Arunachalam, V.
Computer Programmes for Some Problems in Biometrical Genetics-III. Factor Analysis-centroid Method.
Indian Journal of Genetics and Plant Breeding, 27 (1967): 381-391.

COMPUTER PROGRAMMES FOR SOME PROBLEMS IN BIOMETRICAL GENETICS—III. FACTOR ANALYSIS–CENTROID METHOD

V. ARUNACHALAM

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

(Accepted : 22-ix-1967)

FACTOR analysis as a branch of multivariate analysis is very useful in determining the number and nature of causative influences responsible for the inter-correlation of variables in any population. Essentially, it aims in explaining a $p \times p$ matrix of correlations where p is the number of variables, by means of a fewer number k ($k<p$) of meaningful factors (Holzinger and Harman, 1941; Maxwell, 1961; Lawley and Maxwell, 1963; Rao, 1964). Such an approach, employing centroid method of factor analysis, was applied for the first time in crop plants to explain the causal factors for the genetic diversity found in some natural and selected populations of *Sorghum* (Murty and Arunachalam, 1967).

The present paper gives a computer programme (Appendix 1) prepared at the Biometrics Unit of the Division of Genetics for centroid method of factor analysis along with an example from the published data on *Sorghum*.

Method of analysis.—The method of analysis followed in this programme is completely outlined in Appendix E (pages 353–67) of the book "*Factor Analysis*" by Holzinger and Harman. The factors are obtained by centroid method with the variables reflected in the origin, when necessary, to remove the centroid from the origin in the residual factor space and to increase the contribution of the successive factors as well. The following considerations are taken into account during the centroid method of factoring.

(i) The highest correlation coefficient in each array of the correlation matrix, R, is taken as an estimate of the array communality or the diagonal element, as suggested by Thurstone (1935) and Cattell (1965).

(ii) Each element of the residual matrix, other than the estimated communalities, is tested against a desired accuracy level, designated as AQC in the text of the computer programme, to be chosen by the experimenter to suit his experimental material. If any element is greater than AQC, factoring continues: otherwise further extraction of factors is stopped.

(iii) The number of meaningful factors that can be extracted varies from one experimental material to another. However, this programme follows a general rule postulated by Lawley and Maxwell (1963) that $(p+k)$ should be less than $(p-k)^2$ where p is the number of variables and k, the number of factors. In our work on *Sorghum*, we found that this rule was practicable. The number of factors ultimately extracted will be k or m whichever is less, where k is the

number of factors calculated by the above rule and m is the number of factors extracted such that the elements of the residual matrix eliminating the m factors will be less than the desired level of accuracy, AQC. In fact, the programme stops extracting further factors when the number of factors extracted equals L, the lesser of k and m. In any particular case, if the number of factors extracted is only one, a message to that effect will be printed and the job is terminated.

After the number of factors to be extracted are decided as in (iii), the factor matrix of the size $p \times k$ giving the loadings of each variable on the factors is multiplied with its transpose and the resulting $p \times p$ matrix is printed. If the factoring by centroid method is adequate, then this matrix should be equal to the original correlation matrix after allowing for experimental and other errors.

The factor correlation matrix is then obtained primarily to see whether the factors are essentially uncorrelated.

There are several methods of estimating the communalities and the one explained in this programme is the easiest of them all. To see whether the factors obtained are adequate to explain the inter-correlations of the original variables, the communalities are calculated again using these factors and compared with the original estimated communalities. The contribution of each factor to the total original communality is obtained and the cumulative percentage contribution of all these factors to the total original communality serves again as an indication whether the number of factors extracted is adequate.

The programme can work with the control card ≠≠ FORX54 without major modifications up to 15 variables. It can work with a control card ≠≠FORX53 up to 20 variables with necessary changes in the 'Dimension' statement and other output format statements. It can be easily split into a main and a link programme to accommodate more variables.

Language of the programme.—FORTRAN II suitable to work on an IBM 1620 (Model II) computer.

Input.—The following is required as input data. They should be punched in cards and arranged in the order indicated below :

(i) The title of the experiment and other coded details should be arranged to occupy 60 letters at the maximum and punched in one card starting from column 1.

