A COMPARISON OF DISCRIMINANT ANALYSIS METHODS IN THE DIAGNOSTIC OF CORONARY HEART DISEASE

Ana M. Pires*, João A. Branco* and M. Antónia Amaral Turkman**

Abstract: There are a large number of examples of successful applications of discriminant analysis in medical problems, especially in the field of diagnosis.

In this paper a data set concerning 9 variables observed on 113 subjects (83 suffering from coronary heart disease and 30 not suffering from coronary heart disease) is analyzed. The aim of the medical study was to find a simple rule, auxiliary of the diagnosis in the out-patient clinic, able to adequately select the patients to be submitted to coronariography, an invasive test which is able to detect the disease with a high sensitivity but which is quite expensive and risky.

Several linear discriminant rules are compared: Fisher's linear discriminant, logistic discriminant and a modification of Fisher's linear discriminant, recently proposed and based on the projection pursuit method.

As the study involves a large number of variables, relatively to the number of observations, and since it is suspected that some of them are not very important in the detection of the disease, the problem of selecting the most important variables is also addressed. A new methodology based on backwards elimination and bootstrap tests is proposed. The results obtained were found interesting when compared to the usual procedures.

The three different rules are evaluated by the probabilities (total and conditional) of correct classification estimated by three procedures: apparent, cross-validation and bootstrap. It was possible to conclude that the second and third rules lead in this problem to satisfactory results, allowing for correct classification of about 86% of the subjects.

Keywords: Discriminant Analysis; Fisher's Linear Discriminant; Logistic discriminant; Projection Pursuit; Error rate; Bootstrap; Variable selection.

AMS Classification: 62H30

1. INTRODUCTION

Discriminant analysis, as a multivariate statistical procedure that aims to classify an individual in one of previously defined groups, based on the observation of a certain number of features or variables, has important applications in the field of medical diagnosis. Several examples may be found in HERMANS and HABBEMA [1975], ANDERSON [1974], AITCHISON and DUNSMORE [1975], TITTERINGTON *et al.* [1981] and KRUSINSKA *et al.* [1990], to name just a few.

The number of procedures to obtain discriminant rules has increasing largely in recent years, mainly due to the advances in computing facilities, making crucial its

* Departamento de Matemática, Instituto Superior Técnico e Centro de Análise e Processamento de Sinais

** Universidade de Lisboa e Centro de Estatística e Aplicações

adequate choice . Also of great importance is the evaluation of the chosen discriminant as well as the selection of the more important variables.

In this paper we make use of a data set, already studied and published from the medical point of view, MACIEIRA-COELHO *et al.* [1990], with the aim of comparing:

- several discriminant rules;
- several evaluation methods;
- several variable selection procedures.

We intend to obtain results and give some hints useful in future similar applications.

Particular attention is given to the use of bootstrap methods — which in the evaluation process is if not common, at least known — also at the stage of testing the importance of the variables.

In the next section a brief description of the data is given together with an explanation of all the methods under consideration. In section 3 the main results are presented and in section 4 the main conclusions are discussed.

2. DATA AND METHODS

2.1. DATA

The study and the data under concern are described in more detail in MACIEIRA-COELHO *et al.* [1990].

The main aim was to find a fast and inexpensive method for the adequate selection — based on 4 clinical variables (age, sex, risk factor and chest pain) and 5 variables obtained in the stress test (ST depression, appearance of arrhythmias, pain during the test, variations in blood pressure and R wave changes) — of the patients to be submitted to coronariography. This invasive test is able to detect the disease with high sensitivity but is very expensive and involves a certain amount of risk, therefore its rotine use is not advised.

113 subjects were selected and had the 4 clinical variables and the 5 stress test variables observed. All of them were after submitted to coronariography, from which it was concluded that 30 of them did not suffer from coronary heart disease (from now on referred as "not sick" or Group 1, G_1) and that 83 suffered from coronary heart disease ("sick" or Group 2, G_2).

The complete data matrix may be found in the above cited reference, MACIEIRA-COELHO *et al*. [1990], from where Table 1, with the description of the variables, is reproduced.

Table 1 - Description of the variables

Variable		Name	Levels	Codes/values
	x1	age		years
C				
L	x2	sex	female	0
I			male	1
N				
I	x3	risk	absent	0
C		factor	present	1
A				
L	x4	chest pain	apontaneous	0
			stress	1
S	x5	pressão	none	0
T		arterial	raise in diastolic	1
R			decrease in systolic	2
E			both	3
S				
S	x6	appearance	no	0
		of arrythmias	yes	1
T				
E	x7	ST	no	0
S		depression	yes	1
T				
	x8	R wavw changes		real
	x9	pain during	no	0
		the test	yes	1

2.2. COMPUTATION OF THE DISCRIMINANT RULE

We describe next the three estimation procedures under consideration in this study.

While the first and second methods are well known and had their efficacy often checked, the second, on the other hand, is a new method recently proposed (PIRES E BRANCO, 1994) which has performed very well on simulated data and in our opinion that fact *per si* justifies its inclusion in this work.

Much more methods could be considered (see e.g. MCLACHLAN, 1992 or HAND, 1981) but it has already been concluded, OLIVEIRA [1986], that Fisher's discriminant seems to yield much better results than quadratic discriminant and slightly better than logistic discriminant (it was nevertheless decided to try the logistic discriminant once more because it seems the most reasonable choice for this kind of data, mixture of

discrete and continuous variables without distribution assumptions, and because better methods of evaluation are now available).

