Listy Biometryczne - Biometrical Letters
Vol.XXVI (1989), No.1,2, 29-39

THE INFLUENCE OF VARIABLE SELECTION TO IDENTIFICATION RESULTS OBTAINED VIA LOCATION MODEL TECHNIQUE

Ewa Krusińska, Teresa Ciesielska

Institute of Computer Science, University of Wrocław,
51-151 Wrocław, Przesmyckiego 20

Summary

The paper presents the influence of variable selection in location model to the results of identification. The selection of variables is performed by the Akaike information criterion using the modified method of Daudin (1986). It is compared with the procedure of Krzanowski (1983). For some medical data it appeared that the identification results after the model choice may be better than for the complete set of predictor variables. Thus the variable selection should be recommended before discrimination.

1. INTRODUCTION

The discrimination problem, i.e., the problem of identifying the individual described by the observational vector of predictor variables and assigning it to one of several populations considered on the basis of the observed values of these variables is common in many practical applications. The predictor variables are often of both a continuous and a discrete character. Besides nonparametric methods and logistic discrimination the location model approach has been developed to solve the problem. Introduced by Krzanowski (1975) to the dichotomous discrimination with both continuous and binary variables it assumes the multivariate normal distribution for continuous variables with the common covariance matrix but the means different for all groups and cells of the contingency table defined by binary variables values. The generalizations of the method to mixtures of continuous and discrete variables with more than two states (by transformation to the series of binary variables) and to the polychotomous problem are possible (Krzanowski (1980), (1986), Krusińska (1988)). The appropriate model choice is an important problem in mixed-variables discrimination. The procedure of Krzanowski (1983) enables

Key words: Akaike criterion, location model, variable selection

to choose the discrete variables to the model. The procedure of Daudin (1986) is more general one. These methods are compared in the paper basing on the example of medical data.

2. LOCATION MODEL

The model introduced by Krzanowski (1975) to the dichotomous problem (g=2) assumes that each individual drawn out of population π_1 or π_2 is described by the vector $\boldsymbol{x}$ of q binary variables and the vector $\boldsymbol{y}$ of p continuous ones. Binary variables define the contingency table of $k=2^q$ cells. The multivariate normal distribution $N(\mu_i^{(m)}, \Sigma)$ with the mean vectors $\mu_i^{(m)}$ (i=1,2; m=1,2,...,k) different in each cell and group and the common covariance matrix Σ is assumed for continuous variables. The problem is in classifying the individual $(\boldsymbol{x}, \boldsymbol{y})$ falling into the mth cell to π_1 or π_2. The optimal classification rule is: classify $\boldsymbol{y}$ to π_1 if

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

and otherwise to π_2 (where p_{1m}, p_{2m} are the cell probabilities). Thus it is equivalent to the linear Fisherian discrimination but performed separately for each cell of the contingency table defined by binary variables values. Therefore it is easily generalized to the polychotomous case (g>2) (Krusińska (1988)).

Because the parameters of the model are not known in practice they should be estimated from the data. The cell probabilities p_{im} (i=1,2,...,g; m=1,2,...,k) are estimated after assuming the log-linear model by the iterative scaling algorithm of Haberman (1972) which allows for empty cells in the contingency table. For the parameters related to continuous variables, i.e. $\mu_i^{(m)}$ the linear additive model is assumed to obtain the smoothed estimates:

$$\mu_i = \nu_i + \sum_j \alpha_{j,i} x_j + \sum_j \sum_k \beta_{jk,i}\, x_j x_k + \ldots + \delta_{1\ldots q,i} x_1 \cdot x_2 \cdot \ldots \cdot x_q \quad . \qquad (2)$$

The conditional mean vector $\mu_i^{(m)}$ is obtained by inserting the values of binary variables corresponding to the mth cell to the right side of the formula (2).

Commonly the first order model (only with the main effects of binary variables) or the second order model (additionally with the first order interactions) are used. Thus the group effect ν_i, the main effects $\alpha_{j,i}$ and the interactions $\beta_{jk,i}$ should be estimated. In the present paper we focus only on the first order model. The estimation procedure is as follows.

