SELECTION OF VARIABLES AND MULTIVARIATE ANALYSIS OF MULTI-CLASSIFIED NON-ORTHOGONAL DATA

BY

S.C. AGARWAL AND S.N. KAUSHIK

IVRI, Izatnagar—243122

(Received : November 1980)

SUMMARY

The methods for selecting the important discriminating variables and analysing the multiclassified multivariate non-orthogonal data as a factorial experiment have been described in the present paper which were hitherto lacking in livestock researches by now. The methodology so described has also been used to analyse the data on Hariana taurus crosses of cattle.

INTRODUCTION

The multivariate analysis, which elucidates the structure of entire set of data, has widely been utilized in livestock researches by Taneja (10), Narain and Garg (7) and Acharya and Mani Mohan (1) on the lines of Rao (8) in order to compare different groups which is other wise not possible through univariate analysis due to lack of consideration of dependence between a number of variables. In these studies the multi-classified data were converted into a single classification so as to follow discriminant or D^2-approach available for one way classified data by adjusting the data for effects of other factors. This is, however, not valid when some interactions among the factors exist, which is very usual in livestock experiments. Also no consideration was given in these studies to select out important variables from a number of variables as emphasised by Lubischew (6) as could be seen from the work of Taneja *et. al.* (11) and Bhat *et. al.* (2) where the same study was repeated twice by taking into consideration 9 and 3 variables each at a time respectively.

This work aims at describing the application of the method of selecting important discriminating variables, presented by Horton *et. al.* (4) and Kendall (5) in case of balanced (equal sub-class numbers) data and of multivariate analysis of multi-classified non-orthogonal data after developing the multivariate extension of Federer and Zelen's (3) univariate ANOVA method and using the multiclassification extension of discriminant analysis by Tatsuoka (12) in a cattle crossbreeding experiment for comparing halfbreds. The method, described here, requires that all subclasses due to effects are filled. Though we describe exclusively the statistical reasoning but other subject matter criteria may also be considered while selecting important discriminating variables. The paper only demonstrates multivariate procedure and does not necessarily consider the animal breeding aspects in selection of variables. Further, the statistical properties of unbiasedness and efficiency by the recommended procedure are not known.

2. Data Used

Data on characters (1) birth weight, (2) age at first calving (3) first lactation milk yield, (4) first lactation length and (5) first calving interval observed on 98 animals of Friesian *x* Hariana ($F\times H$), 68 animals of Brown Swiss *x* Hariana ($B\times H$) and 82 animals of Jersey *x* Hariana ($J\times H$) maintained at this Institute under All India Coordinated Research Project on Cattle during 1970-75 were used. Fifty animals, 18 in $F\times H$, 22 in $B\times H$ and 10 in $J\times H$ crossbreds not having information on these five characters were discarded from this study. The data fell into $3\times 2\times 5$ classifications due to factorial effects of breed groups (3), period of births (2) and season of births (5). The two periods consisted of 1970-72 and 1973-75 respectively because of similar managemental practices over these years. The year was delineated into five seasons based on climatic conditions classified as C_1 (Dec.-Feb.), C_2 (Mar-Apr.), C_3 (May-June), C_4 (Jul.-Sep.) and C_5 (Oct.-Nov.)

3. Methods

3.1. Selection of Variables

The data was considered initially as one way classified into k $(=a\times b\times c)$ cells. The between cells and within cells *S.S.C.P.* (sums of squares and sums of products) matrices were computed using the multivariate analogue of between and within cells *S.S.* in univariate one way classifications with unequal number of observations. The interative test procedure (Horton (4) ; Kendall (5)) for selecting the important discriminating variables, which is des:ribed below for completness, was used.