The fact that only linear discriminant procedures were selected reflects the great practical importance of this type of procedures, mainly due to their easier interpretation.

2.2.1. FISHER'S DISCRIMINANT

The first approach to linear discriminant analysis is due to FISHER [1936] and is briefly the following: being $z = \alpha_0 + \underset{\sim}{\alpha}^T \mathbf{x}$ the desired linear discriminant function, search for the orthogonal projection of the observations, $\underset{\sim}{\alpha}^T \mathbf{x}$, which maximizes the ratio between the squared difference of the projected locations for the two groups by the within (common) groups variance, that is

$$\frac{\left[\underset{\sim}{\alpha}^T \left(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2\right)\right]^2}{\underset{\sim}{\alpha}^T \underline{\Sigma} \underset{\sim}{\alpha}}, \qquad (1)$$

where $\underset{\sim}{\mu}_1$ e $\underset{\sim}{\mu}_2$ denote the group means and $\underline{\Sigma}$ the common covariance matrix. As these parameters are usually unknown, they are replaced by their training sample counterparts, $\bar{\mathbf{x}}_1$, $\bar{\mathbf{x}}_2$ and $\mathbf{S}$. The solution, easy to obtain by standard algebra, is $\underset{\sim}{\hat{\alpha}} = \mathbf{S}^{-1}\left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right)$.

The classification rule is then:

- classify an individual, characterized by the vector of observations $\mathbf{x}$, in group 1 if $\underset{\sim}{\hat{\alpha}}^T \mathbf{x} + \hat{\alpha}_0 > 0$, (2)

where

$$\hat{\alpha}_0 = -\frac{1}{2} \underset{\sim}{\hat{\alpha}}^T \left(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2\right) + \log\left[\frac{\pi_2 C(1|2)}{\pi_1 C(2|1)}\right] \qquad (3)$$

and π_i is the *a priori* probability that an individual belongs to G_i while $C(i|j)$ is the cost associated to a misclassification in group i of an individual actually belonging to group j .

We remark that in his distribution free approach FISHER [1936] considered only the first part of the second member of (3). The second part is due to WALD [1944] who deduced the same linear discriminant function using a decision theoretic approach and the hypothesis of normality of the populations. Under that hypothesis, rule (2)

minimizes the total expected cost of misclassification or, in the case of equal costs, the total probability of misclassification.

Under mixture sampling — that is, without attending to the group of origin — like in this case, π_i can be estimated by $\hat{\pi}_i = n_i/n$, then $\hat{\pi}_1 = 30/113 \cong 0.27$ and $\hat{\pi}_2 \cong 0.73$.

As usual in problems concerning the life of human beings, costs are hardly quantified and in spite of the fact that $C(1|2)$ should probably be bigger than $C(2|1)$, we have adopted equal costs.[1]

2.2.2. PROJECTION-PURSUIT DISCRIMINANT

Several robustness considerations have led to the proposal of the use of a projection-pursuit method in discriminant analysis (see BRANCO AND PIRES, 1994, about this method in general).

It can be said that this method leads to a modification of Fisher's method in the sense that it still tries to find the direction which maximizes expression (1). The important difference is that in spite of plugging in the multivariate usual estimators, robust univariate estimators are used to estimate the location and dispersion of the projected data. Among many possibilities the simultaneous M-estimators with Huber type weights were chosen because of their good properties of efficiency, regularity and robustness (known as "Huber-proposal 2", HUBER, 1964; see also HAMPEL *et al.*, 1986). The function to maximize is then[2]

$$I\left(\underset{\sim}{\alpha}\right) = \frac{\left[T\left(\underset{\sim}{\alpha}^T \mathbf{x}_1\right) - T\left(\underset{\sim}{\alpha}^T \mathbf{x}_2\right)\right]^2}{S^2\left(\underset{\sim}{\alpha}^T \mathbf{x}_1\right) + S^2\left(\underset{\sim}{\alpha}^T \mathbf{x}_2\right)}, \qquad (4)$$

where T and S denote the M-estimators respectively of location and dispersions, and $\underset{\sim}{\alpha}^T \mathbf{x}_i$ represents the univariate sample corresponding to the projection of the observations of group i on direction $\underset{\sim}{\alpha}$.

After a convenient transformation to spherical coordinates a solution can be obtained by numerical methods (we use a quasi-Newton conjugate gradient algorithm).

---

[1] Note that the consideration of different costs is equivalent, for two groups and in mathematical terms, to a readjustment of the *a priori* probabilities.

[2] In Pires and Branco [1994] we have considered a slightly different version of this expression but which leads to similar results.

2.2.3. LOGISTIC DISCRIMINANT

The previous methods can be applied without any distributional assumptions. They are however, especially the first, linked to the normal multivariate distribution since they lead to the optimal solution under that assumption.

In the logistic model it is assumed that the *a posteriori* probabilities are of the form

$$q_1(\mathbf{x}) = P(G_1|\mathbf{x}) = \frac{\exp\left(\alpha_0 + \underset{\sim}{\alpha}^T \mathbf{x}\right)}{\exp\left(\alpha_0 + \underset{\sim}{\alpha}^T \mathbf{x}\right)+1} \qquad (5)$$

$$q_2(\mathbf{x}) = 1 - q_1(\mathbf{x})$$

and that an individual is classified in group 1 if $q_1(\mathbf{x}) > 1/2$, which is equivalent to $\alpha_0 + \underset{\sim}{\alpha}^T \mathbf{x} > 0$. The parameters can be estimated by maximum likelihood using an iterative procedure.

