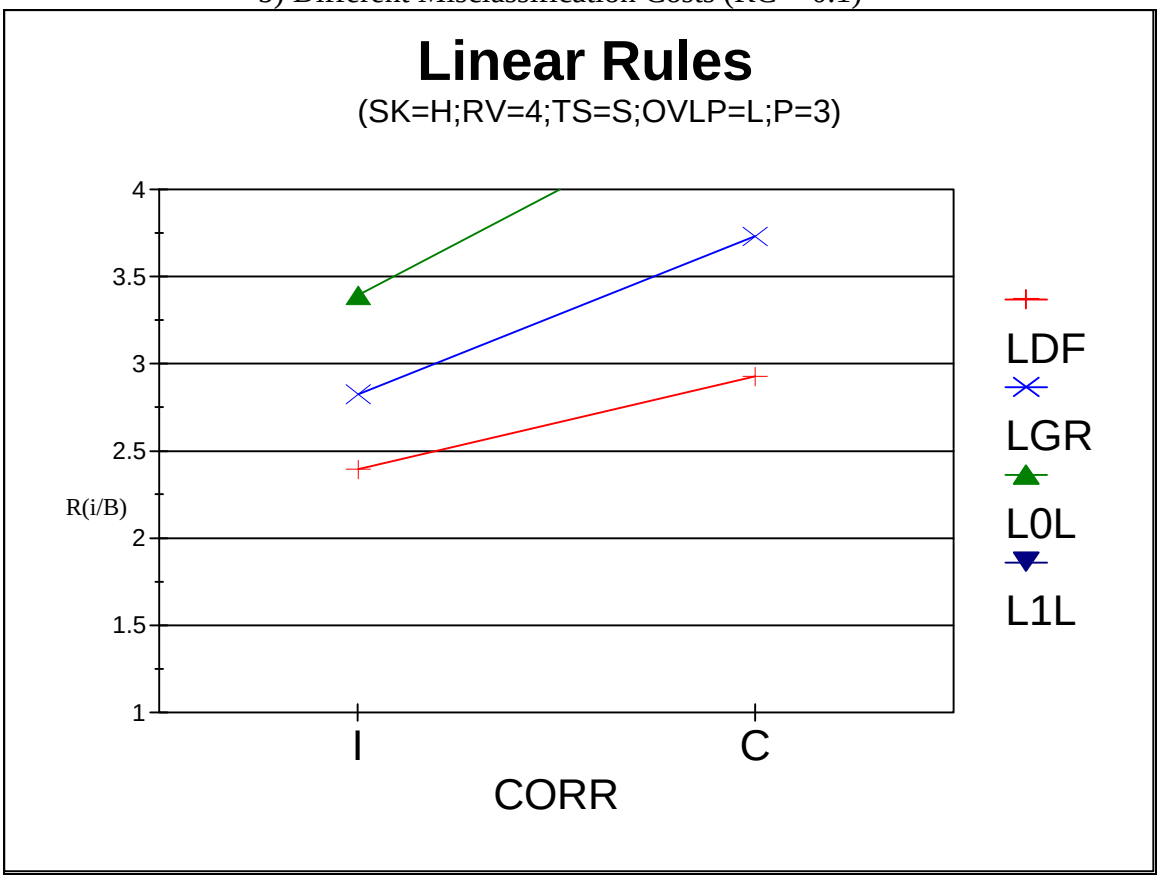
Figure 9: Correlation Structure Effect (CORR)

