Arunachalam, V. and Murty, B. R.
Computer Programmes for Some Problems in Biometrical Genetics-I. Use of Mahalanobis' D^2 in classificatory problems.
Indian Journal of Genetics and Plant Breeding, 27 (1967): 60-69.

COMPUTER PROGRAMMES FOR SOME PROBLEMS IN BIOMETRICAL GENETICS—I. USE OF MAHALANOBIS' D^2 IN CLASSIFICATORY PROBLEMS

B. R. Murty and V. Arunachalam

Division of Genetics, Indian Agricultural Research Institute, Delhi-12

(Accepted : 16-xii-1966)

During the various types of investigations on population dynamics carried out at the Division of Genetics, Indian Agricultural Research Institute, New Delhi, the assessment of genetic divergence between populations by multivariate analyses was found to be useful for classificatory problems as well as choice of parents for breeding work in crops with diverse breeding systems (Murty and Arunachalam, 1966).

While the initial work was limited to a few populations in each crop, the fundamental information on the factors influencing genetic divergence means scanning a large collection with as many characters related to fitness under natural and human selection included in the analysis. Similarly, the extraction of characteristic roots from characteristic equations involving several variables means several iterations. Since the magnitude of computations increase several-fold with an increase in the size of the matrices and the number of populations, the assessment of a large collection of germ plasm can be undertaken only with the facilities of a computer for accurate and rapid analysis.

During the past three years, programmes have been developed in the Biometrics Unit for problems involving multivariate analysis, canonical analysis, factor analysis, fractional and full diallel analyses for combining ability, serial analysis over different environments for assessment of genotype-environment interaction and similar problems encountered frequently in biological investigations.

Computer programmes which are of special interest, were recently published by some workers (Littlewood *et al.*, 1964, Kobetich, 1964). Since the investigations carried out here are of particular interest to biological workers, in general, it is felt desirable to make these programmes available for such investigations.

These programmes are written in FORTRAN language for an IBM 1620 (Model II) computer and can easily be modified to suit the occasion in fields other than biology. For most of the plant breeding investigations of these types we come across, the programme is general enough to be used without any modification. The first of these programmes is on an assessment of divergence by the use of Mahalanobis' D^2-statistic.

DESCRIPTION OF THE PROGRAMME

This programme is intended to compute D^2 values between all possible combinations for a maximum of 80 varieties and 10 characters. Upto 100 varieties with 10 characters (or any combination such that number of varieties × number of characters is less than 1000), the programme will work without any alteration by changing the card No. 2 as For ×53. For investigations involving more varieties and characters the programme can easily be modified. The methodology used in this programme is completely given by C. R. Rao (1952, 1958).

COMPUTATIONAL METHOD

The calculation of D^2 values involves three major steps :

(i) A set of uncorrelated linear combinations (y's) is obtained by the pivotal condensation of the common dispersion matrix (Rao, 1952) of a set of correlated variables (x's).

(ii) Using the relations between y's and x's, the mean values of different varieties for different characters are transformed into the mean values of a set of uncorrelated linear combinations (Y's).

(iii) The D^2 between the ith and jth variety for k characters is calculated as $D^2_{ij}=\sum_{t=1} (Y_{it}-Y_{jt})^2$. The k component D-squares are calculated separately and added up to give D^2_{ij}.

(iv) The k component D-squares for each combination are ranked in descending order of magnitude, equal values, if occur, which are very rare, receiving same ranks.

(v) The ranks are added up for each component D-square over all combinations and the rank totals are got.

Input Data—The input medium for all these programmes is 80-column punch cards. The following data are required as input data.

(i) Title of the experiment not exceeding 80 letters punched in one card from columns 1–80.

(ii) The number of characters and the number of varieties each occupying column 1 to 3 and 4 to 6 respectively of one punch card.

(iii) The upper half of the common dispersion matrix designated as A-matrix i.e., $((a_{ij}))$, $(j\geq i, i=1, 2, \ldots n)$

$$\text{(e.g.)}\begin{bmatrix} a_{11} & a_{12}\ldots\ldots a_{1n} \\ & a_{22}\ldots\ldots a_{2n} \\ & 3_{nn} \end{bmatrix}$$