The family of distributions satisfying (5) is quite general, including in particular certain cases where some of the variables are discrete and some are normal, for more details see for instance SEBER [1984: chap. 6].

2.3. EVALUATION OF A DISCRIMINANT RULE

Since its final aim is to classify, the *performance* of a discriminant rule must be evaluated by its ability to classify well, or equivalently, to misclassify.

In theory one can define several types of error rates such as the optimal error rate or the unconditional error rate. However the most interesting in applications is the conditional error rate (also called actual), where conditional is referred to the training sample on which the rule was built:

$$e_{act} = \pi_1 e_{1,act} + \pi_2 e_{2,act}.$$

π_i has the previously defined meaning and $e_{i,act}$ denotes the probability of (conditional to the estimated discriminant rule) misclassifying an individual from group i. As the distribution of $\mathbf{X}$ is generally not available it is necessary to estimate these probabilities.

From a lot of proposed methods those of a nonparametric nature are presently preferred as they do not rely on distributional assumptions. We have chosen three of them, described next. Only the case of the error rate, e, is addressed, obviously the estimate of the probability of correct classification is 1-e. Note that any of the methods is independent of the procedure used to obtain the discriminant rule.

2.3.1. APPARENT RATE

The most obvious rate is the so called apparent error rate obtained by reclassifying the training sample and counting the number of misclassified individuals in each group, m_i:

$$e_{ap} = \pi_1 e_{1,ap} + \pi_2 e_{2,ap} = \pi_1 \frac{m_1}{n_1} + \pi_2 \frac{m_2}{n_2},$$

and, in the case of $\hat{\pi}_i = n_i / (n_1 + n_2)$, it becomes $e_{ap} = (m_1 + m_2)/(n_1 + n_2)$.

Although based on intuition this estimate is biased towards optimistic values, since it uses the same sample on which the rule was built.

2.3.2. CROSS VALIDATION RATES

The ideal solution to estimate the error rate would be to have samples large enough so that part of the sample could be used to build the rule and the rest to evaluate it (this is usually called the test sample in opposition to the training sample) leading to independent estimates of the error rate. As the sample sizes do not usually permit this scheme the idea of cross validation has appeared: obtain a discriminant rule with all the observations, next, and only to estimate the rates, split the sample in two subsamples, one with n-k_1 individuals and other with k_1 individuals, use the first to obtain a discriminant rule and classify the remaining k_1 with that rule, counting the number of misclassified observations. Repeat this procedure a large number of times, or as many as possible, $\binom{n}{k_1}$, and finally compute the mean of the observed misclassifications. The more usual value for k_1 is 1 leading to the so called "leaving-one-out" method also named, incorrectly though, "jackknife".

EFRON [1983] has concluded that this method reduces substantially the bias when compared with the apparent rate, however its variability is rather high.

Therefore the same author has proposed the next method which, according to him, also reduces the bias but has a smaller dispersion.

2.3.3. BOOTSTRAP RATES

As any bootstrap method it begins by resampling from the initial sample. So, in this technique, a new sample of size n_i (i=1,2) is extracted with replacement from the n_i original observations. Every time this is done a new discriminant rule is obtained by the current procedure. Let m_i^* observations from the new sample and m_i^{**} from the original sample be misclassified by the new rule and let $d_i = \left(m_i^{**} - m_i^*\right)/n_i$. Repeat the scheme a large number of times (in this application we have used 500) and let $\bar{d}_i$ be the mean of the d_i. The bootstrap estimates of the probabilities of misclassification, for each group, are then,

$$e_{i,boot} = \frac{m_i}{n_i} + \bar{d}_i$$

and of the total probability,

$$e_{boot} = \pi_1 e_{1,boot} + \pi_2 e_{2,boot} \ ,$$

as usual.

2.4. VARIABLE SELECTION

It is common in applied problems like this one to try to include a large number of explanatory variables. On one hand, because there is the illusion that the greater the number of variables the better the fit and, on the other hand, because in many cases it is not known exactly which variables are really important.

However, the inclusion of a too large number of variables may cause troubles. These problems may be purely mathematical — collinearity, indetermination, convergence, etc. — or of a statistical nature, particularly overfitting.

In the special case of discriminant analysis and although, in theory, it can be demonstrated that the inclusion of new variables never harms the classification, in practice, that is, with finite samples, the contrary seldom occurs, the exclusion of one or more variables may in fact increase the estimates of the probabilities of correct classification.

Most of the published work on variable selection in linear discriminant analysis is based in the general results produced for the selection of variables in linear

regression. Namely the methods: "all subsets" and "stepwise selection", including the versions "forward selection" and "backward elimination". The "all subsets" method becomes prohibitive even for a reasonable number of variables, in our example it would be necessary to analyze $2^9 - 1 = 512$ possible subsets. The stepwise selection with a forward initial step as well as the forward selection have the deficiency, seldom referred, that they are not able to select highly correlated variables which are not important when isolated but can be important jointly. Under these circumstances our preference goes to the "backwards elimination" with a small modification consisting of: once selected individually two or more variables to eliminate that elimination is only effective after a joint test. It is thus not a "pure" "backwards" since it allows for the reentrance of variables.

In the next subsections some of the possible test statistics for each discriminant method are described. The criteria used for the selection of the final subset of variables are based on the p-values obtained by those statistics. In order to make it easier a more detailed explanation is postponed until section 3, in face of the real results presented there.

2.4.1. F TESTS

A precise F-test only exits for Fisher's discriminant under multinormality of the populations.