Let us denote

$$u_{ij} = (1, x_{1ij}, x_{2ij}, \ldots, x_{qij})$$

($i=1,2,\ldots,g$; $j=1,2,\ldots,n_i$; n_i - the number of individuals in the ith group; x_{lij} ($l=1,2,\ldots,q$) - the value of the lth binary variable for the jth individual from the ith group)

$$B_i = (\nu_i, \alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{iq}) \qquad (i=1,2,\ldots,g)$$

The estimate of the parameter matrix B_i is given as

$$\hat{B}_i = C_i A_i^{-1}$$

and the estimate of Σ as

$$\hat{\Sigma} = \frac{1}{n_1 + n_2 + \ldots + n_g} \sum_{i=1}^{g} \left(\sum_{j=1}^{n_i} y_{ij} y_{ij}' - \hat{B}_i A_i \hat{B}_i' \right)$$

with the matrices C_i and A_i defined as

$$C_i = \sum_{j=1}^{n_i} y_{ij} u_{ij}' \quad , \quad A_i = \sum_{j=1}^{n_i} u_{ij} u_{ij}' \quad .$$

When identifying the data with the rule (1) we use the "leaving-one-out" method. It means that when classifying the jth individual we use for estimation of parameters the whole sample except for this jth individual. Such a procedure recommended by Krzanowski (1975) enables to obtain the unbiased estimates of misclassification probabilities.

3. STEPWISE LOCATION MODEL SELECTION BASED ON THE DISTANCE MEASURE

Krzanowski (1983) has introduced the procedure of selecting discrete variables to the location model for the two groups of data problem. The algorithm is based on the distance measure

$$\Delta^2 = 2(1-\rho) \tag{3}$$

where

$$\rho = \sum_{m=1}^{k} \{ (p_{1m} p_{2m})^{1/2} \exp[-\frac{1}{8} (\mu_1^{(m)} - \mu_2^{(m)})' \Sigma^{-1} (\mu_1^{(m)} - \mu_2^{(m)})] \} \quad .$$

The distance D^2 (which is an estimate of Δ^2) is obtained after replacing the unknown parameters by their estimates. It can be done using the procedure described in Section 2 but Krzanowski (1983) recommends to use the classical estimates (non-smoothed cell means and covariances) for selection. Then after selecting the model he advices to apply the whole estimation procedure in the identification process.

The main idea of selection is that in the discrimination set these variables should remain which give the greatest distance between the groups examined. Applying the stepwise procedure we act as follows:

1° We have the whole set of discrete (binary) variables $S1 = \{1,2,\ldots,q\}$ in the discrimination set. The actual number of discrete variables equals $r=q$.

2° We calculate distances D^2_j after deleting the jth discrete variable $(j=1,2,\ldots,r)$. We use the whole set of p continuous variables for means and covariances estimation.

3° Let $D^2_{j_o} = \max_j D^2_j$. We eliminate the variable no.j_o from the model. The remaining discrete variables give the greatest distance between π_1 and π_2 possible on the set of r-1 variables obtained in the stepwise way.

4° If $r>1$ we set $r := r-1$, $S1 = S1\setminus\{j_o\}$ and proceed to 2° otherwise we stop the elimination process.

The procedure enables only to select binary (discrete) variables because the distances (3) computed for various subsets of continuous and discrete variables are uncomparable.

Acting as in the algorithm described we can eliminate all discrete variables. Krzanowski (1983) do not introduce any stopping criterion. He chooses the subset of discrete variables heuristically after analysing the whole selection process.

4. SELECTION PROCEDURE BASED ON AKAIKE CRITERION

The method for selection of continuous as well as discrete variables and the interaction terms in the linear additive model for means estimation which is based on the Akaike information criterion was introduced by Daudin (1986). The original Akaike criterion (Akaike (1973)) is given as

$$AIC(i) = l_i - k_i \quad , \tag{4}$$

where l_i is the log-likelihood for the ith model considered, k_i is the number of degrees of freedom for the ith model, i.e., the number of independent parameters in the model.

Among the variants of selection given by Daudin (1986) we use the following two suitable for the first order model:

a) select binary variables, then continuous ones;

b) first select continuous variables, then binary ones.

