ZASTOSOWANIA MATEMATYKI
APPLICATIONES MATHEMATICAE
20, 4 (1990), pp. 551–560

E. KRUSIŃSKA and J. LIEBHART (Wrocław)

ON THE WILKS Λ STATISTIC AS A TOOL FOR MODEL SELECTION IN MIXED VARIABLE DISCRIMINANT ANALYSIS

Abstract. In the paper a new method for selection of the most discriminative variables in the location model for mixed variable discriminant analysis is presented. It is based on the Wilks Λ statistic modified to the location model case. It enables the simultaneous choice of discrete as well as continuous variables to the model. A medical example of application is given and the comparison with the selection procedure based on the classical Λ is added.

1. Introduction. The discriminant problem is of the frequent occurrence in many fields, e.g., in medicine, biology, agriculture, technical and economical researches. It consists in assigning a given individual to one of the populations considered on the basis of the observed values of predictor variables. The predictor variables may be of both continuous and discrete character. Such mixtures are very often present in practical problems, while there are only few methods of discriminant analysis especially elaborated to handle them. We should list here logistic approach, nonparametric density estimation methods and the so-called location model. This last one was introduced to the problem of discrimination with mixtures of continuous and binary variables by Krzanowski [9] in the two groups of data problem. The model assumes that the continuous variables follow the different multivariate normal distribution for each possible combination of binary variables. Under the additional assumption about homogeneity of covariance matrices this gives the classification by the linear discriminant function different for each cell of the contingency table defined by binary variables values. So the method is simple enough and computationally feasible.

The location model received in the last years a considerable attention. Krzanowski [10], [13] described generalizations of the method to the mixtures of continuous and discrete variables with more than two possible states and to the several groups of data problem. Some other extensions (e.g., introducing the quadratic classification rule and generalized multiple discrimination algorithm) were given by Krusińska [4], [8].

Moreover, in practice the problem of model choice arises. The methods of dimensionality reduction in the location model may be divided according to their possibilities into three groups. The first method (Krzanowski [12]) enables to select only discrete variables to the model. The second one (Krusińska [4], [7]) is based on the T^2 statistic — one of the possible test statistics in the multivariate analysis of variance — and enables the simultaneous choice of discrete and continuous variables to the model. The third group of methods gives a possibility to select not only single variables to the location model but also the terms in the linear additive model imposed to the mean vectors for their estimation. The method of Daudin [2] based on Akaike's criterion and procedures of Krusińska [5] based on various estimates of the total probability of misclassification are of the last type. A comparative study of some of the methods mentioned is given by Krusińska [6].

In the paper another procedure of the second group is presented in detail. It is based on the Wilks Λ criterion. The distributional approach is possible and the distribution free one as well.

2. Location model approach in detail. Suppose that each individual is described by a vector

$$y' = (y_1, y_2, \ldots, y_p)$$

of p continuous variables and a vector

$$x' = (x_1, x_2, \ldots, x_q)$$

of q binary ones. If we consider discrete variables with more than two states possible, the decoding into binary ones is made for them before discrimination. The problem is in classifying an individual $w = (x, y)$ to one of two populations Π_1 or Π_2 (generally we can consider more populations) on the basis of the observed values of x and y. It is assumed that

$$y \sim N(\mu_i^{(m)}, \Sigma) \quad (i = 1, 2;\ m = 1, 2, \ldots, 2^q).$$

The covariance matrix Σ is assumed to be equal for all 2^q "locations" (cells of the contingency table).

It is easy to prove that the optimal classification rule in that model is: allocate w falling into the m-th cell to Π_1 if

$$(1) \qquad (\mu_1^{(m)} - \mu_2^{(m)})' \Sigma^{-1} \{y - \tfrac{1}{2}(\mu_1^{(m)} + \mu_2^{(m)})\} \geqslant \log(p_{2m}/p_{1m})$$

and otherwise to Π_2.

p_{im} are the a priori probabilities of obtaining an individual from the i-th population in the m-th cell of the contingency table.

It is obvious that the classification rule (1) is equivalent to the classical linear Fisherian discrimination but performed separately for each cell of the contingency table.