Let S_w denote the within cells *S.S.C.P.* matrix and S_b the between cells *S.S.C.P.* matrix for all variables. Let $S=S_w+S_b$. Partition the matriccs S_w, S_b and S as follows :

$$S=\begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} = \begin{bmatrix} S_{b11} & S_{b12} \\ S_{b21} & S_{b22} \end{bmatrix} + \begin{bmatrix} S_{w11} & S_{w12} \\ S_{w21} & S_{w22} \end{bmatrix}$$

where the suffix "1" pertains to the variable (*s*) to be included in the retained subset and the suffix "2', to those whose usefulness as discrimators is to be tested, and from the residual matrices

$$R=S_{22}-S_{21}\ S_{11}^{-1} S_{12}$$

$$R_w=S_{w22}-S_{w21}\ S_{w\,11}^{-1}\ S_{w12}$$

Let L denote the ratio of the determinants of R_w and R i.e.

$$L=|R_w|/|R|$$

The test criterion is then

$$C=-(N-K-M)\log_e L$$

The statistic is approximately distributed as X^2 with $(k-1)p$ degree of freedom, where N is number of measurement vectors, k is the number of cells, p is the dimension of complementry subset (matrix S_{22}) and M is the dimension of retained set (matrix S_{11}). A significant value of criterion C indicates that the complementry subset still includes variables with independent power to differentiate between the cells.

The computation beigns with the smallest subset of independent variables *e.g.* variable I is taken in retained set and $(p-1)$ remaining variates in complementry subset and criteria C is computed and then the same procedure is followed for each variable subsequently and a variable is selected which leaves the smallest residual, as indicated by the smallest value of the criterion. if the value of the criterion of the selected subset is significant, a further variable is added to the subset of independent variable (*s*) using the same procedure as before. This cycle of operations is terminated when no significant residual between cells variance remains.

The validity of set of independent variables as important discriminators thus selected in this study was also verificd from the pooled within cells product moment correlations.

3.2. Comparison of Groups

After selecting the variables the data were analysed for factorial effects and the *S.S.C.P.* matrices for main effects of breed groups (*A*), periods (*B*), seasons (*C*) and their respective interactions were computed using the multivariate analogue of univariate *ANOVA* method of Federer and Zelen (3), which was developed for this study and is given in Appendix. For testing the significance of effects, multivariate *ANOVA* (*MANOVA*) summary table following Tatsouka (12, Table 7.1) was formed as below :

Multivariate ANOVA (MANOVA) Table

Source	*d.f.*	*S.S.C.P. matrix*	Λ
A	$a-1$	S_A	$\mid S_W \mid / \mid S_W+S_A \mid$
B	$b-1$	S_B	$\mid S_W \mid / \mid S_W+S_B \mid$
C	$c-1$	S_C	$\mid S_W \mid / \mid S_W+S_C \mid$
$A\times B$	$(a-1)(b-1)$	S_{AB}	$\mid S_W \mid / \mid S_W+S_{AB} \mid$
$B\times C$	$(b-1)(c-1)$	S_{BC}	$\mid S_W \mid / \mid S_W+S_{BC} \mid$
$A\times C$	$(a-1)(c-1)$	S_{AC}	$\mid S_W \mid / \mid S_W+S_{AC} \mid$
$A\times B\times C$	$(a-1)(b-1)(c-1)$	S_{ABC}	$\mid S_W \mid / \mid S_W+S_{ABC} \mid$
Within cells	$N-abc$	S_W	

$\mid M \mid$ denotes the determinant of *M*-matrix and significance of each g^{th} ($g=A, B, C, A\times B, B\times C, A\times C, (A\times B\times C)$) effect was tested using the multi-classification extension (Tatsouka (12), Sec. 7.2) of Rao's (8) approximate test criteria

$$R=\frac{1-\Lambda^{1/s}}{\Lambda^{1/s}}\cdot\frac{ms-\frac{1}{2}p.\,V_g+1}{p.V_g}$$

where, p=number of variables

V_g=degree of freedom for g^{th} effect

V_w=within cells degree of freedom

$m=V_w+V_g-\frac{1}{2}(p+V_g+1)$

$$S^2=\frac{(p\,V_g)^2-4}{p^2+V_g^2-5}$$

The statistic R is approximately distributed as an F-variate with $p\,V_g$ and $ms-\frac{1}{2}p\,V_g+1$ degree of freedom.