The selection of binary variables is performed with the criterion (4) especially adopted to location model case. It is given as

$$AIC = n - \log|\hat{\Sigma}_s^{(0)}| - 2\,u\,s \quad , \tag{5}$$

where $\hat{\Sigma}_s^{(0)}$ is the estimate of the covariance matrix Σ for the general model with the group factor (obtained by the procedure described in Section 2), s is the number of continuous variables in the current discrimination set, n is the number of individuals in the sample, u is the number of degrees of freedom in the model of one continuous variable.

Let us consider r binary variables. The value of u is simply obtainable (Ciesielska (1988)). It is so because the location model may be written in the terminology of the multivariate analysis of variance. In the case of the first order model we should consider the effects of each binary variable, the group factor and interactions between the group factor and binary variables. This gives $r + (g-1)(r+1)$ degrees of freedom.

The optimal subset $S1_{opt}$ of binary variables is such that

$$AIC(S1_{opt}) = \sup_{S1} AIC(S1) \quad .$$

We use the stepwise algorithm to obtain semi-optimal subset. It is as follows:

1^o Let $S1 = \{1,2,\ldots,q\}$. The actual number of binary variables equals $r=q$.

2^o We calculate the value $AIC(j)$ for all subsets obtained after deleting the jth binary variable from $S1$.

3^o We eliminate variable number j_o for which we have

$$AIC(j_o) = \sup_{j} AIC(j)$$

and

$$AIC(j_o) < AIC_o \quad .$$

4^o If $r>1$ and if it exists such j for which $AIC(j) < AIC_o$ we set $r := r-1$, $S1 = S1\setminus\{j_o\}$ and proceed to 2^o otherwise we stop the selection process.

The value AIC_o is the upper bound of the value of AIC. If it is surpassed for all subsets analysed in 3^o we assume that the subset found has sufficiently large value of AIC and is already sufficiently good for discrimination.

In the case of selection of continuous variables the values of AIC are not directly comparable and therefore Daudin (1986) has used the modified criterion DAIC which may be expressed in the form:

$$DAIC = 2(AIC(0) - AIC(1)) \tag{6}$$

where $AIC(0)$ is the Akaike criterion for the full model with the group factor on the set of the s considered continuous variables, $AIC(1)$ is the Akaike criterion for the model without the group factor and without all interactions in which this factor is involved.

The location model is easily written in the terminology of multivariate analysis of variance, thus its log-likelihood is given as

$$l = -\frac{1}{2}\,(n \log |\hat{\Sigma}| + n\,s \log(2\pi) + s) \quad . \tag{7}$$

Using formulae (4) and (6) the modified criterion DAIC may be written as

$$DAIC = -2(l_1 - l_0 + k_0 - k_1) \quad ,$$

where l_i, k_i are the log-likelihood of the model (i) and the number of degrees of freedom in (i) (i=0,1).

Applying (7) DAIC may be explicitly written as

$$DAIC = n \log(|\hat{\Sigma}^{(1)}|/|\hat{\Sigma}^{(0)}|) - 2\,s(k_0 - k_1) \quad . \tag{8}$$

It may be easily seen (Ciesielska (1988)) that the difference $t = k_0 - k_1$ of degrees of freedom between the model (0) with the group factor and its interactions with binary variables and model (1) without these terms equals for the set of s continuous variables

$$t = s(g-1)(1+r) \quad ,$$

because the number of degrees of freedom for the model (1) equals sr (r is the number of binary variables in the current model).

The optimal subset S_{opt} of continuous variables is such that

$$DAIC(S_{opt}) = \sup_S DAIC(S) \quad .$$

As previously we use the stepwise algorithm to obtain the semi-optimal subset. It is the modification of the algorithm of Daudin (1986) and is as follows (Ciesielska (1988)):

1° Let $S = \{1,2,\ldots,p\}$. The actual number of continuous variables is $s=p$.

2° We calculate the value $DAIC(j)$ for all subsets obtained after deleting the jth continuous variable from S.

3° We eliminate variable no.j_o for which we have

$$DAIC(j_o) = \sup_j DAIC(j) \quad \text{and} \quad DAIC(j_o) < DAIC_o$$

4° If it exists such j for which $DAIC(j) < DAIC_o$ we set $s := s-1$ and $S = S \setminus \{j_o\}$. If also $s>1$ we proceed to 2° otherwise we stop the elimination process.

