Adv. & Appl. in Stat. **7**(2) (2007), 255-280

ASSESSMENT AND COMPARISON OF DISCRIMINANT ANALYSIS METHODS AND NON-PARAMETRIC ERROR ESTIMATORS WITH MIXED DATA

FERRAN GALÁN[1], **FRANCESC OLIVA**[2] **and JOAN GUÀRDIA**[1]

[1]Department of Methodology of Behavior Sciences
School of Psychology
Universitat de Barcelona
Barcelona, Spain
e-mail: fgalan@ub.edu

[2]Department of Statistics
School of Biology
Universitat de Barcelona
Barcelona, Spain

Abstract

In the field of social sciences and health, it is frequent that the information available comes from mixed data, i.e., a mixture of qualitative and quantitative variables. In this context, it is not easy to discriminate patterns: published studies show that no method is superior to the rest and it seems risky to base the decision solely by means of the error of the training data. In the present study we have identified interactive behaviors between the variables distribution and the classification error of several discriminant rules, as well as the bias and the mean square error from non-parametric error estimators. The protocol of the simulation study, dealing with mixed random vectors with a controlled dependence structure, approaches both problems at once and permits to extract relevant information.

2000 Mathematics Subject Classification: 62H12, 62H30, 68T10.

Keywords and phrases: discriminant analysis, non-parametric error estimators, mixed data.

Received January 21, 2007

I. Introduction

Discriminating patterns is a task widely used in decision making processes. There are numerous studies in the fields of health sciences and social sciences which include discriminant methods with diagnostic aims [18], [25] and [28]. Particularly in these fields, pattern recognition problems require the inclusion of both qualitative and quantitative variables, i.e., mixed random vectors $\mathbf{x} = (\mathbf{x}_Q, \mathbf{x}_q)$, where $\mathbf{x}_Q$ is the quantitative variable vector and $\mathbf{x}_q$ is the qualitative one. Even knowing the multivariate distribution of $\mathbf{x}_Q$, the joint distribution of $\mathbf{x}$ is usually unknown. The problem has basically been ignored until the 1970's and in most cases the studies have focused on the robustness of some classic methods dealing with qualitative variables as quantitative ones.

The behavior of Fisher's *linear discriminator* [17] (LD) on discrete and mixed data has been described in detail by Krzanowski [32]. Linearity is the main disadvantage, which manifests itself in the problem of inversion [36]. Several studies show that LD can tolerate distributions slightly asymmetric, with heavier tails than usual, and normal mixtures, as well as non-severe violations of the covariance matrices equality hypothesis [2], [3], [6], [9] and [33]. Schmitz et al. [40, 41] describe that the dependence structure among quantitative variables has a greater impact than that of the qualitative ones, and point out the poor results obtained if the covariance matrices are very different. Likewise, they describe unsatisfactory results when there is a strong dependence between quantitative and qualitative variables, except when the dependence structure is similar for every class.

The *quadratic discriminant* (QD) [42] has been suggested as an alternative when the covariance matrices are very different, even when the normality hypothesis is not met. According to Clarke and Subrahmaniam [7], QD is relatively robust if the distributions are not highly asymmetric. However numerous studies show that a small sample size is decisive, performing in this case worse results than LD, except when the covariance matrices are very different [4], [16], [29], [35], [43]

and [44]. Finally, in the case of mixed models, QD only solves inversions caused by first order interactions (and therefore reflected on the covariance matrix).

The *logistic discriminator* (LOD) has been presented as an alternative to LD, specially in the case of mixed data, due to the wide range of probability models for which it is adequate. Several studies [26], [27] and [37] have shown that parameter estimates of the linear function obtained with LD have a higher bias than those obtained with LOD for mixed data. Notwithstanding, the same study by Halperin et al. [23] and others like Ameniya and Powell [1] and Krzanowski [31] show that the differences in the classification error are slightness. This relative tie and the higher computational cost of LOD make it an unattractive alternative for the mixed data problem.

Studies about applying methods based on kernel functions (PARZEN) and nearest neighbors (k-NN) are scarce. Out of these, the comparative studies by Vlachonikolis and Marriot [43] and Schmitz et al. [41], which compare the performances of LD, QD, k-NN and PARZEN, among others, present contradictory and non-clarifying conclusions.

Another alternative for mixed data is the distance based method (DB) suggested by Cuadras [10] and Cuadras et al. [11]. According to the simulation work of Oliva [38], its behavior depends greatly on the metrics used, as well as on the incorporation of first order interactions among the qualitative variables.

Beyond the mixed data problem, if we focus our attention on tasks of assessment and comparison of discriminant methods, numerous studies have been carried out with real data and they provide us with contradictory results [24]. Because no method is inherently superior to other fact derived from the *Not Free Lunch Theorems* [45] and [46], superiority depends on the random vector distribution, sample size, etc. [12], hence the results of the studies are affected by the characteristics of its own design. It seems then necessary to delve into the knowledge of interactive behaviors [30] which occur among the factors considered in simulation studies. Bearing in mind the points formulated, the simulation protocol carried out in this study intends to:

1. assess and compare different discriminant methods when they are applied to mixed data.

2. assess and compare the behavior of different conditional non-parametric error estimators, identifying interactive behaviors with the variables distribution and with the discriminant method.

II. The DB Method and a Distance Function for Mixed Data

The DB method [10] and [11] is based on proximities between individuals. Given g classes $G_1, \ldots, G_g$ and a distance function d_l defined for class G_l, the proximity measure between the class G_l and the pattern $\mathbf{x}_0 = \mathbf{x}(\omega_0)$ is

$$\phi_l(\mathbf{x}_0) = V_l(\mathbf{x}_{G_l} \mid \mathbf{x}_0) - V_l(\mathbf{x}_{G_l}), \tag{1}$$

where

$$V_l(\mathbf{x}_{G_l}) = \frac{1}{2} \mathrm{E}_{G_l G_l}[d_l^2(\mathbf{x}_{G_l}, \mathbf{x}_{G_l})] \tag{2}$$

and

$$V_l(\mathbf{x}_{G_l} \mid \mathbf{x}_0) = \mathrm{E}_{G_l}[d_l^2(\mathbf{x}_0, \mathbf{x}_{G_l})] \tag{3}$$

are the geometric variability and the geometric variability relative to the $\mathbf{x}_0$ pattern. The DB method assigns ω_0 to G_l if

$$\phi_l(\mathbf{x}_0) = \min_t[\phi_t(\mathbf{x}_0)]. \tag{4}$$

In practice,

$$\hat{V}_l(\mathbf{x}) = \frac{1}{2n_l^2} \sum_{j,j'=1}^{n_l} d^2(\mathbf{x}_{lj}, \mathbf{x}_{lj'}) \tag{5}$$

and

$$\hat{V}_l(\mathbf{x} \mid \mathbf{x}_0) = \frac{1}{n_l} \sum_{j=1}^{n_l} d^2(\mathbf{x}_0, \mathbf{x}_{lj}) \tag{6}$$

are adequate estimates, where $\mathbf{x}_{lj}$ is the sample j of class l with a sample size n_l. Thus, the estimate of the proximity function is

$$\hat{\phi}_l(\mathbf{x}_0) = \frac{1}{n_l}\sum_{j=1}^{n_l} d^2(\mathbf{x}_0, \mathbf{x}_{lj}) - \frac{1}{2n_l^2}\sum_{j,j'=1}^{n_l} d^2(\mathbf{x}_{lj}, \mathbf{x}_{lj'}). \tag{7}$$