(ii) The number of variables N and the desired level of accuracy (AQC) to test the elements of residual matrix as explained before should be punched in one card, the former occupying columns 1 and 2 and the latter 3 and 4. AQC usually is a small decimal quantity with two decimal digits, e.g., N=9, AQC=.07 should be punched in one card from column 1 to 4 as 0907. The decimal point should not be punched.

(iii) The correlation matrix designated as R-matrix with the diagonal elements filled up with the estimated original communalities should be fed as follows : Since this matrix is symmetric, only one half of the matrix along with the diagonal elements is required as input. Each array of this triangular matrix is punched in cards, each card containing 10 quantities, each quantity

occupying six columns with 4 decimal digits. If the number of quantities in an array exceeds ten, then the remaining is punched in another card starting from column 1 and so on. It is important that the quantities in each array should begin with a fresh card. The following 5 × 5 R-matrix would be punched, for example, as follows :

$$(R) = \begin{bmatrix} \cdot 1387 & -\cdot 0015 & \cdot 1387 & -\cdot 1830 & \cdot 0508 \\ & \cdot 1234 & -\cdot 1234 & \cdot 1234 & \cdot 0008 \\ & & \cdot 1387 & -\cdot 1138 & \cdot 1001 \\ & & & \cdot 1234 & -\cdot 0763 \\ & & & & \cdot 1001 \end{bmatrix}$$

Card I : 001387 (Cols. 1 to 6)
Card II : b−0015001234 (Cols. 1 to 12)
Card III : 001387b−1234001387 (Cols. 1 to 18)
Card IV : b−1830001234b−1138001234 (Cols. 1 to 24)
Card V : 000508000008001001b−0763001001 (Cols. 1 to 30)

The decimal points should not be punched. ' b ' stands for blank.

Output.—The following is rendered as printed output with underlined subtitles in a neat form.

(i) The title of the experiment as given in the input (i).

(ii) The desired accuracy level, AQC.

(iii) The original correlation matrix R (the lower triangular matrix).

(iv) The centroid factor 1.

(v) The residual matrix after eliminating factor 1, only if the sense switch 1 is put on before the data are fed.

(vi) With reflected variables are indicated as " Variable..... is reflected ".

The results are obtained in the printer in the same sequence from (iv) to (vi) with the corresponding factor number and residual matrices, when required until all the factors as explained earlier are extracted.

(vii) The factor matrix with the loadings of the variables on the factors.

(viii) The factor matrix multiplied by its transpose.

(ix) The lower half of the matrix of correlations between the factors without the diagonal elements.

(x) The original and calculated communalities along with the percentage contributions of the factors to the total original communality.

The output formats of this programme are suitable for ten variables. When the number of variables differ from ten, the Format and Dimension statements need to be verified before executing the programme.

The procedures outlined in this paper are supplemented below with an example taken from the published data on *Sorghum*.

ACKNOWLEDGEMENT

Thanks are due to the staff of the M. T. Unit of the Institute of Agricultural Research Statistics for computer facilities.

REFERENCES

Cattell, R. B. (1965). Factor Analysis. An introduction to essentials I. The purpose and underlying models. *Biometrics*, **21:** 190–215.

———— (1955). ———— II. The role of factor analysis in research. *Biometrics*, **21:** 405–35.

Holzinger, K. J. and Harman, H. H. (1941). *Factor Analysis.* The University of Chicago Press, Chicago, U.S.A.

Lawley, D. N. and Maxwell, A. E. (1963). *Factor Analysis as a Statistical Method.* Butterworths, London.

Maxwell, A. E. (1961). Recent trends in factor analysis. *J. roy. Statis. Soc.*, **A, 124:** 49–59.

Murty, B. R. and ARUNACHALAM, V. (1967). Factor Analysis of diversity in the genus *Sorghum. Indian J. Genet.*, **27:** 123–35.

Rao, C. R. (1964). The use and interpretation of principal component analysis in applied research. *Sankhya*, **26A:** 329–58.

Thurstone, L. L. (1935). *The Vectors of Mind.* University of Chicago Press, Chicago, U.S.A.

———— (1947). *Multiple factor Analysis.* University of Chicago Press, Chicago, U.S.A.