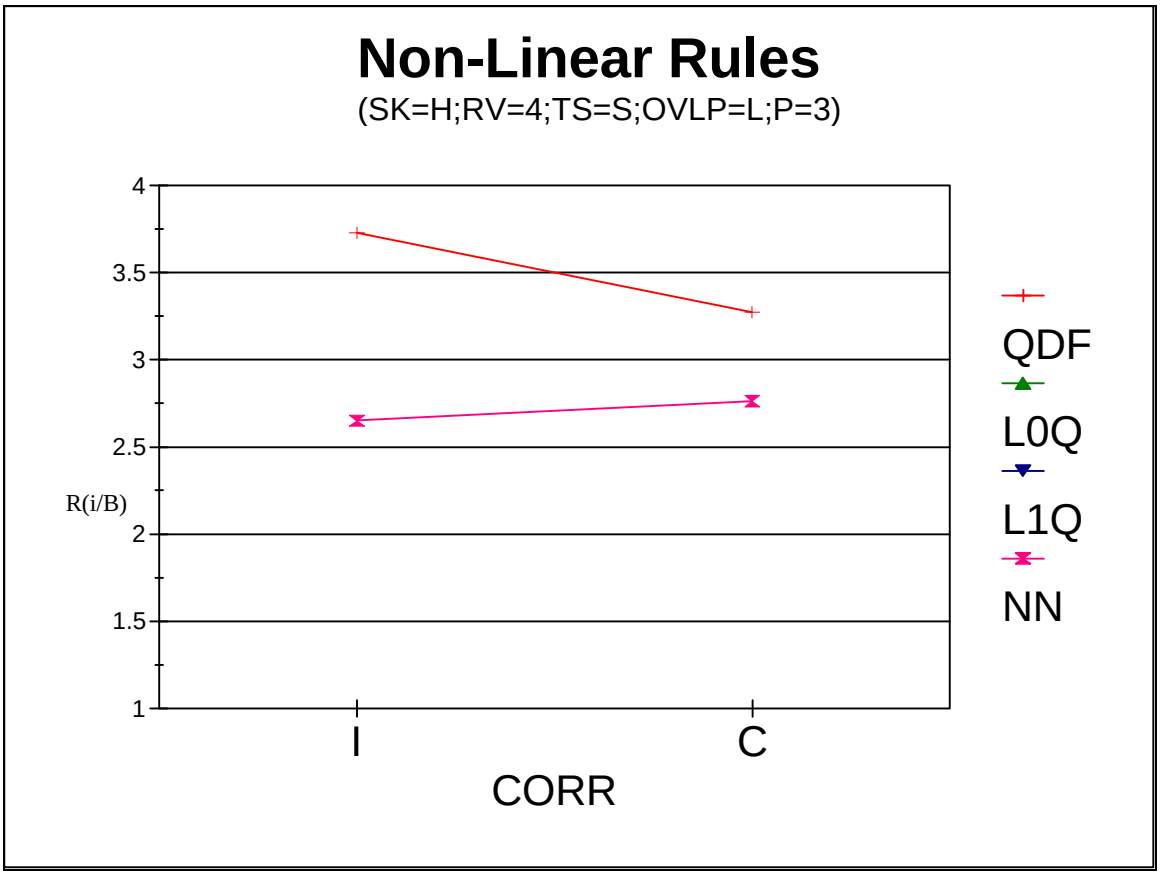
a) Equal Misclassification Costs (RC = 1.0)





b) Different Misclassification Costs (RC = 0.1)





If the differences in misclassification costs across groups are not strong, highly skewed attribute distributions favor logistic regression and certain MP-based methods. Under these conditions, the logistic regression tends to give the best results if the training sample is large. If the training sample is small and the problem has few attributes, the L_0-norm MP-based methods used in this study, with a secondary objective to break ties, appear to give the best results. While the use of data transformations to reduce the degree of skewness can improve the performance of all classification methods, the relative gains achieved by the LDF and QDF tend to exceed those of the remaining classification methods, at least for the data conditions analyzed in this study.

In the majority of data conditions analyzed, methods using linear classification rules tend to yield better results than nonlinear classification methods. The only data condition analyzed in this study for which nonlinear methods are clearly found to be superior is characterized by moderate skewness and high ratios of attribute variances across groups. This condition is particularly favorable for the QDF and the NN methods. When the misclassification costs are similar across groups the QDF is the preferred method. When these costs are strongly different the relative performance of the QDF and the NN depends on whether or not it is possible to achieve important improvements over the NAIVE rule which assigns all observations to the group with the highest cost. For problems where it is difficult to beat the NAIVE rule, the QDF gave the best results. For problems where, the NAIVE rule was not competitive, the NN was the best performing method.

REFERENCES

Agrawala, A. K. 1977. Machine Recognition of Patterns, New York, NY: IEEE Press.

Altman, E., Haldeman, R. and Narayanan, P. 1977. Zeta Analysis: A New Model to Identify Bankruptcy Risk of Corporations. Journal of Banking and Finance,1: 29-54.