Consider a partition of the *d* variables in $\mathbf{x}^T = \left(\mathbf{x}^{(1)T}, \mathbf{x}^{(2)T}\right)$, with *k* and *d-k* elements, respectively. The basic hypothesis to test is that only $\mathbf{x}^{(1)}$ is important for discrimination, or in other words, that the conditional distribution of $\mathbf{x}^{(2)}$ given $\mathbf{x}^{(1)}$ is the same for both groups. It is possible to prove that this equivalent to test the equality of the Mahalanobis distances between the two groups using the smaller set $\mathbf{x}^{(1)}$ or all the variables, $H_0: \Delta_k^2 = \Delta_d^2$. An adequate statistic for this propose is

$$F = \frac{n-d-1}{d-k}\left(\frac{D_d^2 - D_k^2}{c + D_k^2}\right), \qquad (6)$$

where $c = n(n-2)/n_1 n_2$ and D_d^2 and D_k^2 are the corresponding sampling distances. Under H_0 $F \cap F_{d-k,n-d-1}$.

Note that the maximum of expression (1) is the Mahalanobis distance, Δ, for the population parameters, or D, for the sample parameters. This implies that it is possible in the projection-pursuit method to estimate D_d and D_k and therefore to compute an observed value of F, but the exact distribution of F under H_0 is no longer

known. It is thought that for normal populations that distribution may be reasonably approximated by the same F distribution, since, given the properties of the M-estimators, projection-pursuit and Fisher's methods are asymptotically coincident. In any case this statistic will be applied with caution and only with the objective of trying to reach some conclusions about its behavior in the present case of obviously non normal data.

A test that is not an F test but plays a similar role for logistic discrimination is the likelihood ratio test, to test the hypothesis that d-k of the α coefficients are null, based on the statistic

$$y = -2\log\left[\hat{L}_{(k)} / \hat{L}_{(d)}\right]$$

where $\hat{L}_{(k)}$ is the maximum of the likelihood function with k variables. When the null hypothesis is true y is approximately distributed as a χ^2_{d-k} (y is the "deviance" in General Linear Models terminology).

2.4.2. WALD TESTS

Another possibility to obtain test statistics is to use the asymptotic distribution of the coefficients of the linear discriminant function to test a null hypothesis of the type "a subset $\underset{\sim}{\alpha}$ is identically null", $\underset{\sim}{\alpha}^{(2)} = \mathbf{0}$.

For Fisher's discriminant, and again for multinormal populations, it is known that $\hat{\underset{\sim}{\alpha}} \overset{\cdot}{\frown} N_d\left(\underset{\sim}{\alpha}, \mathbf{V}\left(\underset{\sim}{\alpha}\right)\right)$ (KENDALL and STUART, 1973, Vol.3, pag. 343), where, by the δ–method,

$$\mathbf{V}\left(\underset{\sim}{\alpha}\right) = \frac{1}{n}\left[\underset{\sim}{\alpha}\,\underset{\sim}{\alpha}^T + \left(\frac{n^2}{n_1 n_2} + \Delta^2\right)\underline{\Sigma}^{-1}\right], \qquad (7)$$

and for standardized coefficients, $\underset{\sim N}{\alpha} = \underset{\sim}{\alpha} \Big/ \left\|\underset{\sim}{\alpha}\right\|$,

$$\mathbf{V}\left(\underset{\sim N}{\alpha}\right) = \frac{1}{n}\frac{\mathbf{V}\left(\underset{\sim}{\alpha}\right)}{\left\|\underset{\sim}{\alpha}\right\|}\left(\mathbf{I} - \underset{\sim N}{\alpha}\,\underset{\sim N}{\alpha}^T\right).$$

In order to test $H_0: \underset{\sim}{\alpha}^{(2)} = \mathbf{0}$ one can use

$$W = \left(\hat{\underset{\sim}{\alpha}}^{(2)} - \underset{\sim}{\alpha}^{(2)}\right)^T \hat{\mathbf{V}}^{-1}\left(\underset{\sim}{\alpha}^{(2)}\right)\left(\hat{\underset{\sim}{\alpha}}^{(2)} - \underset{\sim}{\alpha}^{(2)}\right) \overset{\cdot}{\frown} \chi^2_{d-k}. \qquad (8)$$

For the logistic discriminant and because of the use of the maximum likelihood method we have also $\hat{\underset{\sim}{\alpha}}$ approximately distributed as $N_d\left(\underset{\sim}{\alpha}, \mathbf{I}^{-1}\left(\underset{\sim}{\alpha}\right)\right)$ where $\mathbf{I}\left(\underset{\sim}{\alpha}\right)$ is

Fisher's information matrix. Again (8) is true, with $\hat{\mathbf{V}}^{-1}\left(\underset{\sim}{\alpha}^{(2)}\right)$ replaced by $\hat{\mathbf{I}}\left(\underset{\sim}{\alpha}^{(2)}\right)$. The W statistic is usually called Wald's statistic.

Unfortunately the asymptotic distribution of $\underset{\sim}{\hat{\alpha}}$ for the projection-pursuit method is not known. Moreover it is no longer possible to apply an argument of the kind of the one used in the previous subsection and adapt expression (7) because $\underline{\Sigma}$ is never estimated, only the variances of the projected populations.

2.4.3. BOOTSTRAP TESTS

Although there exist many variants trying to improve the potentialities of bootstrap (see LEPAGE and BILLIARD, 1992) only the simplest version, also known as "naïf bootstrap" is considered here.