APPENDIX 1

```
≠ ≠ JOB 5                                                                    001
≠ ≠ FORX 54                                                                  002
C   FACTOR ANALYSIS-CENTROID METHOD                                         003
C   PROGRAMMED BY V. ARUNACHALAM, BIOMETRICS UNIT, DIVISION OF GENETICS.     004
C   REFER TO CHAPTER 8, PP. 180-98 OF THE BOOK FACTOR ANALYSIS BY HOLZINGER  005
C   AND HARMAN (1941) FOR THE METHOD DESCRIBED IN THIS PROGRAMME.            006
      DIMENSION R (15, 15), G (15, 15), NEGS (15), S (15), AL  (15,10), CS (15), CK (15)   007
     1, REFL (15), D (15)                                                    008
     1 READ 2, (S (I), I=1, 15)                                              009
     2 FORMAT (15A4)                                                         010
       PRINT 3                                                               011
     3 FORMAT (45 X, 31HFACTOR ANALYSIS-CENTROID METHOD/45X, 31 (1H-) //)    012
       PRINT 4, (S (I), I=1, 15)                                             013
     4 FORMAT (30 X, 15 A4/30X, 60 (1H-)//)                                  014
C    THE ABOVE CAUSES PRINTING OF THE TITLE OF THE EXPERIMENT.               015
       READ 5, N, AQC                                                        016
     5 FORMAT (I2, F2. 2)                                                    017
C    N-NUMBER OF CHARACTERS, AQC-ACCURACY DESIRED (SEE TEXT OF THIS PROGRAMME)  018
       PRINT 72, AQC                                                         019
    72 FORMAT (30X, 39HDESIRED ACCURACY FOR TESTING RESIDUALS=,F4. 2//)      020
       DO 6 J=1, N                                                           021
     6 READ 7, (R (I, J), I=1, J)                                            022
     7 FORMAT (10 F6.4)                                                      023
C THE ABOVE CAUSES READING OF THE UPPER HALF INCLUDING THE DIAGONAL          024
C ELEMENTS OF THE CORRELATION MATRIX R.                                      025
       NO=N-1                                                                026
       NOO=N/2                                                               027
       FN=N                                                                  028
       L=0                                                                   029
       DO 8 I=1, NO                                                          030
       IA=I+1                                                                031
       DO 8 J=IA, N                                                          032
     8 R(J, I)=R (I, J)                                                      033
       DO 80 I=1, N                                                          034
    80 D (I)=R (I, I)                                                        035
       PRINT 800                                                             036
   800 FORMAT (45X, 27 H ORIGINAL CORRELATION MATRIX/45X, 27 (1H-)/)         037
       DO 9 J=1, N                                                           038
     9 PRINT 10, (R (I, J), I=1, J)                                          039
    10 FORMAT (20 F 6.3)                                                     040
C THE ABOVE CAUSES THE PRINTING OF THE CORRELATION MATRIX R.                 041
       K=1                                                                   042
    11 MA=N+K                                                                043
```

```
      MB=(N-K)* (N-K)                                                    044
      IF (MA-MB) 12, 12, 13                                              045
   12 K=K+1                                                              046
      GO TO 11                                                           047
   13 K=K-1                                                              048
   14 DO 15 J=1, N                                                       049
      DO 15 I=1, N                                                       050
   15 G(I,J)=R(I, J)                                                     051
      DO 16 I=1, N                                                       052
   16 REFL (I)=1.                                                        053
   17 DO 20 J=1, N                                                       054
      NEGS (J)=0                                                         055
      DO 20 I=1, N                                                       056
      IF (J-I) 18, 20, 18                                                057
   18 IF (R (I, J)) 19, 20, 20                                           058
   19 NEGS (J)=NEGS (J)+1                                                059
   20 CONTINUE                                                           060
      DO 22 J=1, N                                                       061
      IF (NEGS (J)-NOO) 22, 22, 21                                       062
   21 GO TO 23                                                           063
   22 CONTINUE                                                           064
      GO TO 28                                                           065
   23 MAX=NEGS (1)                                                       066