In practice the parameters of the model, i.e., $\mu_i^{(m)}$, $p_{im} = (i = 1, 2;$ $m = 1, 2, \ldots, 2^q)$ and Σ, should be estimated from the data. The probabilities p_{im} are estimated by the iterative scaling procedure of Haberman [3] which allows for empty cells in the contingency table. The parameters $\mu_i^{(m)}$ and

Σ related to continuous variables are estimated by imposing the linear additive model on the mean vectors:

(2)
$$\mu_i = \nu_i + \sum_{j=1}^{q} \alpha_{j,i} x_j + \sum\sum_{j<k} \beta_{jk,i} x_j x_k + \sum\sum\sum_{j<k<l} \gamma_{jkl,i} x_j x_k x_l + \ldots + \delta_{12\ldots q,i} \cdot x_1 \cdot x_2 \cdot \ldots \cdot x_q.$$

The components of the model (2) may be interpreted as the main effects of each binary variable and interactions between binary variables of all orders up to q. In practice the first order model and the second order model are used.

Classification of an individual $w = (x, y)$ is performed by the leaving-one-out method, i.e., that the parameters p_{im}, $\mu_i^{(m)}$ ($i = 1, 2$; $m = 1, 2, \ldots, 2^q$) and Σ are calculated after throwing away the actually classified individual and then the classification is performed on their basis.

3. Selection of variables via hypotheses testing in linear models. Let us consider a multivariate linear model of the form

$$\underset{n\times p}{Y'} = \underset{n\times k}{X}\ \underset{k\times p}{B} + \underset{n\times p}{E}, \quad n > k,$$

where Y is the observation matrix, and X is the design matrix.

In this model we test the hypothesis on the parameters B:

(3)
$$H_0\colon \underset{l\times k}{K}\ \underset{k\times p}{B} = \underset{l\times p}{0}, \quad 1 \leqslant l \leqslant k,\ \mathrm{rk}(K) = l.$$

Let $M = KB$, $X = [X_1 X_2]$, $\mathrm{rk}(X_1) = r > 0$, and X_1 is nonsingular; $\hat{M} = X_1(X_1'X_1)^{-1}X_1'Y$ is the least squares estimate of M. Now let us define the matrix of residual sums of squares and products as

$$G = (Y - \hat{M})'(Y - \hat{M})$$

and the matrix of sums of squares and products due to H_0 as

$$H = \hat{\Delta}'[K_1(X_1'X_1)K_1']^{-1}\hat{\Delta},$$

where

$$\underset{l\times p}{\hat{\Delta}} = K_1(X_1'X_1)^{-1}X_1'Y.$$

The partition $K = [K_1 K_2]$ corresponds to the partition of the design matrix X.

The hypothesis H_0 (formula (3)) may be tested e.g., by the Lawley–Hotelling T^2 and the Wilks Λ statistic. The use of the Lawley–Hotelling T^2 to the location model choice was described by Krusińska [7]. Now we describe the application of the Wilks Λ to the same problem.

The Wilks Λ statistic (e.g., [14]) is defined as a ratio of two determinants:

(4)
$$\Lambda = \frac{|G|}{|G+H|}.$$

Ahrens and Läuter [1] summarize the results connected with the distribution of the Λ statistic. For some special cases the transformed Wilks Λ has the Snedecor F distribution:

$$
(5)\qquad
\begin{aligned}
p&=1 && \text{any } l\colon \frac{n-r}{l}\cdot\frac{1-\Lambda}{\Lambda}\sim F_{l,n-r},\\
p&=2 && \text{any } l\colon \frac{n-r-1}{l}\cdot\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}\sim F_{2l,2(n-r-1)},\\
l&=1 && \text{any } p\colon \frac{n-r-p+1}{p}\cdot\frac{1-\Lambda}{\Lambda}\sim F_{p,n-r-p+1},\\
l&=2 && \text{any } p\colon \frac{n-r-p+1}{p}\cdot\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}\sim F_{2p,2(n-r-p+1)},
\end{aligned}
$$

where p is the number of the variables considered, n is the number of individuals in the sample, l and r are defined as in (3).