The DB method is susceptible of being used for mixed data. Oliva [38] suggests the weighted sum of two square distances, one for the quantitative variables and another for the qualitative ones

$$d^2(\mathbf{x}_i, \mathbf{x}_{i'}) = w_{(Q)}d^2_{(Q)}(\mathbf{x}_i, \mathbf{x}_{i'}) + w_{(q)}d^2_{(q)}(\mathbf{x}_i, \mathbf{x}_{i'}). \tag{8}$$

Likewise, he reveals a reasonable choice of the relative weights:

$$w_{(Q)} = \frac{p}{1\Big/ g\sum_{l=1}^{g} V_{d_Q}(\mathbf{x}_{G_l(Q)})},$$

$$w_{(q)} = \frac{p}{1\Big/ g\sum_{l=1}^{g} V_{d_q}(\mathbf{x}_{G_l(q)})}$$

which perform as 'normalizing' factors [21], where $V_{d_Q}(\mathbf{x}_{G_l(Q)})$ is the geometric variability for the quantitative variables of class G_l and $V_{d_q}(\mathbf{x}_{G_l(q)})$ is the analogous for the qualitative ones. Let us see a proposal of this mixed distance. Suppose that $\mathbf{x} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the random vector of quantitative variables. We can calculate the absolute value distance (AV)

$$d_Q(\mathbf{x}_i, \mathbf{x}_{i'}) = \left\{\sum_{h=1}^{p} | \mathbf{y}_{ih} - \mathbf{y}_{i'h} |\right\}^{1/2}, \tag{9}$$

where $\mathbf{y} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$ is the Mahalanobis transformation that guarantees the invariance for scale changes, besides considering the correlations between the variables. For the estimate of $\boldsymbol{\Sigma}$, we can opt for the joint estimate (pooled covariance matrix) (PC) or the separated estimate of the within covariance matrices (WC).

On the other hand, given q qualitative variables, we can define a similarity index based on the sum of multinomials with first order interactions (SMI)

$$s_{ii'} = \sum_{u=1}^{q} w_u s_{ii'u} + \sum_{u,u'=1,u<u'}^{q} w_{(u,u')} s_{ii'(u,u')}, \tag{10}$$

where

$$s_{ii'\bullet} = z'_{i\bullet} z_{i'\bullet}, \quad z'_{i\bullet} z_{i'\bullet} = \begin{cases} 0 & \text{if } z_{i\bullet} \neq z_{i'\bullet} \\ 1 & \text{if } z_{i\bullet} = z_{i'\bullet} \end{cases}$$

with the restriction $\sum_{u=1}^{q} w_u + \sum_{u<u'} w_{(u,\,u')} = 1$ due to $0 \leq s_{ii'} \leq 1$. Since $q(q-1)/2$ double interactions are identified, which is higher than the number of variables if $q > 3$, then the weights $w_u = 2/3q$, $w_{(u,u')} = 2/3q(q-1)$, $u, u' = 1, ..., q$, are suggested (notice that the weight of interactions is inversely proportional to the number of interactions in which a variable participates), leading to $\sum_{u=1}^{q} w_u = 2/3$ and $\sum_{u<u'} w_{(u,u')} = 1/3$. In order to obtain distances we can apply to (10) the transformation [20]

$$d_q(\mathbf{x}_i, \mathbf{x}_{i'}) = (1 - s_{ii'})^{1/2}. \tag{11}$$

III. Conditional Error Estimators

Let us consider a discriminant rule, where $r(\mathbf{x}) = l$ symbolizes that the individual is assigned to the class G_l. For a training data t, the rule assigning probabilities is defined by $P[r(\mathbf{x}; t) = j \mid \mathbf{x} \in G_l; t]$. Therefore, the G_l class conditional error is

$$ce_l = P[r(\mathbf{x}; t) \neq l \mid \mathbf{x} \in G_l; t] = \sum_{j \neq l}^{g} ce_{lj} = 1 - ce_{ll} \tag{12}$$

and the global conditional error is

$$ce = \sum_{l=1}^{g} \pi_l ce_l, \tag{13}$$

where π_l is the prior probability of belonging to the G_l class.

The unconditional probabilities ue_{lj} are the expected values of ce_{lj}

$$ue_l = \mathrm{E}(ce_l) = P[r(\mathbf{x}; T) \neq l \,|\, \mathbf{x} \in G_l; T]$$

$$= \sum_{j \neq l}^{g} ue_{lj} = 1 - ue_{ll}. \tag{14}$$

Hence the global unconditional error is

$$ue = \sum_{l=1}^{g} \pi_l ue_l. \tag{15}$$

The analytic calculation of these errors, even when we know the probability model, is not easy and we can face insurmountable difficulties. In such cases, or when the probability model is unknown, we have several non-parametric error estimators.

A. Apparent Error (AE)

The apparent error [42] is the proportion of patterns of G_l in t wrongly classified by $r(\mathbf{x}; t)$. It is defined by

$$AE_l = \frac{1}{n_l} \sum_{j=1}^{n_l} I[r(\mathbf{x}_{lj}; t) \neq l], \tag{16}$$

where

$$I[r(\mathbf{x}; t) \neq l] = \begin{cases} 1 & \text{if true} \\ 0 & \text{if false.} \end{cases}$$

The downside is that using the same patterns of t to assess the discriminant rule results in a negative (optimistic) bias.

B. Leave-One-Out (LOO)

The leave-one-out error estimate [34] of the G_l class is

$$LOO_l = \frac{1}{n_l} \sum_{j=1}^{n_l} I[r(\mathbf{x}_{lj};\, t_{(lj)}) \neq l], \tag{17}$$

where $t_{(lj)}$ expresses the exclusion of the pattern $\mathbf{x}_{lj}$ from the training set. Also known as *cross-validation method*, it avoids AE's optimistic bias and is the extreme variation of the *split method*, maintaining the training set size largest as possible. Nevertheless, Glick [19] describes an increase in the mean square error and several other studies show a slightly positive (pessimistic) bias.

C. Leave-One-Out Bootstrap (LOO-B)

The *leave-one-out bootstrap* estimator [13] is defined as the rate of expected *bootstrap* error in an original pattern not included in the *bootstrap* training set. If we obtain B *bootstrap* training sets $(t^1, ..., t^B)$, then the G_l *class error estimate* is defined as

$$Err_l^1 = \sum_{j=1}^{n_l} \sum_{h=1}^{B} \zeta_{jh} I[r(\mathbf{x}_{lj};\, t^h) \neq l] \Big/ \sum_{j=1}^{n_l} \sum_{h=1}^{B} \zeta_{jh}, \tag{18}$$

where

$$\zeta_{jh} = \begin{cases} 1 & \text{if } x_{lj} \not\subset t^h, \\ 0 & \text{if } x_{lj} \subset t^h. \end{cases}$$

LOO-B has revealed a tendency toward pessimistic estimates in simulation studies carried out by Efron and Tibshirani [14].