      INDEX=1                                                            067
      DO 25 J=2, N                                                       068
      IF (MAX-NEGS (J)) 24, 25, 25                                       069
   24 MAX=NEGS (J)                                                       070
      INDEX=J                                                            071
   25 CONTINUE                                                           072
      PRINT 26, INDEX                                                    073
   26 FORMAT (/8X, 9HVARIABLE   , I2, 14H IS REFLECTED.)                 074
      REFL (INDEX)=-REFL (INDEX)                                         075
      DO 27 I=1, N                                                       076
      R (I, INDEX)=REFL (INDEX)* R(I, INDEX)                             077
   27 R(INDEX, I)=REFL (INDEX)* R (INDEX, I)                             078
      GO TO 17                                                           079
   28 T=0.                                                               080
      DO 29 I=1, N                                                       081
      DO 29 J=1, N                                                       082
   29 R(I, J)=REFL (I)* REFL (J)* G (I,J)                                083
      DO 31 J=1, N                                                       084
      S(J)=0.                                                            085
      DO 30 I=1, N                                                       086
   30 S(J)=S(J)+R(I, J)                                                  087
```

```
   31 T=T+S(J)                                                                  088
      IF (T) 503, 503, 310                                                      089
  310 T=SQRTF (T)                                                               090
      L=L+1                                                                     091
      DO 32 J=1, N                                                              092
      S(J)=REFL (J)* S (J)                                                      093
   32 AL (J, L)=S(J)/T                                                          094
      PRINT 33, L                                                               095
   33 FORMAT (//50X, 16HCENTROID FACTOR (, I2, 1H) /50X,19 (1H-)/)              096
      PRINT 34, (AL (J,L), J=1, N)                                              097
   34 FORMAT (15F8.3)                                                           098
   35 DO 36 J=1, N                                                              099
      DO 36 I=J,N                                                               100
   36 R(I,J)=G(I,J)-(AL(I,L)* AL(J,L))                                          101
      DO 38 I=1,NO                                                              102
      IA=I+1                                                                    103
      DO 38 J=IA, N                                                             104
   38 R(I,J)=R(J, I)                                                            105
  360 IF (SENSE SWITCH 1) 370, 40                                               106
  370 PRINT 37, L                                                               107
   37 FORMAT (/41X, 35 HRESIDUAL MATRIX ELIMINATING FACTOR (I2, 1H)/41X, 381    108
      1H-)//)                                                                   109
      DO 39 J=1,N                                                               110
   39 PRINT 10, (R (I, J), I=1, J)                                              111
C  SENSE SWITCH 1 SHOULD BE PUT ON TO GET THE  RESIDUAL R MATRIX  PRINTED.      112
      IF (T) 43, 43, 40                                                         113
   40 DO 41 J=1, NO                                                             114
      IA=J+1                                                                    115
      DO 41 I=IA,N                                                              116
      TEST=ABSF (R(I, J))                                                       117
      IF (TEST-AQC) 41, 41, 42                                                  118
   41 CONTINUE                                                                  119
      GO TO 43                                                                  120
   42 IF (L-K) 14, 43, 43                                                       121
   43 PRINT 44                                                                  122
   44 FORMAT (/42X, 13HFACTOR MATRIX/42X, 13 (1H-)//26X, 71 (1H-)/26X, 8 H VARIAB 123
      1LE, 17X,26HCOMMON FACTOR COEFFICIENTS/37X, 60 (1H-))                     124
      PRINT 440, (I, I=1, L)                                                    125
  440 FORMAT (37X, 6 (5X, I1, 4X))                                              126
      PRINT 441                                                                 127
  441 FORMAT (26X, 71 (1H-)//)                                                  128
      DO 45 J=1, N                                                              129
   45 PRINT 46, J, (AL (J, I), I=1, L)                                          130