The value $DAIC_o$ is upper bound for DAIC. If it is surpassed for all subsets analysed in 3° we understand that the subset found is already sufficiently good for discrimination. We used $DAIC_o$ and AIC_o equal to 0. As indicated in the step 4° of the above algorithm at least one continuous variable should remain in the discrimination set especially when we then want to perform the elimination of binary variables.

5. EXAMPLE OF APPLICATION

The material used in the study was described by Krzanowski (1975) as the fourth data set. It concerned medical data of 186 subjects with advanced breast cancer who underwent the ablative surgery. In 99 cases the treatment was "non-failure", in 87 remaining cases it was "failure". Six continuous variables and three binary ones were measured.

The results of selection are presented in Table 1 and 2. Table 1 concerns the selection process in the sequence: first binary, then continuous variables. On the complete set of continuous variables the binary variables do not possess a sufficient discriminatory power (AIC

Table 1. Selection on the basis of Akaike criterion in the sequence: first binary, then continuous variables

continuous variables	binary variables	binary variable deleted	AIC
1,2,3,4,5,6	1,2,3	3	$-8.36_{10}\ 3$
1,2,3,4,5,6	1,2	2	$-8.23_{10}\ 3$
1,2,3,4,5,6	1	1	$-8.10_{10}\ 3$
1,2,3,4,5,6	-		$-7.92_{10}\ 3$
		continuous variable deleted	**DAIC**
1,2,3,4,5,6	-	4	$-4.93_{10}\ 1$
1,2,3,5,6	-	1	$-2.78_{10}\ 1$
2,3,5,6	-		$-9.86_{10}\ 0$

lower than - 8000). Thus they are deleted. Then from the complete set of continuous variables features no.4 and 1 are eliminated. The subset consisted of continuous variables no.2,3,5,6 remains. After eliminating succeeding variable the value of DAIC is already positive. The first part of Table 1 concerns also the problem of eliminating only binary variables. Because they do not possess discriminatory ability and all are eliminated, the continuous variables no.1-6 remain in the discrimination set. Table 2 concerns the elimination process in the sequence: first continuous then binary variables. In this case at least one continuous variable should remain in the discrimination set at the first stage of selection. This is the variable no.2. On this feature binary variables have sufficient discriminatory ability and they all remain in discrimination set, thus in

Table 2. Selection on the basis of Akaike criterion in the sequence: first continuous, then binary variables

continuous variables	binary variables	continuous variable deleted	DAIC
1,2,3,4,5,6	1,2,3	4	$-2.51_{10}\ 2$
1,2,3,5,6	1,2,3	1	$-1.69_{10}\ 2$
2,3,5,6	1,2,3	3	$-9.81_{10}\ 1$
2,5,6	1,2,3	6	$-4.49_{10}\ 1$
2,5	1,2,3	5	$-1.00_{10}\ 1$
2	1,2,3		$-2.73_{10}\ 0$
		binary variable deleted	**AIC**
2	1,2,3		$4.13_{10}\ 1$

fact the elimination of binary variables is not performed. The first part of Table 2 concerns also the case when we eliminate only continuous variables, saving all binary in the discrimination set. We have checked the discriminatory ability for the subset of continuous variables no. 2,3,5,6 and all binary variables with the value of DAIC over 2.5 times greater than for the complete set. We have also examined the subset of continuous variables no.2 and 5 and all binary variables for which the DAIC value increases almost 4.5 times in comparison with the set of three continuous variables.

The identification results are given in Table 3. It consists of four parts. The first part concerns the results of identification using the linear discriminant function and the location model of the first order on the complete set of continuous and binary variables. The location model is superior in comparison with the linear discrimination. The second part of the table concerns identification on subsets chosen in various variants by Akaike criterion. The best results of 58 and 59 misclassifications (better than for the complete set) are obtained for subset of continuous variables no. 2,3,5,6 and binary variables no.1-3 and for continuous variables no.2,5 and all binary variables, too. Besides it should be stressed that the linear discrimination on the set of only continuous variables has given 68 misclassifications (when for the complete set - 71). The subset of variables no.2,3,5,6 chosen in Table 1 is even better (66 misclassifications) and at the level of the full location model of the first order (64 misclassification). In the third part of Table 3 the