D. Bootstrap .632 (B.632)

The .632 estimator [13] is a weighted mean between AE and LOO-B

$$Err^{.632} = .368AE + .632Err^1. \tag{19}$$

Several studies show that this ingenious estimator is superior to the previous ones, describing a moderate bias and the smallest mean square error.

E. Bootstrap .632+ (B.632+)

Breiman et al. [5] describe a negative bias of the B.632 estimator in overfitting discriminant rules with $AE = 0$, as is the case of k-NN or RBP methods. In order to solve the bias, Efron and Tibshirani [14] suggest the .632+ estimator

$$Err^{632+} = Err^{632} + (Err^{1'} - AE)\frac{.368 \times .632\hat{R}'}{1 - .368\hat{R}'}, \tag{20}$$

where

$$\hat{R} = \begin{cases} (Err^{1'} - AE)/(\hat{\gamma} - AE) & \text{is } Err^{1}, \hat{\gamma} > AE \\ 0 & \text{otherwise} \end{cases} \tag{21}$$

is the relative overfitting rate, $Err^{1'} = \min(Err^{1}, \hat{\gamma})$, $\hat{\gamma} = \hat{p}_l(1 - \hat{q}_l) + (1 - \hat{p}_l)\hat{q}_l$, $\hat{p}_l$ is the prior probability of G_l and $\hat{q}_l$ is the proportion of patterns classified in G_l. The simulation studies of Efron and Tibshirani [14] show that B.632+ provides estimates with smaller bias.

IV. Method

A. Simulation study

(1) Simulation algorithm

The algorithm is computationally demanding but it permits obtaining the necessary information to give an answer to the goals. It is structured as follows:

1. Generate the training data: random samples of size n_l for each G_l class

$$\left(l = 1, ..., g; \sum n_l = n\right).$$

2. Construct the discriminant rules with the training data.

3. Estimate the conditional error of the discriminant methods by means of the different non-parametric estimators.

4. Random generation of test data, obtaining samples of size n_l for each G_l class.

5. Estimate the conditional error of every rule classifying the test data.

6. Repeat k times the steps 1 to 5.

7. Process the results.

Notice that the double estimate of the conditional error by means of training data and test data let us estimate the bias and the mean square error of each estimator.

(2) Data generation

The study is meant for $g = 2$ classes with $p = 3$ quantitative variables, out of which $p_0 = 2$ is ordinal, and $q = 3$ binary variables. We have chosen $\pi_1 = \pi_2$ prior probabilities, equal sample sizes $n_1 = n_2 = 50$ and $k = 200$.

The mixed data has been generated following Schmitz et al. [41] and Oliva [38], consists in generating $(p + q)$-dimensional multivariate normal vectors and, afterwards, discretizing the p_0 and q variables. The process is as follows:

1. Generate 6-dimensional vectors with distribution $\mathbf{Y}_{G_l} \sim N_6(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_i)$, $l = 1, 2$. Covariance matrices have been defined with the structure:

$$\begin{pmatrix} 1 & \kappa\rho & \rho & \kappa\rho & \rho & \kappa\rho \\ \kappa\rho & 1 & \kappa\rho & \rho & \kappa\rho & \rho \\ \rho & \kappa\rho & 1 & \kappa\rho & \rho & \kappa\rho \\ \kappa\rho & \rho & \kappa\rho & 1 & \kappa\rho & \rho \\ \rho & \kappa\rho & \rho & \kappa\rho & 1 & \kappa\rho \\ \kappa\rho & \rho & \kappa\rho & \rho & \kappa\rho & 1 \end{pmatrix} \times \sigma^2,$$

where σ^2 is the variance parameter, ρ is the correlation coefficient and κ is an indicator of alternate correlations ($\kappa = 1$, $\kappa = -1$). Tables 1 and 2 show the covariance matrices and mean vectors used.

Table 1. Covariance matrices structure

$\boldsymbol{\Sigma}_1$			$\boldsymbol{\Sigma}_2$		
σ^2	ρ	κ	σ^2	ρ	κ
1	.5	1	1	.5	1
1	.5	1	1	.5	–1
1	.5	1	4	.5	1
1	.5	1	4	.5	–1

Table 2. Mean vectors (centroids)

$\boldsymbol{\mu}_1 = (\mu^1\mu^2\mu^3\mu^4\mu^5\mu^6)'$	$\boldsymbol{\mu}_2 = (\mu^1\mu^2\mu^3\mu^4\mu^5\mu^6)'$
$(0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)'$	$(0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)'$
$(0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)'$	$(1.5 \quad 1.5 \quad 1.5 \quad 1.5 \quad 1.5 \quad 1.5)'$

2. Obtain ordinal and binary variables by means of the transformations:

$$X_2 = \begin{cases} 1 & \text{if } Y_2 < \frac{1}{3}\mu_2 \\ 2 & \text{if } \frac{1}{3}\mu_2 \le Y_2 < \frac{4}{9}\mu_2 \\ 3 & \text{if } \frac{4}{9}\mu_2 \le Y_2 < \frac{1}{2}\mu_2 \\ 4 & \text{if } \frac{1}{2}\mu_2 \le Y_2 < \frac{3}{4}\mu_2 \\ 5 & \text{if } Y_2 \ge \frac{3}{4}\mu_2, \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{if } Y_3 < -\frac{1}{2}\mu_3 \\ 2 & \text{if } -\frac{1}{2}\mu_3 \le Y_3 < -\frac{1}{8}\mu_3 \\ 3 & \text{if } -\frac{1}{8}\mu_3 \le Y_3 < \frac{1}{3}\mu_3 \\ 4 & \text{if } \frac{1}{3}\mu_3 \le Y_3 < \frac{1}{2}\mu_3 \\ 5 & \text{if } Y_3 \ge \frac{1}{2}\mu_3, \end{cases}$$

$$X_4 = \begin{cases} 0 & \text{if } Y_4 < \frac{3}{2}\mu_4 \\ 1 & \text{if } Y_4 \geq \frac{3}{2}\mu_4, \end{cases}$$

$$X_5 = \begin{cases} 0 & \text{if } Y_5 < \frac{1}{2}\mu_5 \\ 1 & \text{if } Y_5 \geq \frac{1}{2}\mu_5 \end{cases}$$

and

$$X_6 = \begin{cases} 0 & \text{if } Y_6 < \frac{1}{2}\mu_6 \\ 1 & \text{if } Y_6 \geq \frac{1}{2}\mu_6. \end{cases}$$

B. Comparison of discriminant methods

The discriminant analysis methods considered in the study have been the following:

- Fisher's Linear Discriminator (LD).
- Quadratic Discriminator (QD).
- Discriminator based on the estimate of the density function by means of kernel functions (PARZEN). We have chosen the *Epanechnikov* function and a window width that minimizes the mean integrated square error of the estimated density [15].
- 1-NN and 3-NN nearest-neighbors. The distances used have been the same as for the DB discriminator.
- DB discriminator based on the mixed distance (8), with the absolute value distance (AV) for the quantitative variables and the SMI similarity coefficient for the qualitative variables, both described in Section II.
- *Resilient Backpropagation* [39] (RBP), learning algorithm used on multilayer perceptrons [22] to solve supervised learning tasks. A 6-10-1 multilayer perceptron with $a = \tanh(net)$ has been used as a transference function of the neurons integrated in the hidden layer and the output layer. On the other hand, the parameters have been set according to the values recommended by Riedmiller and Braun [39]: $\Delta_0 = .1$, $\Delta_{\max} = 50$, $\Delta_{\min} = 1e^{-6}$, $\eta^+ = 1.2$, $\eta^- = .5$.