   46 FORMAT (29X, I2, 6X, 6 (F7.3, 3X))                                        131
      PRINT 47                                                                  132
```

```
      47 FORMAT (26X, 71 (1H-)//)
         DO 48 J=1, N
         DO 48 IA=J, N
         R(J, IA)=0.
         DO 48 I=1, L
      48 R(J, IA)=R(J, IA)+AL(J, I)* AL (IA, I)
         PRINT 49
      49 FORMAT (45X, 29HFACTOR MATRIX X ITS TRANSPOSE/45X, 29 (1H-)/)
         DO 50 J=1, N
      50 PRINT 10, (R (I,J), 1=1, J)
     503 IF (L-1) 501, 501, 500
     501 PRINT 502
     502 FORMAT (39X, 43HONLY ONE DISTINCT FACTOR CAN BE ELIMINATED./39X, 19HE
         1 ND OF THE PROGRAM./)
         GO TO 71
     500 DO 52 I=1, L
         CS (I)=0.
         SS=0.
         DO 51 J=1, N
         CS (I)=CS (I)+AL (J, I)
      51 SS=SS+AL (J, I)* AL (J, I)
         CF=CS (I)* CS (I)/FN
         SS=SS-CF
         SS=SS/(FN-1.)
      52 R(I, I)=SS
         LL=L-1
         DO 54 I=1, LL
         IA=I+1
         DO 54 J=IA,L
         SS=0.
         DO 53 KK=1, N
      53 SS=SS+AL (KK, I)* AL (KK, J)
         CF=CS(I)* CS (J)/FN
         SS=SS-CF
         SS=SS/(FN-1.)
      54 R(I, J)=SS
         DO 55 I=1, LL
         IA=I+1
         DO 55 J=IA, L
         CF=SQRTF (R (I,I)* R(J, J))
      55 R(I, J)=R(I, J)/CF
         PRINT 56
      56 FORMAT (/29X, 62HFACTOR CORRELATION MATRIX-LOWER HALF WITHOUT DIAGON
       I AL ELEMENTS/29X, 62 (1H-)//)
```

```
      DO 57 J=2, L                                                                    177
      JK=J-1                                                                          178
   57 PRINT 58, (R (I, J), I=1, JK)                                                   179
   58 FORMAT (6 (F 7.4, 3X))                                                          180
      ORG=0.                                                                          181
      CAL=0.                                                                          182
      DO 60 I=1,N                                                                     183
      CK (I)=0.                                                                       184
      DO 59 J=1,L                                                                     185
   59 CK (I)=CK (I)+AL (I,J)* AL (I,J)                                                186
      CAL=CAL+CK(I)                                                                   187
   60 ORG=ORG+D (I)                                                                   188
      PRINT 61, (I, I=1, N)                                                           189
   61 FORMAT (/42X, 37HORIGINAL AND CALCULATED COMMUNALITIES/42X, 37 (1H-)//          190
     1 20X, 10 (5X, I2), 3X, 5HTOTAL/)                                               191
      PRINT 62, (D (I), I=1, N), ORG                                                  192
   62 FORMAT (10X, 10 HORIGINAL, 2X, 10 (F6.3, 1X), F7.3/)                            193
      PRINT 63, (CK (I), I=1, N), CAL                                                 194
   63 FORMAT (10X, 10 H CALCULATED, 2X, 10 (F 6.3, 1X), F7.3)                         195
      DO 65 J=1, L                                                                    196
      CK (J)=0.                                                                       197
      CS (J)=0.                                                                       198
      DO 64 I=1, N                                                                    199
   64 CK (J)=CK (J)+AL(I,J)* AL (I,J)                                                 200
   65 CS (J)=CK(J)* 100./ORG                                                          201
      PRINT 66, (I, I=1, L)                                                           202
   66 FORMAT (//23X, 10 (5X, I2))                                                     203
      PRINT 67, (CK (J), J=1, L)                                                      204
   67 FORMAT (//22X, 1HC, 2X, 10 (F6.1, 1X))                                          205
      PRINT 68, (CS (J), J=1, L)                                                      206
   68 FORMAT (//22X, 1HQ, 2X, 10 (F6.1, 1X))                                          207
      PRINT 69                                                                        208
   69 FORMAT (//18X, 85HC—CONTRIBUTION OF FACTOR, Q—C EXPRESSED AS PERCE              209
     1 NTAGE OF TOTAL ORIGINAL COMMUNALITY)                                           210
      PRINT 70                                                                        211
   70 FORMAT (10 (/))                                                                 212
   71 GO TO 1                                                                         213
C  N.B. ALL FORMAT STATEMENTS NEED BE CHANGED, WHEN THE NUMBER OF                    214
C  VARIABLES ARE MORE THAN 15, TO BE IN CONFORMITY WITH THE SIZE OF THE              215
CORRELATION MATRIX AND THE NUMBER OF FACTORS. DIMENSION STATEMENT IS TO             216
C  BE CHANGED AND THE PROGRAMME TO BE SPLIT INTO PORTIONS AND ONE OR MORE           217
C  OF THEM TO BE TREATED AS LINK PROGRAMMES, FOR MORE THAN 15 VARIABLES.            218
      END                                                                           219
```