In the other cases the approximation by F may be used:

$$
(6)\qquad \frac{1-\Lambda^{1/u}}{\Lambda^{1/u}}\cdot\frac{mu-2\gamma}{pl}\sim F_{pl,mu-2\gamma},
$$

where

$$
m=n-r-\frac{p-l+1}{2},\qquad u=\frac{p^2l^2-4}{p^2+l^2-5},\qquad \gamma=\frac{pl-2}{4}.
$$

The Wilks Λ statistic may be used in the location model case to the model selection. In the terminology of the analysis of variance we have in that problem a nested model of the two-way classification MANOVA with the first factor (A) corresponding to cells and the second one (B) corresponding to groups. The factor A has at most $s=2^q$ levels (some cells may be empty for all groups). The factor B has g_i levels for $i=1,2,\ldots,s$. The model considered can be written as

$$
(7)\qquad \boldsymbol{y}_{ijt}=\boldsymbol{\mu}+\boldsymbol{\alpha}_i+\boldsymbol{\beta}_{j(i)}+\boldsymbol{e}_{ijt}\qquad (i=1,2,\ldots,s;\ j=1,2,\ldots,g_i;\ t=1,2,\ldots,n_{ij}),
$$

where n_{ij} is the number of individuals in the i-th cell and the j-th group, $\boldsymbol{y}_{ijt}$, $\boldsymbol{\mu}$, $\boldsymbol{\alpha}_i$, $\boldsymbol{\beta}_{j(i)}$, $\boldsymbol{e}_{ijt}$ are vectors of p components (the number of continuous variables).

In the model (7) we test the hypothesis on the influence of the factor B corresponding to groups, i.e.,

$$
(8)\qquad H_B\colon
\begin{cases}
\boldsymbol{\beta}_{1(1)}=\boldsymbol{\beta}_{2(1)}=\ldots=\boldsymbol{\beta}_{g_1(1)},\\
\boldsymbol{\beta}_{1(2)}=\boldsymbol{\beta}_{2(2)}=\ldots=\boldsymbol{\beta}_{g_2(2)},\\
\cdots\cdots\cdots\cdots\cdots\\
\boldsymbol{\beta}_{1(s)}=\boldsymbol{\beta}_{2(s)}=\ldots=\boldsymbol{\beta}_{g_s(s)},
\end{cases}
$$

where $\boldsymbol{\beta}_{j(i)}$ corresponds to the influence of the j-th level of the factor B (observed for the i-th level of A) on the values of variables $\boldsymbol{y}$.

We can reformulate our problem in such a way because we want to find variables differentiating in the best way between the g groups considered (for each cell separately).

The matrices $\boldsymbol{H}$ and $\boldsymbol{G}$ can be found in an easy way by analogy to the test statistic in the univariate case. It is the F statistic given by

$$F_B = \frac{\sum_{i=1}^{s}\sum_{j=1}^{g_i} n_{ij}(\bar{y}_{ij.}-\bar{y}_{i..})^2}{g_.-s} : \frac{\sum_{i=1}^{s}\sum_{j=1}^{g_i}\sum_{t=1}^{n_{ij}} (y_{ijt}-\bar{y}_{ij.})^2}{n-g_.}, \tag{9}$$

where

$$n = \sum_{i=1}^{s}\sum_{j=1}^{g_i} n_{ij}, \quad g_. = \sum_{i=1}^{s} g_i,$$

y_{ijt} is the value of the variable y for the t-th observation, the j-th level of B, the i-th level of A; $\bar{y}_{ij.}$ is the mean value for the j-th level of B and the i-th level of A; $\bar{y}_{i..}$ is the mean value for the i-th level of A.