The basic idea of bootstrap is resampling, as explained in subsection 2.3.3. After that, and also has said before, for each bootstrap sample a discriminant linear function is computed e, simultaneously with the misclassifications the estimates of the coefficients $\underset{\sim}{\alpha}$, $\underset{\sim}{\hat{\alpha}}$, are registered as a matrix of observations

$$\mathbf{A} = \begin{bmatrix} \hat{\alpha}_{10} & \hat{\alpha}_{11} & \dots & \hat{\alpha}_{1d} \\ \dots & \dots & \dots & \dots \\ \hat{\alpha}_{k0} & \hat{\alpha}_{k1} & \dots & \hat{\alpha}_{kd} \end{bmatrix},$$

where d is the number of original variables and k is the number of bootstrap replicates.

To test the hypothesis $H_0: \underset{\sim}{\alpha}^{(2)} = \mathbf{0}$ versus $H_1: \underset{\sim}{\alpha}^{(2)} \neq \mathbf{0}$ form the submatrix $\mathbf{A}_*$ with only the coefficients associated to the variables being tested and compute its mean vector, $\underset{\sim}{\hat{\mu}}_*$, and covariance matrix, $\underline{\hat{\Sigma}}_*$. For each vector $\underset{\sim i}{\hat{\alpha}}^{(2)}$, i=1,...,k, compute the squared distance

$$d_i^{*2} = \left(\underset{\sim i}{\hat{\alpha}}^{(2)} - \underset{\sim *}{\hat{\mu}}\right)^T \underline{\hat{\Sigma}}_*^{-1} \left(\underset{\sim i}{\hat{\alpha}}^{(2)} - \underset{\sim *}{\hat{\mu}}\right)$$

and for the point (0,0,...,0), $d_0^{*2} = \underset{\sim *}{\hat{\mu}}^T \underline{\hat{\Sigma}}_*^{-1} \underset{\sim *}{\hat{\mu}}$.

The p-value of the test of H_0 versus H_1 is then given by k^*/k, where $k^* = \#\left\{i: d_i^{*2} > d_0^2\right\}$. Note that if $\underset{\sim}{\hat{\alpha}}^{(2)}$ has only one component the variance is not necessary and $k^* = \#\left\{i: \left|\underset{\sim}{\hat{\alpha}}_i - \underset{\sim}{\hat{\mu}}_*\right| > \left|\underset{\sim}{\hat{\mu}}_*\right|\right\}$.

If by any reason it is possible to assume that the distribution of $\hat{\underset{\sim}{\alpha}}$ is approximately normal (which can be checked from the sample consubstanced in the matrix **A**) then alternatively the p-values can be obtained using the fact that $D_i^{*2} \overset{\cdot}{\sim} \chi^2_{d-k}$. We will call the first procedure "nonparametric" and the second "normal".

As a precaution, if the observed distribution of $\underset{\sim}{\alpha}$ seems to have "outliers" robust methods must be used to estimate $\hat{\underset{\sim}{\mu}}_*$ e $\hat{\underline{\Sigma}}_*$ and it will be no longer advisable to use the second procedure. Other departures from normality, namely asymmetry, may be overcome, when testing only one component, by letting $k^* = 2\times\#\{i:\hat{\alpha}_i > 0\}$ if $\hat{\alpha} < 0$ or $k^* = 2\times\#\{i:\hat{\alpha}_i < 0\}$ if $\hat{\alpha} > 0$. Other solutions to this problem, applicable with any number of components, may involve an adequate transformation.

3. RESULTS

Finally we describe the application of all the above procedures to our data set, beginning with the selection of the variables.

3.1. VARIABLE SELECTION

3.1.1. FISHER'S METHOD

First we must say that the final selection of the variables must be based not only on the p-values but also on the estimated classification rates when the variables are omitted.

The global results obtained for Fisher's method are presented in Table 2.

The discriminant function with all the variables is the following[3]

$$z = 1.673 - 0.015x_1 - 0.803x_2 - 0.296x_3 - 0.096x_4 - 0.142x_5 - 0.360x_6 - 0.289x_7 - 0.007x_8 - 0.154x_9. \quad (9)$$

In this case there are no p-values to compute, and only the estimates of the correct probabilities of classification are shown. Consider next the elimination of the variables one at a time.

---

[3] In all the discriminant functions given we have divided all of the coefficients by $\left\|\underset{\sim}{\alpha}\right\|$, in order to make the comparisons across methods easier.

If we adopt a significance level of 5% (in any test), we arrive to the conclusion that the candidates to elimination are variables x_8, x_4, x_9, x_5 and x_1 (in decreasing order of p from the F test) although in relation to x_5 and x_1 there are certain doubts. By a Bonferroni argument the remaining for variables can not be removed neither isolately ($p_i < 0.05$), neither simultaneously because $p_{total} \leq \sum p_i < 0.05$.

We considered next the joint elimination of x_4 and x_8 obtaining $p > 0.05$ (if this has not happened we would have to consider x_4 with x_8 and x_8 with x_9). Adding x_9 still verifies $p > 0.05$. In the next step we opted, since the several tests do not agree in the ordination of x_1 and x_5, for the consideration of two possibilities: x_4, x_5, x_8, x_9 or x_1 , x_4, x_8, x_9. In terms of p-value all the tests seem now to point to the second set. This precautions are justified by the fact that the estimators of each coefficient in $\underset{\sim}{\alpha}$ are not independent and as a consequence, the information given by single elimination may not be valid.