After getting the test of significance in MANOVA and finding some effects significant, the mean vectors for each level of the effects were calculated using the estimated effects (see Appendix). For each significant effect the discriminant functions, which in numbers are equivalent to the degree of freedom of the effect in MANOVA table or the number of variables in study whichever is less, were computed following Tatsuoka (12, Sec. 7.4) by solving the characteristic equation

$$(S_w^{-1}\,S_g-\lambda I)V=0 \qquad \ldots(3.2.1)$$

where S_w and S_g are within cells and among g^{th} ($g=A, B, C$) effect $S.S.C.P.$ matrix given in MANOVA table, and λ is eigen value and V eigen vector of dimension p (number of variables) for each discriminant function, the number of non-zero eigen values and eigen vectors being equal to number of discriminant functions for each g^{th} effect. The value of eigen vectors were calculated following Tatsuoka (12, Sec. D.1) and Searle (9, Ch. 7). The discriminant scores, also termed sometimes as compound values, for each level of g^{th} effect were obtained putting the corresponding mean vector value in g^{th} effect's discriminant functions. Then the Mahalanobis distance-D^2 between two levels of each g^{th} effect were computed as the sum of the differences of the corresponding discriminant scores over V_g number of discriminant functions. Significance of D^2-values were tested using the statistic

$$\frac{N_1 N_3 \,(V_w - p + 1)}{p(N_1 + N_2)\; V_w} D^2$$

which follows F-distribution with p and V_w degrcee of freedom. N_1 and N_2 are the sample sizes for the two populations (levels of an effect) characterized by p variables and V_g and V_w defined earlier.

4. Results

4.1. Selection of Variables

The important discriminating variables, obtained after the method described above, included (1) birth weight, (2) age at first calving and (3) first lactation length (Table 1). In first iteration birth weight (Var. No. 1) with the smallest and statistically significant value of criterion C was moved to the top of the next column so as to add another independent variable in the subset. In the second iteration the subset of birth weighl and age at first calving (Var. No. 1 and 2) having the smallest and significant value of criterion were moved to the top of the next column. Similarly in third iteration lactation length along with the two earlier selected variables was selected. Since the value of criterion for the selected subset consisting of variables (No. I, 2 and 4) was non-significant, the cycle of operations for selecti on any more independent variable was terminated.

The similar selection of variables was observed from the pooled within cells correlations (Table 2) where birth weight (Var. 1) is uncorrelated with all variables and so it is to be selected as first independent variable, and second independnet variable seems to be age at first calving (Var. 2) which is uncorrelated wrth all the variables except milk yield (Var. 3) and the third independent variable is lactation length which is uncorrelated with birth weight, age at first calving and significantly correlated with milk yield and first calving interval.

4.2. Comparison of Groups

Testing the factorial effects by MANOVA (Table 3) revealed that effects due to breed groups, periods and seasons of birth were significant and so also the interactions due to breed groups x seasons, periods x scasons and breeds x periods x seasons were significant ($p < 0.05$).

Having shown the existence of significant differences due to breed groups, periods and seasons, the eigen vectors corresponding to each significant effects were computed using equation (3.2.1), Sec. 3.2 and thus the discriminant functions were obtained. The discriminant functions numbered for breed groups and periods as 2 and 1 respectively equal to its degree of freedom and for seasons as 3 equal to the number of variables characterising the discriminant function. The distances between two groups of each factor (breeds and periods) were estimated by computing the three variable mean veetors and the discriminant scores on the lines described earlier. The number of scores for genetic groups, periods and seasons of birth were obtained as 2, 1 and 3 respectively. The Mahalanobis distance D^2 were then calculated (Table 4) which revealed the Friesian and Brown Swiss halfbred groups not to differ significantly ($p<0.05$) from each other but the Jersey halfbreds significantly differed from either of them ($p<0.01$) due to inherent breed differences in the 3 exottc breeds *viz.* Holstein Friesian, Brown Swiss and Jersey used on Hariana.

The period I (1970-72) appeared better than period 2 (1973-75) obviously because of lesser number of animals in the former than the latter under the same managemental regime and hence better supervision and care.

All the seasons of birth except in the season I and 3 differed significantly from each other ($p<0.05$) indicating thereby marked seasonal variation in the inputs and the general managerial conditions, be either due to feeds/fodder supply or other environmental fluctuations.

ACKNOWLEDGEMENT

The authors are thankful to the referees for their help to improve the original manuscript of this paper.