C. Error estimators

By means of the s testing sets we can estimate the conditional errors

$$\widehat{ce}_l^{(s)} = \frac{1}{n_l} \sum_{j=1}^{n_l} I\left[r\left(\mathbf{x}_{lj}^{(s)}; t\right) \neq l\right] \tag{22}$$

so that

$$\widehat{ce}^{(s)} = \sum_{l=1}^{g} \pi_l \widehat{ce}_l^{(s)}. \tag{23}$$

The estimate of the unconditional errors is immediate

$$\widehat{ue}_l = \frac{1}{k} \sum_{s=1}^{k} \widehat{ce}_l^{(s)}, \quad \widehat{ue} = \sum_{l=1}^{g} \pi_l \widehat{ue}_l. \tag{24}$$

On the other hand, for each training data t we will obtain the non-parametric estimators presented in Section III. In this sense, if we consider t and the estimator θ, then the bias is the following:

$$b_\theta = \mathrm{E}[\theta - ce] = \mathrm{E}[\theta] - ce. \tag{25}$$

Given the differences

$$\hat{\theta}^{(j)} - \widehat{ce}^{(j)}, \tag{26}$$

the bias estimate becomes evident

$$\hat{b}_\theta = \frac{1}{k} \sum_{j=1}^{k} (\hat{\theta}^{(j)} - \widehat{ce}^{(j)}), \tag{27}$$

where j refers to the correspondent t training sets and s testing sets.

On the other hand, θ's mean square error is

$$MSE_\theta = \mathrm{E}[(\theta - ce)^2]. \tag{28}$$

The following relation is known

$$MSE_\theta = \mathrm{E}\left[(\theta - \widehat{ce})^2\right] - V\left[\widehat{ce}\right]. \tag{29}$$

From $n_l \widehat{ce}_l \sim B(n_l, ce_l)$, it can be inferred that

$$V[\widehat{ce}_l] = ce_l(1 - ce_l)/n_l.$$

Thus

$$V[\widehat{ce}] = \sum_{l=1}^{g} \pi_l^2 \frac{ce_l(1 - ce_l)}{n_l}. \tag{30}$$

According to the previous results, the following differences can be obtained:

$$(\hat{\theta}^{(t)} - \widehat{ce}^{(s)}) - \sum_{l=1}^{g} \pi_l^2 \frac{\widehat{ce}_l^{(s)}(1 - \widehat{ce}_l^{(s)})}{n_l}, \tag{31}$$

therefore the mean square error estimate is

$$\widehat{MSE}_\theta = \frac{1}{k} \sum_{j=1}^{k} \left[(\hat{\theta}^{(j)} - \widehat{ce}^{(j)}) - \sum_{l=1}^{g} \pi_l^2 \frac{\widehat{ce}_l^{(j)}(1 - \widehat{ce}_l^{(j)})}{n_l} \right]. \tag{32}$$

D. Analysis

In order to identify interactive behaviors between the random vector distribution and the discriminant rules error, the conditional error estimates obtained by means of the test data have been analyzed with a $(2 \times 2 \times 2) \times 10$ mixed factorial model, where the between subjects factor are the vector features (means M, structure of correlations SC and variance V) and the discriminant methods (DM) is the within subjects factor.

On the other hand, aiming to identify interactive behaviors between the random vector distribution and the bias and mean square error of the non-parametric conditional error estimators, two $(2 \times 2 \times 2) \times (10 \times 5)$ mixed factorial models have been used. The factors are the same aforementioned, with the addition of the estimators (ESTIM) as a new within subjects factor. The dependent variables in every model have been the expressions (26) and (31).

In order to deal with multiple comparisons, the minimum significant difference (MSD) has been fixed taking into account an overall significance level $\alpha_F = 1-(1-\alpha)^k < .05$ for every family of comparisons, where k is the number of contrasts. We have only carried out contrasts that refer to the large-size and medium-size effects (according to Cohen's judgment [8]) on the $(2\times 2\times 2)\times 10$ model ($k = 450$), and large-size effects on the $(2\times 2\times 2)\times(10\times 5)$ models ($k = 620$ and $k = 220$, respectively).

E. Software

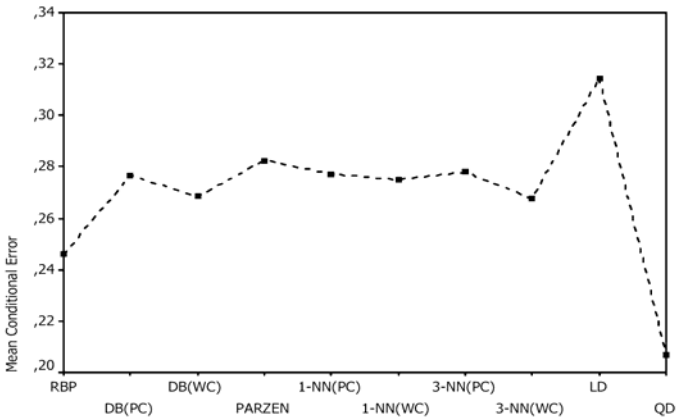
The discriminant methods, the conditional error estimators, as well as the whole simulation algorithms, have been implemented on MATLAB. Statistical analyses have been carried out on SPSS.

V. Results

A. Identifying interactive behaviors between the random vector distribution and the discriminant rules error

Multivariate contrasts show interactive patterns among all the $\mathrm{DM}\times\mathrm{SC}\times\mathrm{V}\times\mathrm{M}$ factors (Lambda Wilks = .063; $F_{9/1581} = 11.824$; $p < 1e^{-7}$; $\omega^2 = 0.006$; observed power = 1). The effects showing a larger size are those referring to the discriminant methods (Lambda Wilks = .879; $F_{9/1581} = 1280.754$; $p < 1e^{-7}$; $\omega^2 = 0.419$; observed power = 1); to the $\mathrm{DM}\times\mathrm{V}$ interaction (Lambda Wilks = .557; $F_{9/1581} = 221.436$; $p < 1e^{-7}$; $\omega^2 = 0.110$; observed power = 1); and to the $\mathrm{DM}\times\mathrm{SC}$ interaction (Lambda Wilks = .426; $F_{9/1581} = 130.610$; $p < 1e^{-7}$; $\omega^2 = 0.068$; observed power = 1).

Figure 1 shows mean conditional error (unconditional error estimate) for every discriminant rule. The methods with smaller error are QD (mean = 0.207, standard error = $9.576\cdot 10^{-4}$) and RBP ($m = 0.246$; s.e. = $1.086\cdot 10^{-3}$); whereas the worst method is LD ($m = 0.134$; s.e. = $9.105\cdot 10^{-4}$). The smaller error is evident on methods with non-joint covariance matrices estimate (WC).