Table 2 - Selection of variables for Fisher's method

Variables to remove	p-values				Correct class. without the variables (in %)		
	F Test	Wald Test	Bootstrap Test Non par.	Bootstrap Test Normal	Apparent	Cross-val.	Bootstrap
None	-	-	-	-	84.1	83.2	81.8
x_1	0.0511	0.0446	0.046	0.0458	82.3	80.5	80.7
x_2	0.0000	0.0000	0.000	0.0000	-	-	-
x_3	0.0153	0.0090	0.016	0.0233	-	-	-
x_4	0.5108	0.4909	0.602	0.6131	85.8	82.3	83.4
x_5	0.0579	0.0405	0.012	0.0172	83.2	82.3	79.8
x_6	0.0111	0.0047	0.006	0.0029	-	-	-
x_7	0.0128	0.0058	0.014	0.0088	-	-	-
x_8	0.7626	0.7524	0.804	0.8210	85.0	83.2	83.3
x_9	0.2291	0.2084	0.274	0.2528	84.1	83.2	82.6
x_4, x_8	0.7395	0.7194	0.790	0.8170	85.8	81.4	**83.9**
x_4, x_8, x_9	0.3520	0.3064	0.418	0.4397	85.0	84.1	83.0
x_4, x_5 x_8, x_9	0.0977	0.0505	0.060	0.0480	85.0	82.3	83.6
x_1, x_4 x_8, x_9	0.1044	0.0928	0.104	0.0965	83.2	82.3	81.9

When comparing the p-values given by the different statistics it can be said that, in general, there is a good agreement between the results, in the sense that all of them give the same kind of information concerning rejection or acceptance. This is a bit surprising since two of the methods assume normality. However, if we try to be a little more precise some differences can be found. First of all the Wald test gives (with

one exception) p-values always smaller than the other tests. This may be justified by the known fact that the δ-method tends to underestimate variability. Comparing the p-values of the F-test with those of the bootstrap test it is easily seen that in almost all cases the p-values of the F-test are smaller to the ones obtained from the bootstrap test for higher p's (indicating acceptance) and greater for lower p's (indicating rejection). This seems to point to a smaller power of the F-test relatively to the bootstrap test. At last we can conclude that the two procedures to compute the p-values of the bootstrap test are very similar, which may give a good indication of approximate normality of these estimators, even for non-normal data. This must also be the explanation for the general agreement between all the tests.

We discuss now the classification rates. The discrepancies between the different rates will be considered only in section 4. For the moment we would like to point out the already mentioned fact that the elimination of variables can indeed improve the classification results. If by the reasons already stated we trust more the bootstrap rate than we can conclude, although the differences are probably not significant, that the best set of variables is obtained by eliminating x_4 and x_8. This will be the final decision for Fisher's method, leading to the following discriminant function

$$z = 1.611 - 0.015x_1 - 0.815x_2 - 0.286x_3 \\ - 0.146x_5 - 0.334x_6 - 0.292x_7 - 0.190x_9 .$$

We must remark that if at the beginning a more conservative significance level was adopted, for instance 25%, as it is recommended in some literature on linear regression, the selection procedure would have stopped after the elimination of three variables, which is closer to the decision based on the rates.

3.1.2. PROJECTION-PURSUIT METHOD

With all the variables the projection-pursuit method led to the following discriminant function

$$z = 1.571 - 0.011x_1 - 0.599x_2 - 0.377x_3 - 0.222x_4 \\ - 0.176x_5 - 0.483x_6 - 0.388x_7 - 0.011x_8 - 0.185x_9 ,$$

which is, at least qualitatively, comparable to (9) as all the coefficients have the same sign and the same order of magnitude, however the angle between the two directions is 17.4^0.

Table 3 presents the results of the several steps performed for the final selection of variables. The criteria followed were the same as in the previous section.

Table 3 - Selection of variables for the projection-pursuit method

Variables to remove	p-values F Test	p-values Bootstrap Test Non par.	p-values Bootstrap Test Normal	Correct class. without variables (in %) Apparent	Cross-val.	Bootstrap
None	-	-	-	88.5	82.3	84.6
x_1	0.0942	0.151	0.1798	85.0	80.5	81.3
x_2	0.0000	0.000	0.0000	-	-	-
x_3	0.0015	0.021	0.0201	-	-	-
x_4	0.1128	0.269	0.2987	88.5	83.2	85.1
x_5	0.0151	0.027	0.0215	-	-	-
x_6	0.0081	0.005	0.0035	-	-	-
x_7	0.0011	0.008	0.0035	-	-	-
x_8	0.5853	0.705	0.7936	85.8	83.2	82.3
x_9	0.0859	0.214	0.2329	85.0	84.1	81.4
x_4, x_8	0.2329	0.450	0.5465	88.5	83.2	**85.7**
x_4, x_8, x_9	0.0563	0.209	0.2575	85.0	83.2	82.7
x_1, x_4 x_8, x_9	0.0336	0.124	0.1008	82.3	76.1	80.5

Again it can be verified that the two methods for the bootstrap test give similar p-values, leading to the conclusion of approximate normality of the estimators.

A comparison with the F-test shows greater discreapancies than before but this is not surprising given the tated doubts about the validity of this statistic. It is possible that with a minor modification, for instance in the constants or in the number of degrees of freedom, one can obtain better results.