REFERENCES

[1] Acharya, R. M. and Mani Mohan (1979) : Genetic considerations in crossbreeding for evolving new breeds of sheep-Results of a crossbreeding experiment. The *Ind. J. Anim. Genet, and breed.* 1(1), 37-45.

[2] Bhat, P.N., Taneja, V.K. and Garg, R.C. (1979) : Genetic divergence as a method for classifying Sahiwal and Sahiwal *x* Holstein crossbred grades. The *Indian J. Anim. Genet and Breed.* 1(2) 72-76.

[3] Federer, W.T. and Zelen, M. (1966) : Analysis of multifactor classifications with unequal number of observations. *Biometrics* **22**, 525-552.

[4] Horton, I.F., Russel J.S. and Moore, A.W. (1968) : Multivariate-covariance and canonical analysis. *Biometrics* **24**, 845-858.

[5] Kendall, M.G. (1957) : *A Course in Multivariate Analysis.* Griffin, London.

[6] Lubischew A.A. (1962) : On the use of discriminant functions in Taxonomy. *Biometrics* **18**, 455-476.

[7] Narain, P. and Garg, L.K. (1965) : A possible use of discriminant function and D^2 statistic for comparing different grades of sheep in a crossbreeding programme. The *Indian J. Anim. Sciences* **45**, 243-247.

[8] Rao, C.R. (1952) : *Advanced Statistical Methods in Biometries Research*, Wiley, New York.

[9] Searle, S.R. (1966) : *Matrix Algebra for the Biological Sciences*, Wiley, New York.

[10] Taneja, V.K. (1973) : Genetic analysis of Holstein *x* Zebu crosses. Ph. D. Thesis, *Agra University, Agra.*

[11] Taneja V.K. Bhat, P.N. and Garg, R.C. (1979) : Genetic divergence in various Sahiwal *x* Holstein crossbreed grades. *Theor. Appl. Genet.* **54**, 69-74.

[12] Tatsuoka, M.M. (1971) : Multivariate Analysis : Techniques for Educational and Psychological Research. Wiley, New York.

APPENDIX

MULTIVARIATE ANALOGUE OF UNIVARIATE ANALYSIS OF NON-ORTHOGONAL DATA

Federer and Zelen's (3) method for univariate analysis of data with unequal subclass numbers can be extended in multivariate situation as :

Supposc n_{ijk} is the number of individuals in $(i, j, k,)^{th}$ cell and $X_{hijkl}(h=1, 2, \ldots p;\ i=1, 2, \ldots, a;\ j=1, 2, \ldots, b;\ k=1, 2, \ldots c;\ 1=1, 2, \ldots n_{ijk})$ is the observation of h^{th} variable for 1^{th} individual in $(i. j, k)^{th}$ cell. Let $\bar{X}_{hijk}$ be the mean of the h^{th} variable for n_{ijk} individuals in $(i, j, k)^{th}$ cell. Also let the fixed effects model for the factorial experiment be :

$$E(\bar{X}_{hijk}) \\ =\mu_h+A_{hi}+B_{hj}+C_{hk}+(AB)_{hij}+(AC)_{hik}+(BC)_{hjk}+(ABC)_{hijk}$$

The parameters included in the model can be estimated as :

$$\hat{\mu}_h = \frac{1}{abc}\sum_i\sum_j\sum_k \bar{X}_{hijk} = \bar{X}_{h\cdots}$$

$$\hat{A}_{hi} = \frac{1}{bc}\sum_j\sum_k \bar{X}_{hijk} - \hat{\mu}_h = \bar{X}_{hi\cdot\cdot} - \bar{X}_{h\cdots}$$

$$\hat{B}_{hj} = \frac{1}{ac}\sum_j\sum_k \bar{X}_{hijk} - \hat{\mu}_h = \bar{X}_{h\cdot j\cdot} - \bar{X}_{h\cdots}$$

$$\hat{C}_{hk} = \frac{1}{ab}\sum_i\sum_j \bar{X}_{hijk} - \hat{\mu}_h = \bar{X}_{h\cdot\cdot k} - \bar{X}_{h\cdot\cdot}$$

$$(\hat{AB})_{hij} = \frac{1}{c}\sum_k \bar{X}_{hijk} - \hat{\mu}_h - A_{hl} - B_{hj}$$