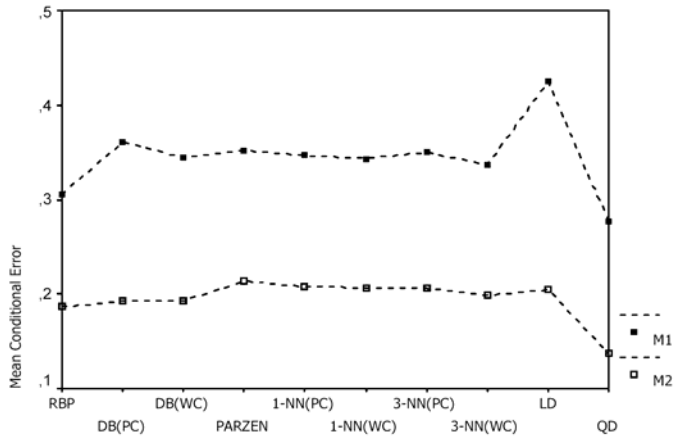


DM

Figure 1. Mean conditional error under DM effect conditions.

DM × M interaction (Figure 2) highlights the differences between methods when class means are equal (M1), specifically the good results of QD.

The DM × SC interaction analysis (Figure 3) points out that LD and DB have peculiar nuances. Whereas the LD has a mean conditional error similar to other methods under equal correlation structures (SC1), errors are clearly bigger on the SC2 situations, as it was expected. On the other hand, DB methods show a smaller error than the k-NN and PARZEN methods in SC1, while they are alike in SC2.



DM-M

Figure 2. Mean conditional error under DM-M effect conditions.

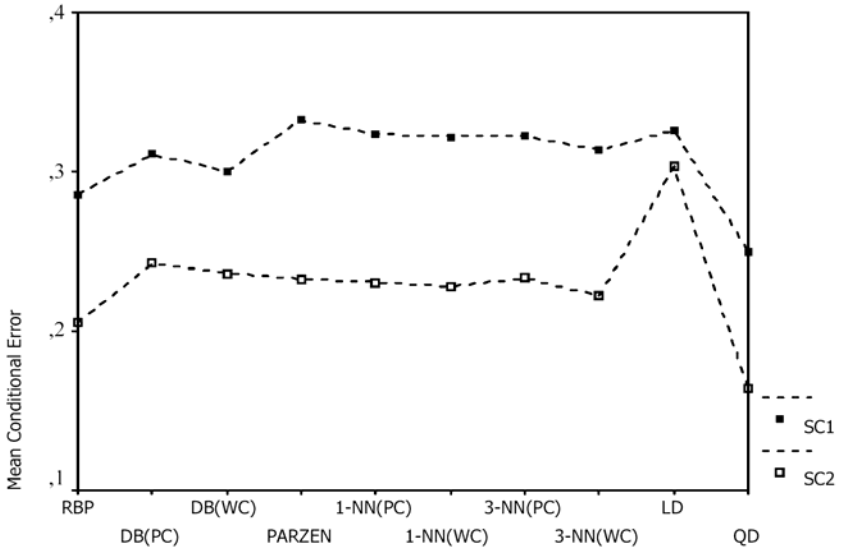


Figure 3. Mean conditional error under DM-SC effect conditions.

B. Identifying interactive behaviors between random vector distribution and conditional non-parametric error estimators

(1) Bias

Multivariate contrasts show a significant interaction effect among all the factors introduced in the model DM × ESTIM × SC × V × M (Lambda Wilks = .520; $F_{27/1566} = 53.574$; $p < 1e^{-5}$; $\omega^2 = 0.022$; observed power = 1). The effects showing the largest size are those referring to the estimators (Lambda Wilks = .005; $F_{3/1590} = 99827.405$; $p < 1e^{-5}$; $\omega^2 = 0.824$; observed power = 1); to the first order interaction between discriminant methods and estimators DM × ESTIM (Lambda Wilks = .007; $F_{27/1566} = 8055.782$; $p < 1e^{-5}$; $\omega^2 = 0.773$; observed power = 1); and to the second order interaction between discriminant methods, estimators and centroids DM × ESTIM × M (Lambda Wilks = .082; $F_{27/1566} = 652.390$; $p < 1e^{-5}$; $\omega^2 = 0.216$; observed power = 1).

Figure 4 shows the bias for the different non-parametric error estimators. AE and B.632 have negative bias, although very different in magnitude (bias = –0.127, s.e. = $6.475 \cdot 10^{-4}$ and bias = $-1.887 \cdot 10^{-2}$, s.e. = $7.318 \cdot 10^{-4}$, respectively); whereas B.632+, LOO and LOO-B show positive bias (bias = $1.390 \cdot 10^{-2}$, s.e. = $8.490 \cdot 10^{-4}$; bias = $2.307 \cdot 10^{-2}$, s.e. = $8.913 \cdot 10^{-4}$; and bias = $4.450 \cdot 10^{-2}$, s.e. = $8.210 \cdot 10^{-4}$). All a posteriori contrasts show significant differences.

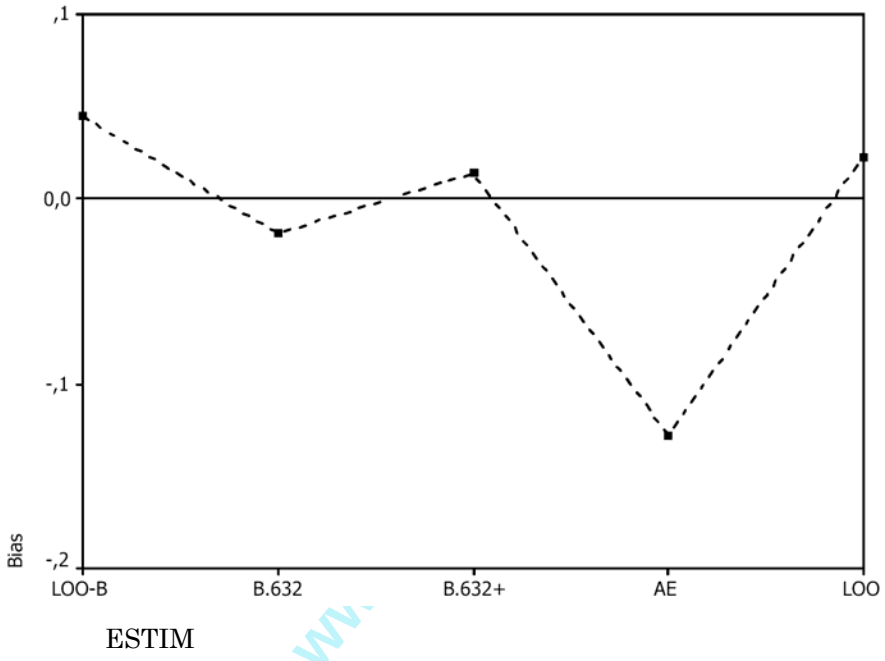


Figure 4. Bias under ESTIM effect conditions.

DM × ESTIM interaction (Figure 5) reveals that the bias depends on the discriminant method. The results obtained from the contrasts of each estimator's bias depending on the discriminant methods are relevant. Out of the 90 possible contrasts, AE shows the biggest number of significant contrasts (86), while the LOO estimator shows the smallest one (34). LOO-B, B.632+ and B.632 estimators present 61, 78 and 84 significant contrasts, respectively.