In terms of a final decision regarding the variables to include and again taking into account the classification rates the conclusion is the same, variables x_4 and x_8 shall be excluded. The discriminant function is then

$$z = 1.527 - 0.013x_1 - 0.647x_2 - 0.367x_3 - 0.192x_5 - 0.423x_6 - 0.396x_7 - 0.272x_9 . \quad (10)$$

3.1.3. LOGISTIC

For this method, with all the variables, the discriminant function is

$$z = 1.506 - 0.012x_1 - 0.635x_2 - 0.290x_3 - 0.121x_4 - 0.142x_5 - 0.578x_6 - 0.329x_7 - 0.001x_8 - 0.188x_9 ,$$

whose coefficients are, again qualitatively, similar to those given by the previous methods.

All the figures necessary for the selection of the variables are shown in Table 4. It can be seen that the p-values of the likelihood ratio test and of the Wald test are in close agreement when testing only one variable but that the first seems to be more powerful when testing simultaneous eliminations which is not a surprising result.

For the bootstrap method it must be mentioned in the first place that some implementation problems have come out, because for some of the bootstrap replicates there was no convergence, due to limitations of the iterative procedure that finds the maximum of the likelihood which does not allow empty marginal for discrete variables or complete separation[4]. It was then necessary to censore the procedure and omit those samples. It is not yet very clear how this affects the global results, however is is expected that it is mot very serious. In the second place a remarkable nonnormality of the $\hat{\underset{\sim}{\alpha}}$ estimates, with high left asymmetry, is also present. This implies, on one hand, that it does not make sense to use the "normal" method for obtaineing p-values and, on the other hand that the symmetric non-parametric method does not work very well. Finally we note that the p-values for the asymmetric bootstrap procedure are closely related to the ones of the other tests when testing individually.

Table 4 - Selection of variables for the logistic discriminant

Varibles to remove	p-values				Correct class. without varibles (in %)		
	Dev. test	Wald test	Bootstrap test				
			Asymm.	Symmet.	Apparent	Cross val.	Bootstrap
None	-	-	-	-	88.5	85.8	85.2
x_1	0.1042	0.1098	0.070	0.090	86.7	83.2	83.5
x_2	0.0001	0.0001	0.000	0.018	-	-	-
x_3	0.0164	0.0230	0.022	0.083	-	-	-
x_4	0.3961	0.3976	0.452	0.436	88.5	85.0	83.3
x_5	0.0902	0.1096	0.039	0.081	-	-	-
x_6	0.0028	0.0196	0.004	0.136	-	-	-
x_7	0.0059	0.0113	0.004	0.035	-	-	-
x_8	1.0000	0.9641	0.833	0.825	88.5	85.8	85.6
x_9	0.1229	0.1260	0.180	0.182	87.6	84.1	84.6
x_4, x_8	0.6873	0.6584	-	0.708	87.6	85.8	**85.7**
x_4, x_8, x_9	0.2001	0.2262	-	0.384	85.0	81.4	82.3

[4] The same happens when in cross-validation individual 15 is omitted. This subject is the only "non-sick" having a value of 1 on variable x_6. The logical solution is to assume that he would always be misclassified.

x_1, x_4 x_8, x_9	0.0855	0.1423	-	0.228	84.1	77.9	82.5

In terms of selection the results are consistent with the previous leading to elimination of variables x_4 and x_8 and to the final discriminant function:

$$z = 1.469 - 0.012x_1 - 0.661x_2 - 0.286x_3 - 0.154x_5 - 0.538x_6 - 0.341x_7 - 0.228x_9. \quad (11)$$

3.2. FINAL RESULTS

It is of some interest the comparison of the classification results for all the methods, before and after the variable selection. In Table 5 those figures are presented, for each group and total.

Table 5 - Comparison of the correct classification rates

Method	Estimates of the prob. of correct classification (in %)								
	Apparent			Cross validation			Bootstrap		
	G1	G2	Total	G1	G2	Total	G1	G2	Total
Fisher (All)	53.3	95.2	84.1	50.0	**95.2**	83.2	48.0	94.4	81.8
Fisher (without x_4, x_8)	56.7	**96.4**	85.8	50.0	92.8	81.4	53.9	**95.0**	83.9
P. pursuit (All)	**70.0**	95.2	**88.5**	56.6	91.6	82.3	60.9	93.3	84.6
P. pursuit (without x_4, x_8)	66.6	**96.4**	**88.5**	53.3	94.0	83.2	**61.5**	94.7	**85.7**
Logistic (All)	66.6	**96.4**	**88.5**	60.0	**95.2**	**85.8**	60.2	94.4	85.2
Logistic (without x_4, x_8)	63.3	**96.4**	87.6	**63.3**	94.0	**85.8**	61.0	94.8	**85.7**

The advantages of selecting the variables is now rather obvious, especially when considering the bootstrap estimates.

Regarding the different discriminant procedures it is possible to conclude that although the results are quite similar, none is chatastrophic, Fisher's is clearly the worst, for all the estimates, mainly for group 1, where it can even be bellow 50%. The projection pursuit and the logistic behave almost equally, except for the cross-

validation rates which have, however, the greatest variation. The apparent conttradiction with the results obtained by OLIVEIRA [1986], who concluded that the performance of the logistic was poorer than that of the Fisher discriminant, can be justified by the difference in the evaluation procedure. She separated the whole sample on a training sample (with 60 subjects) and on a test sample (with the remaining 53 subjects) in the way described in the first paragraph of 2.3.2. This method is not very reliable with the small number of observations, especialy for group 1.

Having to choose a final discriminant function we would hesitate between (10) and (11). In figure 1 the projection of the observations on the direction given by (10), the discriminant scores, are represented. There is a mild nonnormality that may justify the worst performance of Fisher's discriminant.