The matrices $\boldsymbol{H}$ and $\boldsymbol{G}$ are now obtainable by analogy to the sums of squares in the numerator and the denominator of the statistic (9). Thus

$$\boldsymbol{H} = \sum_{i=1}^{s}\sum_{j=1}^{g_i} n_{ij.}(\bar{\boldsymbol{y}}_{ij.}-\bar{\boldsymbol{y}}_{i..})(\bar{\boldsymbol{y}}_{ij.}-\bar{\boldsymbol{y}}_{i..})',$$

$$\boldsymbol{G} = \sum_{i=1}^{s}\sum_{j=1}^{g_i}\sum_{t=1}^{n_{ij}} (\boldsymbol{y}_{ijt}-\bar{\boldsymbol{y}}_{ij.})(\boldsymbol{y}_{ijt}-\bar{\boldsymbol{y}}_{ij.})',$$

where $\boldsymbol{y}_{ijt}$, $\bar{\boldsymbol{y}}_{ij.}$, $\bar{\boldsymbol{y}}_{i..}$ are now vectors of p components (p is the number of continuous variables).

To find the distribution of Λ we should also define the quantities r and l appearing in formulae (5) and (6). r is the rank of the design matrix $\boldsymbol{X}$ and l is the rank of the hypothesis matrix $\boldsymbol{K}$. After writing both matrices explicitly it is easily seen that r equals $g_.$ and l equals $g_. - s$ in the considered location model case (s is here the number of nonempty cells, because some of them may be empty for all groups). For the simpler motivation of the values of r and l we may see the numerator and the denominator of F_B statistic (formula (9)) in the univariate case.

Now let us return to the selection procedure. To the subset of the most discriminative variables we should choose those for which the hypothesis H_B (formula (8)) is rejected at the lowest significance level α. Using the distributional-free approach we select those variables for which the value of the Λ statistic is the smallest one. It should be stressed that from the theoretical point of view the distributional approach is more adequate to the problem because having selected different subsets of binary variables we consider different models, so the values of g and s differ for each subset. It is known (see, e.g., [14]) that, in the case where we have only continuous variables the Wilks Λ statistic is monotonous and has clear interpretation. It is equal to 0 when there is a full discrimination between groups considered and 1 when there is no discriminatory power at all. It decreases with the increase of the number of variables in the subset. The same cannot be told in the case of the mixture of binary and continuous variables when we consider different models for subsets of different binary variables. Thus the monotonicity of the Wilks Λ is not fulfilled in such a problem.

There is no theoretical obstacle to find the optimal subset of the most discriminative variables in the location model, using the criteria described (formulae (4)–(6)). But in such a choice it is necessary to investigate all subsets of the size considered, so in practice the stepwise procedures are commonly applied. Totally $p+q$ (continuous and binary) variables are used in the problem under consideration. At the first step of the backward selection algorithm the subsets of size $p+q-1$ (after eliminating the variables no. 1, 2, ..., $p+q$) are investigated. The probabilities $\Pr(F > F_{\text{calc}})$ (using formulae (5) and (6)) or the Wilks Λ (for the distributional-free approach – formula (4)) are computed for all these subsets. The best subset (where $\Pr(F > F_{\text{calc}})$ or Λ is the smallest one) is chosen and the procedure goes down for subsets of size $p+q-2$, $p+q-3$, and so on until 1. At each step of the algorithm one variable is eliminated. When only one continuous variable remains in the subset, it is saved in it and after that only binary variables are deleted because the selection procedure is based on the test statistic for the multivariate (or univariate if $p = 1$) analysis of variance, so at least one continuous variable should remain in the subset. The stepwise procedure described is computationally feasible. When instead of binary variables we omit discrete ones with more than two states possible, the procedure should be only slightly adapted.

4. Example of application. Now we present one medical application of the method described. The data used are the part of the more extended studies in the so-called chronic obturative lug disease. The sample consisted of 164 patients suffering from uncomplicated bronchial asthma ($n_1 = 112$) and bronchial asthma complicated by lung emphysema ($n_2 = 52$). 14 variables – 6 continuous ones called $C_1, C_2, C_3, C_4, C_5, C_6$ (5 spirometric examinations and a smoking index), and 8 binary ones called $B_1, B_2, B_3, B_4, B_5, B_6, B_7, B_8$ (disease symptoms such as cough, dyspnoea, X-ray examination of the chest) – were taken into consideration. The hypotheses on normality of continuous variables and on homogeneity of covariance matrices in both groups were rejected at the level $\alpha = 0.05$. The selection was performed in three variants: by the classical Wilks Λ neglecting that for binary variables the assumption on normality is not fulfilled, by the Wilks Λ criterion established for the location model case, by the distributional approach using formulae (5) and (6) for the location model case.