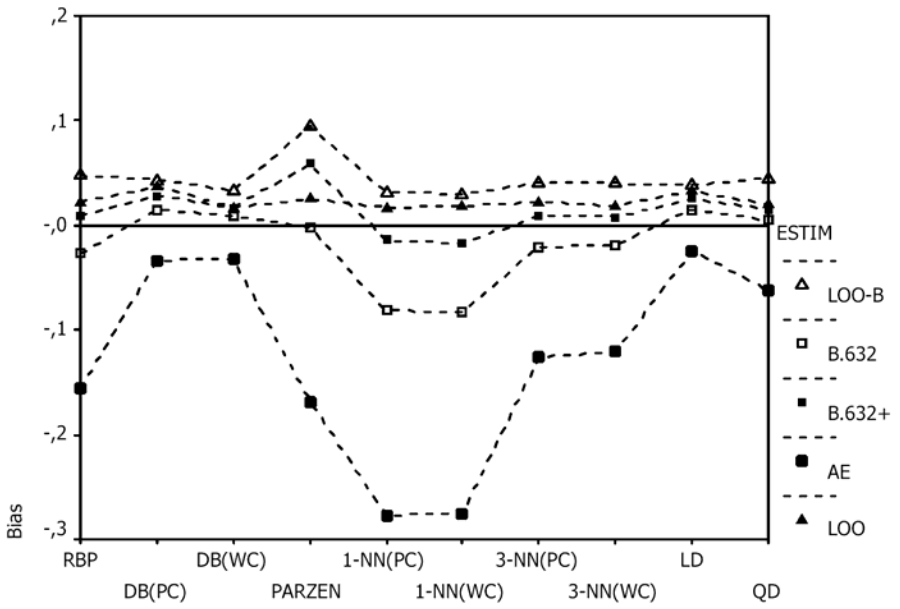


Figure 5. Bias under DM-ESTIM effect conditions.

DM × ESTIM × M interaction points out that class separation can influence the bias depending on the discriminant rule being used (Figures 6 and 7). The significant contrasts found to show a tendency to increase the bias when the centroids are equal (M1) in the same direction as the bias described when these are unequal (M2), specially with AE and B.632. However, B.632+ does not show this behavior with the 1-NN(PC) and 1-NN(WC) (in fact, its behavior is just in the opposite way). Another exception is the sign inversion on the bias introduced by B.632 on the PARZEN discriminant method, it being negative with equal centroids (M1), and positive in M2 situations. On the other hand, LOO and LOO-B only show significant contrasts on 4 out of the 10 discriminant methods: DB(PC), DB(WC), 3-NN(WC) and LD in the first one; RBP, DB(PC), DB(WC) and LD in the second one. B.632, B.632+ and AE estimators show a more heterogenic behavior, with significant contrasts being detected in 7, 8 and 10 out of the 10 discriminant methods, respectively.

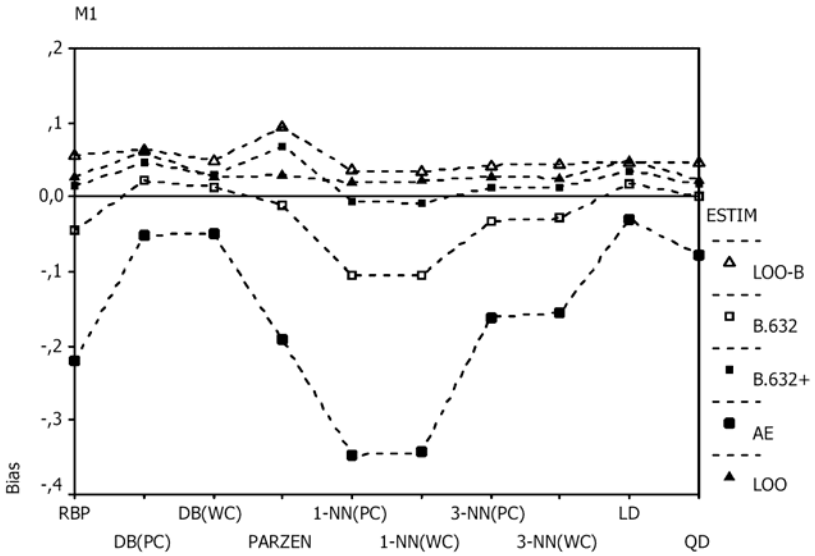


Figure 6. Bias under DM-ESTIM-M effect conditions (M1).

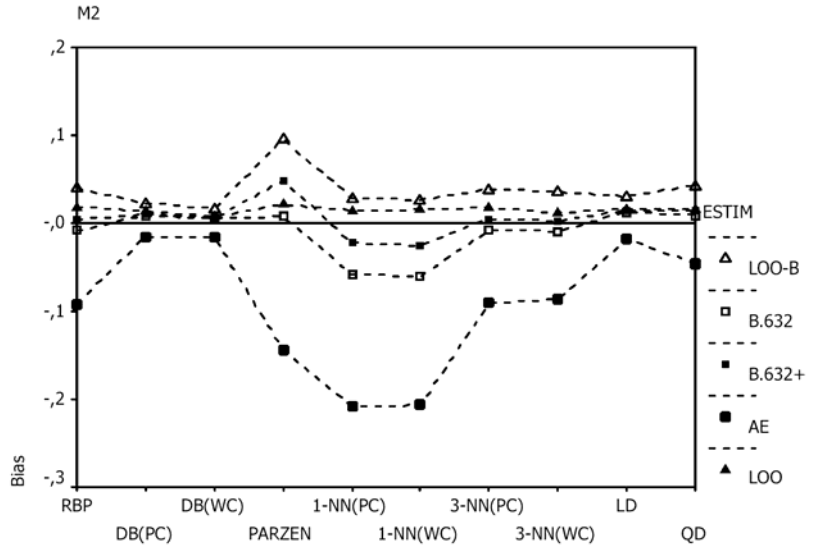


Figure 7. Bias under DM-ESTIM-M effect conditions (M2).

(2) Mean Square Error

Like on the bias analysis, a significant interactive effect is detected among all the DM × ESTIM × SC × V × M factors (Lambda Wilks = .761;

$F_{36/1557} = 13.602$; $p < 1e^{-5}$; $\omega^2 = 0.006$; observed power = 1). Nevertheless, the only large-size effects are those referring to estimators (Lambda Wilks = .024; $F_{4/1589} = 15955.869$; $p < 1e^{-5}$; $\omega^2 = 0.444$; observed power = 1) and to the DM × ESTIM first order interaction (Lambda Wilks = .025; $F_{36/1557} = 1674.213$; $p < 1e^{-5}$; $\omega^2 = 0.429$; observed power = 1).

On Figure 8 we can see how B.632+ is the one with the smallest mean square error (Mean Square Error = $3.408 \cdot 10^{-3}$, s.e. = $6.779 \cdot 10^{-5}$), followed by LOO and B.632 (MSE = $4.167 \cdot 10^{-3}$, s.e. = $8.101 \cdot 10^{-5}$; MSE = $4.018 \cdot 10^{-3}$, s.e. = $6.605 \cdot 10^{-5}$). LOO-B shows a mean square error higher than the previous ones (MSE = $4.911 \cdot 10^{-3}$, s.e. = $9.020 \cdot 10^{-5}$), still far from AE (MSE = $2.898 \cdot 10^{-2}$, s.e. = $2.094 \cdot 10^{-4}$).