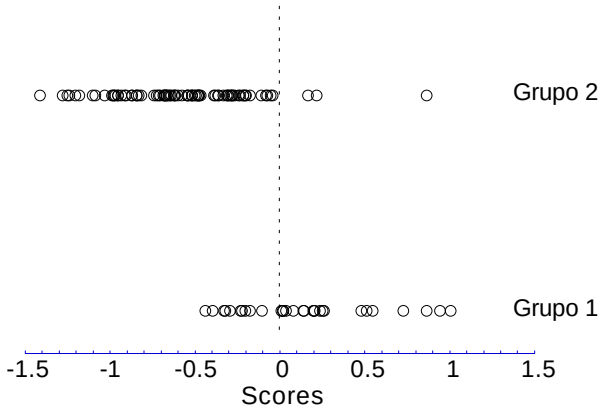

Figure 1 - Discriminant "Scores"

4. CONCLUSIONS

In face of the results presented in section 3 it is possible to conclude that linear discriminant analysis, based on a subset of the variables found important in medical terms, is an adequate method to solve the problem of diagnosing coronary heart disease. The sensitivity achieved is good, around 95 %, however the specificity is rather low, about 60%. We think that it may be difficult to obtain better results without further observations. This conclusion is in agreement with those of the preceeding studies on this data set (OLIVEIRA, 1986; MACIEIRA-COELHO *et al.*, 1990).

As for the main aim of comparison of the different methodologies three levels of conclusions were reached:

Choice of the discriminant method: if a linear discriminant is adequate, as it is thought in this particular problem, than the differences between methods shall not be huge, except in pathologic cases. However it seems safer to use a robust method, like the projection pursuit, or a better method for the specific type of variables, like the logistic, than the blind application of Fisher's discriminant.

Evaluation of the discriminant rule: it was clearly demonstrated the bias, in the optimistic direction, of the apparent rates. It wass also possible to verify the greater variability of the cross validation rates. In conclusion, and until other evidence, the bootstrap rates are preferable.

Selection of the variables: the importance of selecting the variables was shown together with the advantages of the backwards elimination. The bootstrap tests, whose greatest importance is their wide applicability, are apparently reliable and in some cases — Fisher and projection pursuit — really good. For the logistic some doubts remain and probably the best one is the likelihood ratio test.

REFERENCES

AITCHISON, J. & DUNSMORE, I. R. [1975]. *Statistical Prediction Analysis.* Cambridge University Press.

ANDERSON, J. A. [1974]. *Diagnosis by logistic discriminant function: further practical problems and results.* Appl. Statist., 23, 397-404.

BRANCO, J. A. & PIRES, A. M. [1994]. *Projection pursuit: a general strategy for the analysis of multivariate observations.* In Pestana, D. *et al*. (ed.) "Statistics and the Future and the Future of Statistics", Edições Salamandra, Lisboa, pp. 333-342 (in portuguese).

EFRON, B. [1983]. *Estimating the error-rate of a prediction rule: improvement on cross-validation.* J. Am. Statist. Assoc., 78, 316-331.

FISHER, R. A. [1936]. *The use of multiple measurement in taxonomic problems.* Ann. Eugen, 7, 179-188.

HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J. & STAHEL, W. A. [1986]. *Robust Statistics: the approach based on influence functions.* Wiley.

HAND, D. J. [1981]. *Discrimination and Classification.* Wiley.

HERMANS, J. & HABBEMA, J. D. F. [1975]. *Comparison of five methods to estimate posterior probabilities.* EDV in Medizin und Biologie, 6, 14-19.

HUBER, P. J. [1964]. *Robust estimation of a location parameter.* Ann. Math. Statist., 35, 73-101.

KENDALL, M. G. & STUART, A. [1973]. *The Advanced Theory of Statistics. (3rd edition).* Charles Griffin.

KRUSINSKA, E. & LIEBHART, J. [1990]. *Robust discriminant functions in assisting medical diagnosis: application to the chronic obturative lung disease data.* Biometrical Journal, 32, 915-929.

LEPAGE, R. & BILLARD, L. (ED.) [1992]. *Exploring the Limits of Bootstrap.* Wiley.

MACIEIRA-COELHO, E., OLIVEIRA, M. F. & AMARAL-TURKMAN, M. A. [1990]. *Diagnóstico de cardiopatia isquémica no doente ambulatório: análise multivariada de dados clínicos e electrocardiográficos.* Acta Médica Portuguesa, 3, 277-282.

MCLACHLAN, G. J. [1992]. *Discriminant Analysis and Statistical Pattern Recognition.* Wiley.

OLIVEIRA, M. F. [1986]. *Análise Discriminante quando os Dados são de Natureza Mista.* Tese de Mestrado, Universidade de Lisboa.

PIRES, A. M. & BRANCO, J. A. [1994]. *A robust linear discriminant by projection-pursuit.* COMPSTAT 1994, Book of Short Communications, pp. .

SEBER, G. A. F. [1984]. *Multivariate Observations.* Wiley.

TITTERINGTON, D. M., MURRAY, G. D., MURRAY, L. S., SPIEGELHALTER, D. J., SKENE, A. M., HABEMMA, J. D. F. & GELPKE, G. J. [1981]. *Comparison of discrimination techniques applied to a complex data set of head injuried patients (with discussion).* J. R. Statist. Soc. A, 144, 145-175.

WALD, A. [1944]. *On a statistical problem arising in the classification of an individual into one of two groups.* Ann. Math. Statist., 15, 145-162.