Altman E., Avery, R. Eisenbeis, R. and Sinkey, J. 1980. Applications of Classification Techniques in Business, Banking, and Finance, Greenwich, CT: JAI Press.

Anderson, J. A. 1972. Separate Sample Logistic Discrimination. Biometrika, 59: 19-35.

Breiman, L., Meisel, W., and Purcell, E. 1977. Variable Kernel Estimates of Multivariate Densities. Technometrics, 19: 135-144.

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

Clarke, W. R., Lachenbruch, P. A. and Broffit, B. 1979. How Nonnormality Affects the Quadratic Discriminant Function. Communications in Statistics - Theory and Methods, A8: 1285-1301.

Duarte Silva, A. P. and Stam, A. 1994. Second Order Mathematical Programming Formulations for Discriminant Analysis. European Journal of Operational Research, 72: 4-22.

Duarte Silva, A. P. and Stam, A. Forthcoming. A Mixed Integer Programming Algorithm for Minimizing the Training Sample Misclassification Cost in Two-Group Classification. Annals of Operations Research.

Efron, B. 1975. The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis. Journal of the American Statistical Association, 70: 892-898.

Eisenbeis, R.. 1977. Pitfalls in the Application of Discriminant Analysis. Journal of Finance, 32: 875-900.

Enas, G. G. and Choi, S. C. 1986. Choice of the Smoothing Parameter and Efficiency of *k*-Nearest Neighbor Classification.. Computers and Mathematics with Applications, 12A: 235-244.

Erenguc, S. S. and Koehler, G. J. 1990. Survey of Mathematical Programming Models and Experimental Results for Linear Discriminant Analysis. Managerial and Decision Economics, 11: 215-225.

Fatti, L. P., Hawkins, D. M. and Raath, E. L. 1982. Discriminant analysis. In D.M. Hawkins (Ed.), Topics in Applied Multivariate Analysis: 1-71. Cambridge: Cambridge University Press.

Fisher, R. A. 1936. The Use of Multiple Measurements in Taxonomy Problems. Annals of Eugenics. 7: 179-188.

Fix, E. and Hodges, J. L. 1951. Discriminatory Analysis - Nonparametric Discrimination: Consistency Properties. Report No 4. Randolph Texas: U.S. Air Force School of Aviation Medicine. (Reprinted as pp. 261-279 of Agrawala. 1977)

Gessaman, M. P. and Gessaman, P. H. 1972. A Comparison of Some Multivariate Discrimination Procedures.. Journal of the American Statistical Association, 67: 468-472.

Gibbons, Dianne I., McDonald G. C. and Gunst, R. F. 1987. The Complementary Use of Regression Diagnostics and Robust Estimators. Naval Research Logistics, 34: 109-131.

Glorfeld, L. W. and Olson, D. L. 1982. Using the L_1-Metric for Robust Analysis of the Two Group Discriminant Problem. In Proceedings of the American Institute of the Decision Sciences: 297-298.

Glover, F. 1990. Improved Linear Programming Models for Discriminant Analysis. Decision Sciences, 21: 771-785.

Hand, D. J. 1982. Kernel discriminant analysis, New York, NY: Research Studies Press.

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

Johnson, J. R., Leitch, R. A. and Neter, J. 1981. Characteristics of Errors in Accounts Receivable and Inventory Audits. The Accounting Review, 41: 270-293.

Johnson, N. L. and Kotz, S. 1972. Distributions in Statistics: Continuous Multivariate Distributions, New York, NY: Wiley.

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

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

Konishi, S. and Honda, M. 1990. Comparison of Procedures for Estimation of Error Rates in Discriminant Analysis under Nonnormal Populations. Journal of Statistical Computing and Simulation, 36: 105-115.

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

Liittschwager, J. M. and Wang, C. 1978. Integer Programming Solution of a Classification Problem. Management Science, 24: 1515-1525.

Marks, S. and Dunn, O. J. 1974. Discriminant Functions When Covariance Matrices are Unequal. Journal of the American Statistical Association, 69: 555-559.

McCullagh, P. and Nelder, J. A. 1989. Generalized Linear Models, (2nd ed.), London: Chapman and Hall.

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

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

Parzen, E. 1962. On the Estimation of a Probability Density Function and the Mode. Annals of Mathematical Statistics, 33: 1065-1076.