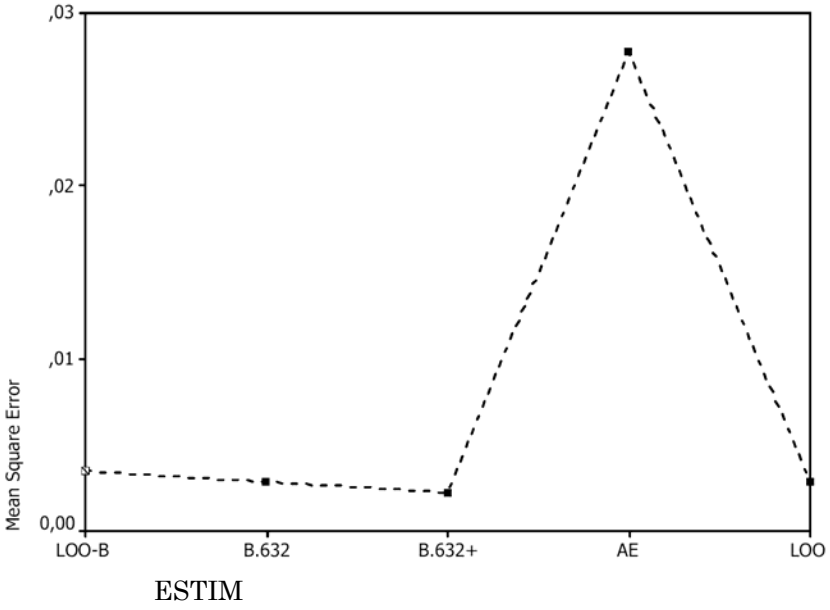


Figure 8. Mean square error under ESTIM effect conditions.

The DM × ESTIM effect (Figure 9) reveals important differences in the mean square error of every estimator depending on the discriminant method. The multiple comparisons show similar values to those obtained with bias. Out of the 90 possible contrasts, AE shows the biggest number

of significant contrasts (82); and LOO, the smallest one (42). LOO-B, B.632+ and B.632 present 50, 52 and 56 significant contrasts, respectively.

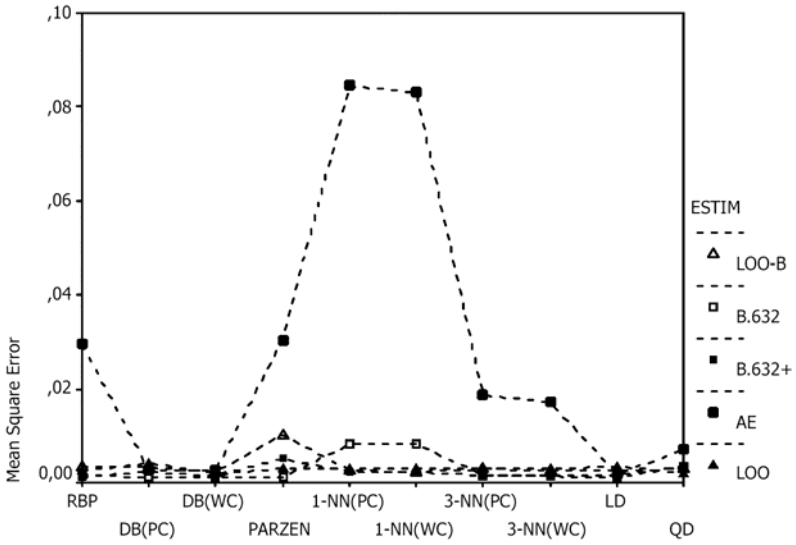


Figure 9. Mean square error under DM-ESTIM effect conditions.

VI. Discussion and Conclusions

A. Random vector distribution and discriminant rules error

QD and RBP show the smallest unconditional error, while LD has the biggest one. While a better behavior can be described of those methods which do not carry out the joint covariance matrix estimate, this fact may vary depending on the different centroid and correlation structure configurations. In this sense, the following conclusions are to be remarked:

1. The differences between methods using a pooled covariance matrix and those using a within covariance matrix decrease with class separation (differentiated means).

2. In comparison to the rest of the methods, LD shows a notably poorer performance when the correlation structures do not match. On the other hand, when compared to NN, PARZEN and LD methods, DB methods show better behavior with equal correlation structures.

It is easily noticed that results depend greatly on the specific situation. Out of the eight situations analyzed, six show classes with differentiated covariance matrices, and this fact explains the generally good performance of those methods that do not use the joint covariance matrix. However, the good performance of QD when dealing with mixed data is a bit surprising. It is possible that generating the data by discretizing multivariate normal random vectors improves its performance, whereas its behavior might worsen with data generated from non-normal random vectors.

B. Random vector distribution and behavior of non-parametric conditional error estimators

The bias and mean square error on the estimators vary ostensibly depending on the discriminant method and the random vector distribution. The following conclusions can be extracted:

1. The use of the AE estimator with assessment and comparison objectives must be rejected, since it presents large negative bias and unacceptable mean square error, besides showing heterogenic behavior depending on the situation and the discriminant rule.

2. The use of B.632 and B.632+ in order to carry out assessment tasks is recommendable, since they present the smallest mean square error in most situations. However, they show a certain heterogenic behavior depending on the discriminant method and random vector distribution, which advises against its use in comparison tasks. B.632+ does not improve substantially B.632, except for overfitting methods such as 1-NN, as it was expected.

3. LOO shows the most homogenous mean square error among the different discriminant methods and random vector distributions. This fact, along with its moderate bias and mean square error, recommends its use for discriminant rules comparison purposes.

Despite the small number of situations evaluated, we consider that this study reflects the need of analyzing the behavior of discriminant methods and non-parametric error estimators jointly with the statistical features of the patterns, in a simultaneous and integrated way. Otherwise, the results may appear as contradictory or even lead us to erroneous conclusions.

References

[1] T. Ameniya and J. L. Powell, A comparison of the logit model and normal discriminant analysis when the independent variables are binary, Studies in Econometrics, Time Series and Multivariate Statistics, S. Karlin, T. Ameniya and L. A. Goodman, eds., Academic Press, New York, 1983, pp. 3-30.

[2] T. Ashikaga and P. C. Chang, Robustness of the linear discriminant function under two-component mixed normal models, J. Amer. Statist. Assoc. 76 (1981), 676-680.

[3] N. Balakrishan and S. Kocherlakota, Robustness to nonnormality of the linear discriminant function: mixture of normal distributions, Comm. Statist. Theory Methods 14 (1985), 465-478.

[4] C. K Bayne, J. J. Bauchamp, V. E. Kane and G. P. McCabe, Assessment of Fisher and logistic linear and quadratic discrimination models, Comput. Statist. Data Anal. 1 (1983), 257-273.

[5] L. Breiman, J. Friedman, R. Olshen and C. Stone, Classification and regression trees, Wadsworth, Pacific Grove, CA, 1984.