Each card contains 7 quantities, each quantity occupying 11 columns with four decimal digits, i.e., quantities 1, 2, 3, etc. occupying columns 1-11, 12–22, 23–33, etc. respectively of a card. The decimal point need not be punched. For e.g., if a 4×4 matrix is used with values

$$\begin{bmatrix} 832\cdot2382 & -0\cdot7632 & -0\cdot0078 & 6\cdot7002 \\ & 11576\cdot2000 & 76\cdot3280 & -6\cdot0007 \\ & & 6766327\cdot1181 & -6650\cdot7634 \\ & & & 181\cdot0705 \end{bmatrix}$$

the input will take the form

bbbb8322382bbbbbb—7632bbbbbb—0078bbbbbb67002bb115762000bbbbb 763280bbbbb—60007

(Card 1, the 7 quantities occupying columns 1–77 of the card, b's representing blanks) 67663271181bb—66507634bbbb1810705 (Card 2, the remaining 3 quantities occupying columns 1–33). The computer will read the quantities as if the decimal point is placed after the 4th digit from the right of each quantity.

The format i.e., the mode of allotting columns to each quantity, may be changed if necessary according to the size of the quantities in individual experiments which requires changing the statement 11 in the programme.

(iv) The mean values for each character for each variety. The values of each of the characters for one variety are punched in one card.

The mean value for each character occupies 4 columns with one decimal digit. The characters are arranged in the same order as in the common dispersion matrix. The mean values for a variety are punched in one card in the case of 10 characters for e.g., from column 1 to 40 as indicated in (iii) above. Thus for each variety one card is used for the input of mean values.

Output.—The programme renders the following printed output in a neat form giving spaces and underlined sub-titles wherever necessary as in the example presented here.

(i) Title of the experiment.

(ii) y, the coefficients of x in the uncorrelated linear combinations.

(iii) Uncorrelated mean values for each variety for each component y_1, y_2, etc.

(iv) D-square values and ranks for each combination. All the component D-squares and the total D-square are printed out.

(v) Rank totals.

In addition, the programme renders the mean values for each of the uncorrelated linear-combinations y_1, y_2 etc. for each variety as punched output in cards. These cards can be used as such for doing the principal component analysis to be published as Part II of the paper. The values y_1, y_2 etc. for one variety is punched in one card, similar to the input (iv), the change being that each quantity occupies 10 columns with four decimal digits.

The programme with the example of the data collected in this laboratory published in Sankhya, Vol. 27 (1965) is presented as an appendix. The input and printed output excluding punched output are given.

ACKNOWLEDGEMENT

Our thanks are due to the officer and members of the Mechanical Tabulation Unit, Institute of Agricultural Research Statistics, New Delhi for their help and for making the computer available for these investigtions.

REFERENCES

Kobetich, E.J. (1964). General Computer plotting subroutine, *Trans. Kansas. Acad. Sci.*, **67:** 640–45.

Littlewood, R.K., Carmer, S.G. and Hittle, C.N. (1964. A computer programme for estimating combining abilities in relation to diallel crossing systems, *Crop Sci.*, **4:** 662–63.

Majumdar, D.N. and Rao C.R. (1958). Bengal anthropometric survey, 1945; A statistical study, *Sankhya*, **19:** 201–408.

Murty, B. R. and Arunachalam, V. (1966). The nature of divergence in relation to breeding system in some crop plants, *Indian J. Genet.*, **26A:** 188–98.

———— Mathur, J.B.L. and Arunachalam, V. (1965). Self-incompatibility and genetic divergence in *Brassica campestris* var. brown *sarson*, *Sankhya* **B, 27:** 271–78.

Rao, C.R. (1952). *Advanced Statistical Methods in Biometrical Research*, John Wiley & Sons, New York.