$$= \bar{X}_{hij\cdot} + \bar{X}_{h\cdots} - \bar{X}_{hi\cdot\cdot} - \bar{X}_{h\cdot j\cdot}$$

$$(\hat{BC})_{hjk} = \frac{1}{a}\sum_i \bar{X}_{hijk} - \hat{\mu}_h - \hat{B}_{hj} - \hat{C}_{hk}$$

$$= \bar{X}_{h\cdot jk} + \bar{X}_{h\cdots} - \bar{X}_{h\cdot j\cdot} - \bar{X}_{h\cdot\cdot k}$$

$$(A\hat{C})_{hik}=\frac{1}{b}\sum_{j}\bar{X}_{hijk}-\hat{\mu}_{h}-\hat{A}_{hi}-\hat{C}_{hk}$$
$$=\bar{X}_{hi\cdot k}+\bar{X}_{h\cdots}-\bar{X}_{hi\cdot\cdot}-\bar{X}_{h\cdot\cdot k}$$

$$(\hat{A}BC)_{hijk}=\bar{X}_{hijk}-\hat{\mu}_{h}-\hat{A}_{hi}-\hat{B}_{hj}$$
$$-\hat{C}_{hk}-(\hat{A}B)_{hij}-(\hat{B}C)_{hjk}-(\hat{A}C)_{hik}$$
$$=\bar{X}_{hijk}+\bar{X}_{hi\cdot\cdot}+\bar{X}_{h\cdot j\cdot}+\bar{X}_{h\cdot\cdot k}$$
$$-\bar{X}_{hij\cdot}-\bar{X}_{h\cdot jk}-\bar{X}_{hi\cdot k}-\bar{X}_{h\cdots}$$

with Harmonic average number of observations for each effects as :

$$n(\hat{\mu})=\bar{n}_{\cdots}=abc/\sum_{i}\sum_{j}\sum_{k}\frac{1}{n_{ijk}}$$

$$n(\hat{A}_{hi})=\bar{n}_{i\cdot\cdot}=bc/\sum_{j}\sum_{k}\frac{1}{n_{ijk}}$$

$$n(\hat{B}_{hj})=\bar{n}_{\cdot j\cdot}=ac/\sum_{i}\sum_{k}\frac{1}{n_{ijk}}$$

$$n(\hat{C}_{hk})=\bar{n}_{\cdot\cdot k}=ab/\sum_{i}\sum_{j}\frac{1}{n_{ijk}}$$

$$n(\hat{A}B)_{hij}=\bar{n}_{ij\cdot}=c/\sum_{k}\frac{1}{n_{ijk}}$$

$$n(B\hat{C})_{hjk}=\bar{n}_{\cdot jk}=a/\sum_{i}\frac{1}{n_{ijk}}$$

$$n(\hat{A}C)_{hik}=\bar{n}_{i\cdot k}=b/\sum_{j}\frac{1}{n_{ijk}}$$

$$n(AB\hat{C})_{hijk}=n_{ijk}$$

The various sums of sqnares and sums of products needed for *SSCP* matrices S_W, S_A, S_B, S_C, S_{AB}, S_{BC}, S_{AC} and S_{ABC} the within cells, between A groups, between B groups, between C groups and their possible interactions respectively can be computed as :

$$(S_W)_{hm}=\sum_{i}\sum_{j}\sum_{k}\left[\sum_{1=1}^{n_{ijk}}X_{hijkl}X_{mijkl}-\frac{\sum_{l}X_{hijkl}\sum_{l}X_{mijkl}}{n_{ijk}}\right]$$

$$(h,\ m=1,\ 2,\ \ldots p)$$

$$(S_A)_{hm}=b\times c\left[\sum_i \bar{n}_{i..}\,\bar{X}_{hi..}\,\bar{X}_{mi..}-\frac{(\sum_i \bar{n}_{i..}\bar{X}_{hi..})(\sum_i \bar{n}_{i..}\bar{X}_{mi..})}{\sum_i \bar{n}_{i..}}\right]$$

$$(S_B)_{hm}=a\times c\left[\sum_j \bar{n}_{.j.}\,\bar{X}_{h.j.}\,\bar{X}_{m.j.}-\frac{(\sum_j \bar{n}_{.j.}\,\bar{X}_{h.j.})(\sum_j \bar{n}_{.j.}\bar{X}_{m.j.})}{\sum_j \bar{n}_{.j.}}\right]$$