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

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

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

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

Rudolph, P. M. and Karson, M. 1988. The Effect of Unequal Priors and Unequal Misclassification Costs on MDA. Journal of Applied Statistics, 15: 69-82.

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

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

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

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.

Van Ness, J. W. 1979. On the Effects of Dimension in Discriminant Analysis for Unequal Covariance Populations. Technometrics, 21: 119-127.

Wald, A. 1939. Contributions to the Theory of Statistical Estimation and Testing Hypothesis. Annals of Mathematical Statistics, 10: 299-326.

Wald, A. 1944. On a Statistical Problem Arising in the Classification of an Individual into One of Two Groups. Annals of Mathematical Statistics, 15: 145-162.

Wald, A. 1949. Statistical Decision Functions. Annals of Mathematical Statistics, 20: 165-205.

APPENDIX A: Accuracy Criteria and Normalization Scheme Used in the MP-Based Classification Methods

A1. Methods Based on L_1-Norm Criteria

The L_1-norm criterion was the first one proposed in the MP literature (Koford and Groner 1966; Smith 1968), as well as among the simplest and most widely used ones. The L_1-norm distance optimization criterion used in our experiments (L1L, L1Q) is defined as follows:

$$\text{Min } Z = C_1 \, S_{G1,i\rightarrow 2}|f(\mathbf{b},\mathbf{x}_i)-c| + C_2 \, S_{G2,i\rightarrow 1}|f(\mathbf{b},\mathbf{x}_i)-c|,$$

where $S_{G1,i\rightarrow 2}$ and $S_{G2,i\rightarrow 1}$ represent the sum over those observations from G_1 that are assigned to G_2, and those from G_2 that are assigned to G_1, respectively. This criterion minimizes the sum of absolute differences, for all misclassified observations, between the classification scores and the threshold value. The components of this criterion can be interpreted as heuristic indications of the "extent" by which observations are misclassified weighted by the appropriate misclassification costs. In our experiments, the parameters of the classification function are normalized by the normalization constraint $n_2\Sigma_{G1}[f(\mathbf{b},\mathbf{x}_i)-c] + n_1\Sigma_{G2}[c-f(\mathbf{b},\mathbf{x}_i)] = 1$, proposed by Glover (1990). Glover shows that applying this normalization guarantees that the classification rule is non-trivial and invariant with respect to linear transformations of the attribute variables.

A2. Methods Based on the Number of Misclassifications

The implementation of the L_0-norm method used in this paper (L0L, L0Q) minimizes Z_1, the total misclassification cost in the training sample as the primary criterion, and includes Z_2 as a secondary criterion to break ties among those rules that yield identical training sample misclassification costs. Z_1 and Z_2 are defined as follows:

$$\text{Min } Z_1 = \Sigma_{G1,i\rightarrow 2}C_1 + \Sigma_{G2,i\rightarrow 1}C_2$$
$$\text{Min } Z_2 = C_1\Sigma_{G1}[c-f(\mathbf{b},\mathbf{x}_i)] + C_2\Sigma_{G2}[f(\mathbf{b},\mathbf{x}_i)-c]$$

Z_2 is an L_1-norm measure that simultaneously minimizes the sum of the absolute differences between the classification scores and the threshold value for the misclassified observations and maximizes the absolute differences for the observations that are correctly classified. The components of Z_2 are weighted by C_1 and C_2. The classification function parameters are normalized by enforcing the following constraints: $|b_l| \leq b^0_l$, (l=1,...,t); $|c| \leq 1$; $|c|=1$ or $|b_{l'}|=b^0_{l'}$ for at least one $l=l'$, where the b^0_l are appropriately chosen constants (Liitschwager and Wang 1978). In this paper, the b^0_l were set to the absolute values of the corresponding coefficients of the LDF (linear rules) or QDF (quadratic rules), which were first normalized by setting the absolute value of the threshold value to 1. Duarte Silva and Stam (forthcoming) remark that computational considerations play a major role in the choice of secondary criteria and normalization schemes for methods that minimize a primary criterion based on the number of misclassifications, and describe an algorithm that computes classification rules that optimize these primary and secondary criteria for moderate size data sets. A software implementation of this algorithm is available upon request from these authors.