APPENDIX

```
‡‡JOB 5
‡‡FORX54
  COMPUTATION OF D-SQUARE VALUES-PROGRAMMED BY V. ARUNACHALAM, I.A.R.I.
      DIMENSION A(10, 10), B (10, 10), PA (10), PB (10, 10), X (20), Y (80, 10), DS (20)
     1, M (10), KT (10)
  250 READ 251, (DS (I), I=1, 20)
  251 FORMAT (20A4)
      PRINT 252, (DS (I), I=1, 20)
  252 FORMAT (1X, 20A4/1X, 80 (1H-)//)
C THE ABOVE CAUSES THE PRINTING OF THE TITLE OF THE EXPERIMENT
C N-NO. OF CHARACTERS, IV-NO. OF VARIETIES
C A(N, N)-COMMON DISPERSION MATRIX, B(N, N)-UNIT MATRIX
C Y(IV, N)-UNCORRELATED MEANS MATRIX OF DIMENSION IV X N
C DS (N)-COMPONENT D-SQUARES
      READ 1, N, IV
    1 FORMAT (2I3)
      NO=N-1
      IVV=IV-1
      FIV=IV
      READ 11, ((A(I,J), J=I, N), I=1
   11 FORMAT (7F 11·4)
      PA (1)=A(1, 1)
      DO 101 I=1, N
  101 B (I, I)=1.
      DO 1010 I=1, NO
      IA=I+1
      DO 10 J=IA, N
      A (J, I)=A(I, J)
      B (I, J)=0.
   10 B (J, I)=B(I, J)
 1010 CONTINUE
      DO 1001 J=1, N
 1001 PB (1, J)=B (1, J)
      DO 2 I=1, NO
      L=I+1
      DO 12  J=L, N
   12 A(I, J)=A(I, J)/A(I, I)
      DO 121 J=1, I
  121 B(I, J)=B(I, J)/A(I, I)
      DO 114 K=L, N
```

```
      DO 112 J=L, N
  112 A(K, J)=A(K, J)-A(K, I)* A(I, J)
      DO 113 J=1, I
  113 B(K, J)=B (K, J)-A(K, I)* B(I, J)
  114 CONTINUE
      PA(L)=A(L, L)
      DO 1200 J=1, N
 1200 PB (L, J)=B (L, J)
    2 CONTINUE
      DO 8 I=1, N
    8 PA (I)=SQRTF (PA(I))
      DO 80 I=1, N
      DO 80 J=1, N
   80 PB(I, J)=PB(I, J)/PA(I)
      PRINT 9
    9 FORMAT (1X, 59HCOEFFICIENTS OF X IN THE UNCORRELATED LINEAR FUNCTION
       1S OF Y/1X, 1H-, 19(3H*-*), 1H-//)
      PRINT 90, 1H
   90 FORMAT (1X, 65 (1H-)/29X, 15HCOEFFICIENTS OF// 10X, 4HX(1), 6X, 4HX(2), 6X,
       14HX(3), 6X, 4HX(4), 6X, 4HX(5), 6X, 4HX(6)/1X, 65(1H-)/)
      DO 91 I=1, N
   91 PRINT 92, I, (PB(I, J), J=1, N)
   92 FORMAT (1X, 2HY(, I2, 1H), 6 (2X,F8.4)),
      PRINT 7000
 7000 FORMAT (/)
      PRINT 222
  222 FORMAT (1×24HUNCORRELATED MEAN VALUES/1X, 24 (1H-)//)
      PRINT 224
  224 FORMAT (1X,74(1H-)/1X,3HVAR, 5X, 4HY(1), 8X, 4HY(2), 8X,4HY(3). 8X. 4HY(4)
       1, 8X, 4HY(5), 8X, 4HY(6)1X, 74(1H-)//)
      DO 215 KD=1, IV
      READ 212, (X(J), J=1, N)
  212 FORMAT (10F 4·1)
      DO 214 I=1, N
      Y(KD, I)=0.
      DO 214 J=1, I
  214 Y(KD, I)=Y(KD, I)+PB (I, J)*X(J)
  215 PRINT 225, KD, (Y (KD, I), I=1, N)
  225 FORMAT (2X, I2, 3X, F8·3, 5(4X,F8·3))
C STATEMENTS 224 AND 225 TO BE CHANGED FOR PROCESSING MORE THAN SIX
```

```
C  CHARACTERS UPTO TEN
        DO 9991 I=1, IV
   9991 PUNCH 9990, (Y(I, J), J=1, N), I
   9990 FORMAT (6F10·4,17X,I2)
        PRINT 7000
        PRINT 216
    216 FORMAT (53X,15HD-SQUARE VALUES/53X,15(1H-)//)
        PRINT 217
    217 FORMAT (21X, 5HCOMBN, 4X, 6HD1-SQR, 4X, 6HD2-SQR, 4X, 6HD3-SQR, 4X, 6HD4-SQR
           C, 4X,6HD5-SQR, 4X, 6HD6-SQR, 4X, 8HD-SQUARE/21X, 77 (1H-)//)
        DO 2213 I=1, N
   2213 KT(I)=0
        DO 221 KM=1, IVV
        KK=KM+1
        DO 221 I=KK, IV
        DSS=0.
        DO 223 J=1, N
        DS(J)=Y(I, J)-Y(KM, J)
        DS(J)=DS(J)* DS(J)
    223 DSS=DSS+DS(J)
       PRINT 218, KM, I,(DS(J), J=1, N), DSS
    21  FORMAT (21X, I2,1H-, I2, 3X, 6(F9·4,1X), 2X,F7·2)
        8 DO 2212 LI=1, N
        M(LI)=1
        DO 2211 LJ=1, N
        IF (DS(LI)-DS(LJ)) 2210, 2211, 2211
   2210 M(LI)=M(LI)+1
   2211 CONTINUE
   2212 KT(LI)=KT(LI)+M(LI)
        PRINT 2180, (M(LI), LI=1, N)
   2180 FORMAT (22X,4HRANK, 2X, 6 (5X, I1, 4X)/)
    221 CONTINUE
        PRINT 2215, (KT(I), I=1, N)
   2215 FORMAT (21X, 77(1H-)/16X, 10HRANK TOTAL, 2X,6 (3X, I4, 3X)/21X,77(1H-)//)
C  STATEMENTS 224, 225, 9990, 218, 2180, AND 2215 TO BE CHANGED FOR
   PROCESSING MORE THAN SIX CHARACTERS UPTO TEN
C  FORMAT AND DIMENSION STATEMENTS SHOULD BE ALTERED FOR PROCESSING MORE
C  THAN 10 CHARACTERS AND 80 VARIETIES
        PRINT 2216
```

```
2216 FORMAT (10(/))
     GO TO 250
     END
DATA ON BRASSICA-SANKHYA, B., VOL. 28, 1966
  6 10
      97500    -56600     2400    -46000    -31600     1500   423000
      37800    162400   147500     24500     13400    31600     8900
       8900    340300    98700     30400    490500     5700    43300
   472   466   55   240   394   167
   485   526   73   242   385   192
   579   597   82   275   435   196
   511   426   66   243   342   167
   594   414   76   328   208   151
   599   427   78   309   211   141
   755   639   77   199   365   217
   544   451   77   260   382   188
   724   583   72   188   320   213
   602   501   76   314   383   171
```