INPUT DATA

DATA ON SORGHUM—ENVIRON. CORRELATION MATRIX-IND. J. GENET., 1967

1039

002812

-4261003357

002129002763007580

001401 -0284003326007878

00281200335700758000424000758O

001434 -0217001407 -0226001798002282

000688000914004667002987005846000912005846

001436000072004900007878005309001815004548007878

-191001156002848 -1780001584001042002219 -1270002848

002735 -338600331200217700306200228200565200379500127200565 2

OUTPUT

FACTOR ANALYSIS-CENTROID METHOD

DATA ON SORGHUM-ENVIRON. CORRELATION MATRIX-IND. J. GENET., 1967

DESIRED ACCURACY FOR TESTING RESIDUALS= .39

ORIGINAL CORRELATION MATRIX

.281									
-.426	.335								
.212	.276	.758							
.140	-.028	.332	.787						
.281	.335	.758	.424	.758					
.143	-.021	.140	-.022	.179	.228				
.068	.091	.466	.298	.584	.091	.584			
.143	.007	.490	.787	.530	.181	.454	.787		
-.019	.115	.284	-.178	.158	.104	.221	-.127	.284	
.273	-.338	.331	.217	.306	.228	.565	.379	.127	.565

CENTROID FACTOR (1)

.222	.070	.818	.557	.871	.253	.692	.734	.196	.536

VARIABLE 8 IS REFLECTED.

VARIABLE 4 IS REFLECTED.

CENTROID FACTOR (2)

.002	.091	.192	-.702	.140	.137	.115	-.515	.458	.078

VARIABLE 6 IS REFLECTED.

VARIABLE 10 IS REFLECTED.

VARIABLE 1 IS REFLECTED.

CENTROID FACTOR (3)

-.415	.771	.162	.065	.127	-.275	0.000	-.065	.076	-.446

FACTOR MATRIX

VARIABLE	COMMON FACTOR COEFFICIENTS		
	1	2	3
1	.222	.002	−.415
2	.070	.091	.771
3	.818	.192	.162
4	.557	−.702	.065
5	.871	.140	.127
6	.253	.137	−.275
7	.692	.115	0.000
8	.734	−.515	−.065
9	.196	.458	.076
10	.536	.078	−.446

FACTOR MATRIX × ITS TRANSPOSE

.222									
−.304	.608								
.114	.200	.732							
.094	.025	.331	.808						
.140	.172	.760	.395	.796					
.171	−.182	.188	.026	.204	.159				
.153	.059	.588	.304	.619	.191	.492			
.189	−.046	.491	.766	.559	.132	.448	.808		
.013	.114	.261	−.207	.245	.091	.189	−.097	.254	
.304	−.299	.381	.214	.421	.269	.380	.382	.107	.492

FACTOR CORRELATION MATRIX—LOWER HALF WITHOUT DIAGONAL ELEMENTS

−.2628	
−.1097	.0631

ORIGINAL AND CALCULATED COMMUNALITIES

	1	2	3	4	5	6	7	8	9	10	TOTAL
ORIGINAL	.281	.335	.758	.787	.758	.228	.584	.787	.284	.565	5.371
CALCULATED	.222	.608	.732	.808	.796	.159	.492	.808	.254	.492	5.375

	1	2	3
C	3.2	1.0	1.0
Q	59.6	19.9	20.4

C—CONTRIBUTION OF FACTOR, Q—C EXPRESSED AS PERCENTAGE OF TOTAL ORIGINAL COMMUNALITY