TABLE 1. Variable selection with the classical Wilks Λ criterion

Subset	Variable to be deleted	Value of the Wilks Λ	$\Pr(F > F_{\text{calc}})$
$C_1, C_2, C_3, C_4, C_5, C_6, B_1, B_2, B_3, B_4, B_5, B_6, B_7, B_8$	B_1	$6.512_{10}-1$	$8.669_{10}-9$
$C_1, C_2, C_3, C_4, C_5, C_6, B_2, B_3, B_4, B_5, B_6, B_7, B_8$	B_7	$6.515_{10}-1$	$5.568_{10}-9$
$C_1, C_2, C_3, C_4, C_5, C_6, B_2, B_3, B_4, B_5, B_6, B_8$	C_5	$6.519_{10}-1$	$1.343_{10}-9$
$C_1, C_2, C_3, C_4, C_6, B_2, B_3, B_4, B_5, B_6, B_8$	B_4	$6.524_{10}-1$	$9.762_{10}-10$
$C_1, C_2, C_3, C_4, C_6, B_2, B_3, B_5, B_6, B_8$	B_8	$6.534_{10}-1$	$1.870_{10}-10$
$C_1, C_2, C_3, C_4, C_6, B_2, B_3, B_5, B_6$	B_5	$6.551_{10}-1$	$1.825_{10}-10$
$C_1, C_2, C_3, C_4, C_6, B_2, B_3, B_6$	C_2	$6.582_{10}-1$	$3.020_{10}-11$
$C_1, C_3, C_4, C_6, B_2, B_3, B_6$	C_6	$6.616_{10}-1$	$4.560_{10}-11$
$C_1, C_3, C_4, B_2, B_3, B_6$	B_2	$6.650_{10}-1$	$4.685_{10}-12$
C_1, C_3, C_4, B_3, B_6	C_1	$6.738_{10}-1$	$1.927_{10}-11$
C_3, C_4, B_3, B_6	B_3	$6.879_{10}-1$	$3.115_{10}-12$
C_3, C_4, B_6	B_6	$7.003_{10}-1$	$3.282_{10}-11$
C_3, C_4	C_4	$7.172_{10}-1$	$2.385_{10}-12$
C_3		$7.403_{10}-1$	$2.539_{10}-10$

TABLE 2. Variable selection with the Wilks Λ for the location model

Subset	Variable to be deleted	Value of the Wilks Λ	$\Pr(F > F_{\text{calc}})$
$C_1, C_2, C_3, C_4, C_5, C_6, B_1, B_2, B_3, B_4, B_5, B_6, B_7, B_8$	B_5	$2.487_{10}-1$	$1.485_{10}-1$
$C_1, C_2, C_3, C_4, C_5, C_6, B_1, B_2, B_3, B_4, B_6, B_7, B_8$	B_7	$2.598_{10}-1$	$8.210_{10}-2$
$C_1, C_2, C_3, C_4, C_5, C_6, B_1, B_2, B_3, B_4, B_6, B_8$	B_8	$2.720_{10}-1$	$8.294_{10}-2$
$C_1, C_2, C_3, C_4, C_5, C_6, B_1, B_2, B_3, B_4, B_6$	C_2	$3.009_{10}-1$	$8.215_{10}-3$
$C_1, C_3, C_4, C_5, C_6, B_1, B_2, B_3, B_4, B_6$	C_5	$3.255_{10}-1$	$9.038_{10}-4$
$C_1, C_3, C_4, C_6, B_1, B_2, B_3, B_4, B_6$	B_4	$3.627_{10}-1$	$1.028_{10}-4$
$C_1, C_3, C_4, C_6, B_1, B_2, B_3, B_6$	C_6	$4.235_{10}-1$	$8.187_{10}-6$
$C_1, C_3, C_4, B_1, B_2, B_3, B_6$	C_1	$4.889_{10}-1$	$5.491_{10}-6$
$C_3, C_4, B_1, B_2, B_3, B_6$	B_3	$5.770_{10}-1$	$5.537_{10}-6$
C_3, C_4, B_1, B_2, B_6	B_1	$6.426_{10}-1$	$4.021_{10}-8$
C_3, C_4, B_2, B_6	B_2	$6.892_{10}-1$	$1.028_{10}-9$
C_3, C_4, B_6	B_6	$7.137_{10}-1$	$5.805_{10}-11$
C_3, C_4	C_4	$7.172_{10}-1$	$2.385_{10}-12$
C_3		$7.403_{10}-1$	$2.539_{10}-10$