$$(S_C)_{hm}=a\times b\left[\sum_k \bar{n}_{..k}\bar{X}_{h..k}\bar{X}_{m..k}-\frac{(\sum_k \bar{n}_{..k}\bar{X}_{h..k})(\sum_k \bar{n}_{..k}\bar{X}_{m..k})}{\sum_k \bar{n}_{..k}}\right]$$

$$(S_{AB})_{hm}=c\left[\sum_i\sum_j \bar{n}_{ij.}(AB)_{hij}(AB)_{mij}-\sum_i\frac{(\sum_j \bar{n}_{ij.}\,(AB)_{hij})(\sum_j \bar{n}_{ij.}\,(AB)_{mij})}{\sum_j \bar{n}_{ij.}}\right]$$

$$(S_{BC})_{hm}=a\left[\sum_j\sum_k \bar{n}_{.jk}\,(BC)_{h.jk}\,(BC)_{m.jk}-\sum_j\frac{(\sum_k \bar{n}_{.jk}\,(BC)_{h.jk})(\sum_k \bar{n}_{.jk}\,(BC)_{m.jk})}{\sum_k \bar{n}_{.jk}}\right]$$

$$(S_{AC})_{hm}=b\left[\sum_i\sum_k \bar{n}_{i.k}\,(AC)_{hi.k}\,(AC)_{mi.k}-\sum_i\frac{(\sum_k \bar{n}_{i.k}\,(AC)_{hi.k})(\sum_k \bar{n}_{i.k}\,(AC)_{mi.k})}{\sum_k \bar{n}_{i.k}}\right]$$

$$(S_{ABC})_{hm}=\sum_i\sum_j\sum_k n_{ijk}\,(ABC)_{hijk}\,(ABC)_{mijk}-\sum_i\sum_j\frac{(\sum_k n_{ijk}\,(ABC)_{hijk})\,(\sum_k n_{ijk}\,(ABC)_{mijk})}{\sum_k n_{ijk}}$$

Note that each of these equations gives sums of squares when $h=m$ and sums of products when $h\neq m$,

TABLE 1

Selection of important variables by iterative procedure from variables (1) birth weight, (2) age at first calving, (3) first lactation yield, (4) first lactation length and (5) first calving interval

Iteration					
116 d.f.		*87 d.f.*		*58 d.f.*	
Var	*C*	*Var*	*C*	*Var*	*C*
1	141.7*@	1	—	1	—
2	195.8 @	2	102.5*@	2	—
3	205.6 @	3	106.4 @	3	64.6* NS
4	208.8 @	4	109.6 @	4	64.7 NS
5	211.1 @	5	244.2 @	5	66.4 NS

C indicates criteria explained in Sec. 3.1

* indicates the selected variable

@ means significant

NS means non-significant.

TABLE 2

Pooled within cells correlations

Var	*2*	*3*	*4*	*5*
1	—0.006	0.020	—0.024	—0.002
2		0.704*	—0.087	—0.060
3			0.679*	0.791*
4				0.704*

* Significant at 5% level of significance.

TABLE 3

MANOVA for variables birth weight, age at first calving and lactation length

Source	*d.f.*	Λ	*R*+
Breed groups (A)	2	0 7382744	57.31*
Periods (B)	1	0.9441694	4.06*
Seasons (C)	4	0.9085098	13.13*
A × B	2	0.9851020	2.09
B × C	4	0.9739378	3.29*
A × C	8	0.8896306	10.06**
A × B × C	8	0.9085077	7.99
Within ceils	208		

+Criteria—See Sec. 3.2.

TABLE 4

Mabalanobis' distance-D^2 among breed groups, periods and seasons

Breed groups

	A_1	A_2
A_2	0.001	
A_3	66.920**	67.102**

Periods

	B_1
B_2	4.396*

Seasons

	C_1	C_2	C_3	C_4
C_2	14.539**			
C_3	0.003	15.150**		
C_4	5.290**	35.715**	5.055**	
C_5	8.817	43.806**	7.533**	6.650**

*P∠.005, **P∠0.01