DATA ON BRASSICA–SANKHYA, B., VOL. 28, 1966

COEFFICIENTS OF X IN THE UNCORRELATED LINEAR FUNCTIONS OF Y

	COEFFICIENTS OF					
	X(1)	X(2)	X(3)	X(4)	X(5)	X(6)
Y (1)	·3202	0·0000	0·0000	0·0000	0·0000	0·0000
Y (2)	·0929	·1600	0·0000	0·0000	0·0000	0·0000
Y (3)	–·0855	–·1035	1·0312	0·0000	0·0000	0·0000
Y (4)	·0908	–·0298	–·4211	·2073	0·0000	0·0000
Y (5)	·0025	–·0517	·1144	–·0299	·1529	0·0000
Y (6)	–·0079	·0035	–·3044	–·0215	·0022	·5199

UNCORRELATED MEAN VALUES

VAR	Y(1)	Y(2)	Y(3)	Y(4)	Y(5)	Y(6)
1	15·116	11·847	–3·192	5·560	3·649	6·370
2	15·532	12·928	–2·069	4·783	3·405	7·127
3	18·542	14·939	–2·680	5·731	3·831	6·951
4	16·365	11·569	–1·977	5·633	3·188	5·972
5	19·023	12·148	–1·531	7·764	1·082	4·552
6	19·183	12·403	–1·502	7·293	1·141	4·014
7	24·179	17·247	–5·136	5·840	2·759	8·219
8	17·421	12·276	–1·384	5·747	3·754	6·684
9	23·186	16·062	–4·806	5·708	2·328	8·182
10	19·279	13·615	–2·501	7·287	3·352	5·686

D-SQUARE VALUES

COMBN	D1–SQR	D2–SQR	D3–SQR	D4–SQR	D5–SQR	D6–SQR	D–SQUARE
1–2	·1733	1·1694	1·2621	·6041	·0597	·5720	3·84
RANK	5	2	1	3	6	4	
1–3	11·7425	9·5588	·2622	·0291	·0330	·3377	21·96
RANK	1	2	4	6	5	3	
1–4	1·5600	·0772	1·4766	·0052	·2129	·1584	3·49
RANK	1	5	2	6	3	4	
..................etc..................							
8–9	33·2307	14·3354	11·7132	·0015	2·0327	2·2438	63·55
RANK	1	2	3	6	5	4	
8–10	3·4502	1·7943	1·2479	2·3712	·1611	·9956	10·02
RANK	1	3	4	2	6	5	
9–10	15·2656	5·9861	5·3145	2·4954	1·0491	6·2290	36·33
RANK	1	3	4	5	6	2	
RANK TOTAL	87	132	179	198	117	152	