The results are summarized in Tables 1–3. In all tables the values of Λ are given as well as the probabilities $\Pr(F > F_{\text{calc}})$. In Table 1 the process of selection with the classical Λ is presented. The values of Λ increase in the process of selection, so the discriminatory power of subsets considered decreases. From the distributional point of view the best is the subset of 2 variables, but we ought to remember that the assumption on normality is not fulfilled for the binary variables. Applying the distributional approach we use that assumption in the considerable way, thus we should carefully analyze the

TABLE 3. Variable selection with the approximative distribution of Λ in the location model

Subset	Variable to be deleted	Value of the Wilks Λ	$\Pr(F > F_{calc})$
$C_1, C_2, C_3, C_4, C_5, C_6, B_1, B_2, B_3, B_4, B_5, B_6, B_7, B_8$	C_2	$2.487_{10}-1$	$1.485_{10}-1$
$C_1, C_3, C_4, C_5, C_6, B_1, B_2, B_3, B_4, B_5, B_6, B_7, B_8$	C_5	$2.739_{10}-1$	$3.110_{10}-2$
$C_1, C_3, C_4, C_6, B_1, B_2, B_3, B_4, B_5, B_6, B_7, B_8$	B_8	$3.064_{10}-1$	$3.622_{10}-3$
$C_1, C_3, C_4, C_6, B_1, B_2, B_3, B_4, B_5, B_6, B_7$	B_3	$3.336_{10}-1$	$7.263_{10}-4$
$C_1, C_3, C_4, C_6, B_1, B_2, B_4, B_5, B_6, B_7$	B_5	$4.325_{10}-1$	$3.472_{10}-5$
$C_1, C_3, C_4, C_6, B_1, B_2, B_4, B_6, B_7$	C_1	$4.422_{10}-1$	$3.190_{10}-6$
$C_3, C_4, C_6, B_1, B_2, B_4, B_6, B_7$	B_7	$4.877_{10}-1$	$5.056_{10}-7$
$C_3, C_4, C_6, B_1, B_2, B_4, B_6$	B_6	$5.061_{10}-1$	$1.477_{10}-7$
$C_3, C_4, C_6, B_1, B_2, B_4$	B_1	$5.961_{10}-1$	$2.117_{10}-8$
C_3, C_4, C_6, B_2, B_4	B_2	$6.422_{10}-1$	$4.702_{10}-10$
C_3, C_4, C_6, B_4	C_6	$6.751_{10}-1$	$1.414_{10}-11$
C_3, C_4, B_4	B_4	$6.968_{10}-1$	$9.232_{10}-12$
C_3, C_4	C_4	$7.172_{10}-1$	$2.385_{10}-12$
C_3		$7.403_{10}-1$	$2.539_{10}-10$

results obtained. The process of selection with the Λ modified to the location model case is presented in Table 2. The values of Λ also decrease, though theoretically this need not be always fulfilled. From the distributional point of view the best result is obtained also for 2 variables. In Table 3 the selection with the approximative distribution of Λ in the location model is presented. The last two variables are the same as in the previous tables, but the whole process of selection has given different results. This is confirmed by the presentation of Table 4 where the results of reclassification for the subsets of 9,