Table 3. Identification results

Method	Continuous variables	Binary variables	Misclassifications π_1	π_2	together
linear discriminant function*	1,2,3,4,5,6	1,2,3	41	30	71
location model of the first order	1,2,3,4,5,6	1,2,3	34	30	64
location model of the first order	2,3,5,6	1,2,3	32	26	58
	2,5	1,2,3	27	32	59
	2	1,2,3	25	51	76
location model reduced to the linear discriminant function	1,2,3,4,5,6	-	34	34	68
	2,3,5,6	-	33	33	66
location model of the first order	1,2,3,4,5,6	1,2	40	27	67
ideal point discriminant analysis**	2	1,3	30	27	57

* after Krzanowski (1975)

** after Takane et al. (1987)

results on the subset obtained by Krzanowski (1983) are reported but for the first order location model (he only gives the outcomes for the second order model which is not studied in the present paper; compare also Krzanowski (1975)). They are better than the linear discrimination but worse than the identification on the subsets chosen by Akaike criterion. For comparison the fourth part of the table concerns the results of the so called ideal point discriminant analysis introduced by Takane et al. (1987). This is the kind of extension of logistic discrimination with the choice of variables by Akaike criterion defined especially for that model. Using it the continuous variable no.2 and binary variables no.1 and 3 were saved in the discrimination set. The number of misclassifications equaled 57 and was at the level of the location model chosen also by Akaike criterion. So it is seen that it is not the superiority of ideal point discriminant analysis and that the appropriate model choice (indicated by Takane et al. (1987) as the considerable advantage of their method) is also possible with the location model technique.

6. CONCLUSIONS

The results obtained indicate that the model choice should be performed before identification. It concerns the linear discrimination as well as the location model, because the noninformative "noise" can contaminate the results of classification. The best location models chosen have given the identification results at the level of the newly developed ideal point discriminant analysis for the subset also selected by Akaike criterion.

ACKNOWLEDGEMENTS

The authors are greatly indebted to Dr. W.J. Krzanowski from the University of Reading, Great Britain for the permission to use his data set in the study.

The paper was supported by the grant I.07 CPBP 02.20.

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, 2nd *International Symposium on Information Theory*, Ed. Petrov B.N., Budapest.

Ciesielska, T. (1988). Redukcja liczby zmiennych w modelu lokacyjnym na podstawie kryterium informacyjnego Akaike. *M.S.thesis, Institute of Computer Science*, University of Wrocław.

Daudin, J.J. (1986). Selection of variables in mixed-variable discriminant analysis. *Biometrics* **42**, 473-482.

Haberman, S.J. (1972). Log-linear fit for contingency tables. Algorithm AS51. *Applied Statistics* **21**, 218-225.

Krusińska, E. (1988). Linear and quadratic classification rule in the location model. A comparison for heterogeneous data. *Biocybernetics and Biomedical Engineering* (to appear).

Krzanowski, W.J. (1975). Discrimination and classification using both binary and continuous variables. *J. Amer. Statist. Ass.* **70**, 782-790.

Krzanowski, W.J. (1980). Mixtures of continuous and categorical variables in discriminant analysis. *Biometrics* **36**, 493-499.

Krzanowski, W.J. (1983). Stepwise location model choice in mixed variables discrimination. *Applied Statistics* **32**, 260-266.

Krzanowski, W.J. (1986). Multiple discriminant analysis in the presence of mixed continuous and categorical data. *Computers and Mathematics with Applications* **12A**, 179-185.

Takane, Y., Bozdogan H., Shibayama T. (1987). Ideal point discriminant analysis. *Psychometrika* **52**, 371-392.

Received 1 August 1988: revised 30 January 1989

WPŁYW WYBORU ZMIENNYCH NA REZULTATY IDENTYFIKACJI UZYSKIWANE PRZY UŻYCIU MODELU LOKACYJNEGO

Streszczenie

W pracy przedyskutowano wpływ wyboru zmiennych w modelu lokacyjnym na wyniki identyfikacji. Wybór zmiennych przeprowadzono przy użyciu kryterium informacyjnego Akaike za pomocą zmodyfikowanej metody Daudina (1986). Została ona porównana z procedurą Krzanowskiego (1983). Na przykładzie danych medycznych pokazano, że wyniki klasyfikacji dla wybranego modelu mogą być lepsze niż dla kompletnego zbioru zmiennych objaśniających. Dlatego też wybór zmiennych powinien być przeprowadzony przed dyskryminacją.