[6] E. F. Chingada and F. Subrahmaniam, Robustness of the linear discriminant function to nonnormality: Johnson's system, J. Planning Inference 3 (1979), 69-77.

[7] W. R. Clarke and K. Subrahmaniam, Nonnormality affects the quadratic discriminant function, Comm. Statist. Theory Methods A8 (1979), 1285-1301.

[8] J. Cohen, Statistical Power Analysis for the Behavioral Sciences, Second Edition, Erlbaum, Hillsdale, New Jersey, 1988.

[9] D. R. Crawley, Logistic discrimination as an alternative to Fisher's linear discriminant function, N. Z. Statist. 14 (1979), 21-25.

[10] C. Cuadras, Distance analysis in discrimination and classification using both continuous and categorical variables, Statistical Data Analysis and Inference, Y. Dodge, ed., North-Holland, Amsterdam, 1989, pp. 459-473.

[11] C. M. Cuadras, J. Fortiana and F. Oliva, The proximity of an individual to a population with applications in discriminant analysis, J. Classification 14 (1997), 117-136.

[12] R. O. Duda, P. E. Hart and D. G. Stork, Pattern Classification, 2nd ed., Wiley, New York, 2001.

[13] B. Efron, Estimating the error rate of a prediction rule: Improvement on cross-validation, J. Amer. Statist. Assoc. 78 (1983), 316-331.

[14] B. Efron and R. Tibshirani, Improvements on cross-validation: The .632+ Bootstrap method, J. Amer. Statist. Assoc. 92 (1997), 548-560.

[15] V. Epanechnikov, Nonparametric estimates of a multivariate probability density, Theory Probab. Appl. 14 (1969), 153-158.

[16] L. P. Fatti, D. M. Hawkings and E. L. Raath, Discriminant analysis, Toppics in Applied Multivariate Analysis, D. M. Hawkings, ed., Cambridge University Press, Cambridge, 1982, pp. 1285-1301.

[17] R. A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugen 7 (1936), 179-188.

[18] M. X. Froján and R. Rubio, Análisis discriminante de la adhesión al tratamiento en la diabetes mellitus insulinodependiente, Psicothema 16(4) (2004), 548-554.

[19] N. Glick, Additive estimators for probabilities of correct classification, Pattern Recognition 10 (1978), 211-222.

[20] J. C. Gower, Some distance properties of latent roots and vector methods used in multivariate analysis, Biometrika 53 (1966), 325-338.

[21] J. C. Gower and P. Legendre, Metric and Euclidean properties of dissimilarity coefficients, J. Classification 3 (1986), 5-48.

[22] M. T. Hagan, H. B. Demuth and M. H. Beale, Neural Network Design, PWS Publishing, Boston, MA, 1996.

[23] M. Halperin, W. C. Blackwelder and J. I. Verter, Estimation of the multivariate logistic risk function: a comparison of the discriminant and maximum likelihood approaches, J. Chron. Dis. 24 (1971), 125-128.

[24] R. J. Henery, Review of previous empirical comparisons, Machine Learning, Neural and Statistical Classificastion, D. Michie, D. J. Spiegelhalter and C. C. Taylor, eds., Ellis Horwood, London, 1994, pp. 125-130.

[25] M. D. Hidalgo and J. Gómez, Comparación de la eficacia de regresión logística politómica y análisis discriminante logístico en la detección del DIF no uniforme, Psicothema 12 (2000), 298-300.

[26] T. A. Hosmer, D. W. Hosmer and L. Fisher, A comparison of the maximum likelihood and discriminant function estimators of the coefficients of the logistic regression model for mixed continuous and discrete variables, Comm. Statist. Simulation Comput. 12 (1983a), 23-43.

[27] T. A. Hosmer, D. W. Hosmer and L. Fisher, A comparison of three methods of estimating the logistic regression coefficients, Comm. Statist. Simulation Comput. 12 (1983b), 577-593.

[28] H. Jo, I. Han and H. Lee, Bankrupcy prediction using case-based reasoning, neural networks, and discriminant analysis, Expert Systems Appl. 13 (1997), 97-108.

[29] E. A. Joachimsthaler and A. Stam, Four approaches to the classification problem in discriminant analysis, Decision Sci. 19 (1988), 322-333.

[30] M. Y. Kiang, A comparative assessment of classification methods, Decision Support Systems 35 (2003), 441-454.

[31] W. J. Krzanowski, Discrimination and classification using both binary and continuous variables, J. Amer. Statist. Assoc. 70 (1975), 782-790.

[32] W. J. Krzanowski, The performance of Fisher's linear discriminant function under non-optimal conditions, Technometrics 19 (1977), 191-200.

[33] P. A. Lachenbruch and L. L. Kupper, Discriminant analysis when one population is a mixture of normals, Biom. J. 15 (1973), 191-197.

[34] P. A. Lachenbruch and M. R. Mickey, Estimation of error rates in discriminant analysis, Technometrics 10 (1968), 1-11.

[35] S. Marks and O. J. Dunn, Discriminant functions when covariance matrices are unequal, J. Amer. Statist. Assoc. 69 (1974), 555-559.

[36] D. H. Moore, Evaluation of five discriminant procedures for binary variables, J. Amer. Statist. Assoc. 68 (1973), 399-404.

[37] T. F. O'Hara, D. W. Hosmer, S. Lemeshov and S. C. Hartz, A comparison of discriminant function and maximum likelihood estimates of logistic coefficients for categorical-scaled data, J. Statist. Comput. Simul. 14 (1982), 169-178.

[38] F. Oliva, Aportacions a l'anàlisi discriminant basada en distàncies, Estudi Comparatiu de Mètodes d'anàlisi Discriminant amb dades Mixtes, Unpublished Doctoral Thesis, Universitat de Barcelona, Spain, 1995.

[39] M. Riedmiller and H. A. Braun, A direct adaptative method for faster backpropagation learning: the RPROP algorithm, Proc. IEEE Int. Conf. Neural Networks, H. Ruspini, ed., San Francisco, 1993, pp. 586-591.

[40] P. M. Schmitz, J. D. F. Habbema and J. Hermans, A simulation study of the performance of five discriminant analysis methods for mixtures of continuous and binary variables, J. Statist. Comput. Simul. 23 (1985), 69-95.

[41] P. M. Schmitz, J. D. F. Habbema, J. Hermans and J. W. Raatgever, Comparative performance of four discriminant analysis methods for mixtures of continuous and discrete variables, Comm. Statist. Simulation Comput. 12 (1983), 727-751.

[42] C. A. B. Smith, Some examples on discrimination, Ann. Eugen. 13 (1947), 272-282.

[43] I. G. Vlachonikolis and F. H. C. Marriot, Discrimination with mixed binary and continuous data, Appl. Statist. 31 (1982), 23-31.

[44] P. W. Wahl and R. A. Kronmal, Discriminant functions when covariance matrices are unequal and sample sizes are moderate, Biometrics 33 (1977), 479-484.

[45] D. H. Wolpert, The lack of a priori distinctions between learning algorithms, Neural Computation 8 (1996), 1381-1390.

[46] D. H. Wolpert and W. G. Macready, No free lunch theorems for optimization, IEEE Transactions on Evolutionary Computation 1 (1997), 67-82.

■