TABLE 4. Comparison of reclassification results

Method of selection	Subset	Number of misclassifications
Classical Λ	$C_1, C_2, C_3, C_4, C_6, B_2, B_3, B_5, B_6$	54
	$C_1, C_3, C_4, B_2, B_3, B_6$	46
	C_3, C_4, B_6	42
Λ in the location model	$C_1, C_3, C_4, C_6, B_1, B_2, B_3, C_4, B_6$	51.5
	$C_3, C_4, B_1, B_2, B_3, B_6$	46
	C_3, C_4, B_6	42
Distributional approach in the location model	$C_1, C_3, C_4, C_6, B_1, B_2, B_4, B_6, B_7$	43
	$C_3, C_4, C_6, B_1, B_2, B_4$	39
	C_3, C_4, B_4	39

6 and 3 variables chosen by three compared methods are given. The reclassification of the sample has been performed by the leaving-one-out method with the smoothed estimates of means and covariances obtained with

the linear additive model of the first order. When the reclassification is not possible (see [9]), 0.5 is added to the total number of misclassifications. The numbers of misclassifications for the subsets chosen by the classical Λ and Λ in the location model are similar. The distributional approach has given considerably better results (e.g., for 6 variables – 46 and 39 incorrectly classified individuals). It should be stressed that not only the numbers of misclassifications are lower but also the probabilities $\Pr(F > F_{\mathrm{calc}})$ (compare Tables 2 and 3) are smaller for the subsets chosen by the distributional approach (the probabilities to be compared are underlined). Thus the latter should be recommended, even when (as in our example) the assumption on normality of continuous variables is not fulfilled.

5. Conclusions. Krzanowski [12] proposing the method of selection based on the comparisons of distance measures (especially defined for mixtures of continuous and discrete variables in [11]) has written that the lack of possibility to judge the change of the distance when deleting continuous variables is its main disadvantage. The reason is that the distribution of the distance measure defined in [11] is not known. This disadvantage is overcome when using the Wilks Λ statistic to the problem of choosing the most discriminative subset. Thus both continuous and discrete variables may be included to the selection process. It should be stressed that the subset chosen can give better results of reclassification than the whole set considered, differently than in the case of discriminating with continuous variables, only.

References

[1] H. Ahrens und J. Läuter, *Mehrdimensionale Varianzanalyse*, Akademie-Verlag, Berlin 1974.

[2] J. J. Daudin, *Selection of variables in mixed-variable discriminant anaylsis*, Biometrics 42 (1986), pp. 473–482.

[3] S. J. Haberman, *Log linear fit for contingency tables. Algorithm AS51*, Applied Statistic 21 (1972), pp. 218–225.

[4] E. Krusińska, *About some generalizations of location model for discrimination with mixed continuous and discrete variables*, pp. 193–203 in: *Proceedings of DIANA II Conference*, Prague 1986.

[5] – *Variable selection in location model for mixed variable discrimination: a procedure based on total probability of misclassification*, EDV in Medizin und Biologie 19 (1988), pp. 14–18.

[6] – *Variable selection in location model for mixed variable-discrimination – a comparative study*, *Data Analysis and Informatics. V*, North-Holland, Amsterdam 1988, 57–67.

[7] – *New procedure for selection of variables in location model for mixed variable discrimination*, Biometrical J. (to appear).

[8] – *Linear and quadratic classification rule in the location model. A comparison for heterogeneous data*, Biocybernetics and Biomedical Engineering 31 (1989), 511–523.

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

[10] – *Mixtures of continuous and categorical variables in discriminant analysis*, Biometrics 36 (1980), pp. 493–499.

[11] — *Distance between populations using mixed continuous and categorical variables*, Biometrika 70 (1983), pp. 235–243.

[12] — *Stepwise location model choice in mixed-variable discrimination*, Applied Statistics 32 (1983), pp. 260–266.

[13] — *Multiple discriminant analysis in the presence of mixed continuous and categorical data*, Comput. Math. Appl. 12a (1986), pp. 179–185.

[14] C. R. Rao, *Linear Statistical Inference and Its Applications*, Wiley, New York 1965.

EWA KRUSIŃSKA
INSTITUTE OF COMPUTER SCIENCE
UNIVERSITY OF WROCŁAW, 51-151 WROCŁAW

JERZY LIEBHART
DEPARTMENT OF INTERNAL DISEASES
MEDICAL ACADEMY OF WROCŁAW, 50-477 WROCŁAW

Received on 1989.01.30