MORPHO FORTRAN COMPUTER PROGRAMS FOR THE COLLECTION AND MULTIVARIATE ANALYSIS OF MORPHOMETRIC DISTANCE CHARACTERS

by
Gary A. Winans

Coastal Zone and Estuarine Studies Division
Northwest and Alaska Fisheries Center
National Marine Fisheries Service
National Oceanic and Atmospheric Administration
2725 Montlake Boulevard East
Seattle, Washington 98112

April 1984

NWAFC/MORPHO

MORPHO is a series of NMFS computer programs designed for the collection and multivariate analysis of morphometric distance characters. MORPHO consists of three FORTRAN programs:

1) DIGITIZE - collects X-Y coordinate data from a remote computer device, e.g., a digitizing pad, for relevant morphological features or landmarks around the outline of a specimen.

2) MMCALC - calculates Euclidean (=morphometric) distances between any pair of landmarks in any geometrical pattern. Output is n untransformed distance measurements.

3) PCA/SHEAR - computes principal components from the overall variance-covariance matrix for continuous and discrete data and shears components II-IV to produce size-free shape components [after Humphries et al. 1981. Syst. Zool.(3)]. Output includes component scores on I-IV and sheared scores on II-IV. Results are viewed on B.M.D. plots and histograms.

DIGITIZE

This program enters X-Y coordinate data from a remote digitizing pad to a Burroughs data file. Landmarks are predesignated and digitized in the same order on each specimen. The program is set up to digitize one group or population of specimens at a time. This program operates with a DigiPad 5 by GTCO Company and with a Lear Siegler ADM5 terminal. Two changes must be made in the program to accomodate other pads or terminals (see program).

A digitizing session is started at a terminal with the command "RUN DIGITIZE". The user is first prompted on the terminal screen to enter the output file name and population-specific information. For each specimen then, the user enters specimen-specific information, such as weight and general comments. If specimens are digitized from a photograph, a reference scale must be included in the photograph and digitized with the specimen. Missing landmarks are recorded by pressing and holding down the "9" key (i.e., MISS = 999.99). The prompt "Are data OK?" provides the user with the opportunity to repeat a particular interactive procedure. Depressing the stylus on the pad is the equivalent to a "yes" response.

Variables that are entered interactively are:

file name	8A6
group code	A1
population name	A4
population number	I2
specimen (fish) number	I3
number of landmarks to be digitized on each specimen	I2

state of the specimen		A4
FRSH = fresh		
FROZ = frozen		
PICK = chemically preserved		
FOTO = photograph		
eye diameter	entered as	I2 but
	changed to	F2.2
weight		F6.2
comments (color, smell, etc.)		A6
sex (optional)		A1

Variables in output file:

Group code, population name, population number, specimen number, state of the specimen, scale (if SCALE = FOTO, then SCALE ≠1), sex, weight, comment, eye diameter, X-Y coordinate pairs.

```
FORMAT  (A1,2X,A4,2X,I3,2X,I3,2X,A4,F10.4,2X,A1,F10.3,
         2X,A6,F7.3,/,1X,16F7.3,/,1X,16F7.3,/,...)
```

MMCALC

This program calculates the Euclidean distance between pairs of X-Y coordinate points. This program is run by a WFL file:

```
START/MMCALC : JOB

?BEGIN JOB MMCALC;
RUN OBJECT/MMCALC;
FILE 1 = INPUT/DIGITIZED/DATA;
FILE 2 = INPUT/RUN/INFORMATION;
FILE 3 = OUTPUT/MM DISTANCES;
?END JOB;
```

The user provides file names for two input files and one output file.

FILE 1 is the digitized data from DIGITIZE (same variables in the same format);

FILE 2 is:

Line 1 -	Number of specimens	I3
	Number of landmarks digitized per specimen	I3
	Number of distance variables to be calculated (NMAX)	I3
Line 2 -	Landmark pair #1*	2I2
Line 3 -	Landmark pair #2	2I2
:		
Line NMAX +1		

*e.g., for an entry of 0102, the distance between landmark 1 and landmark 2 is calculated;

FILE 3 is the output file with the following variables:

group, population name, population number, sex, specimen number, state, scale, weight, comment, eye diameter, and NMAX distance variables (A1,2X,A4,2X,I3,2X,A1,2X,I3,2X,A4,F10.4,F10.3,2X,A6,F7.3,/, (NMAX)*F8.4). Missing distance values = 999.99.

PCA/SHEAR: FORTRAN

This program calculates principal components (PC) from a variance-covariance matrix. Optionally, size-free "shape" components are calculated from the data, i.e., the PC scores are adjusted or sheared after the algorithm of Humphries et al. 1981.

Given NVAR (number of variables) morphometric measures on NFISH specimens or cases, the main program uses the total variance-covariance matrix to construct six PC factors that represent multivariate linear relationships among the observed variables. It is assumed that the data set is subdivided and sorted by group. [Refer to Morrison (1976) and Green (1976) for details of principal component analysis.]

OUTPUT - to printer:

1) variance-covariance matrix, in part.
2) means and standard deviations for each variable by group.
3) first six eigenvectors and eigenvalues.
4) component correlation matrix, i.e., correlation of ith variable with j^{th} component.
5) component scores on first four eigenvectors for each case.
6) optionally, calculates the condition factor index--(100 x weight)/$length^3$.

SHEAR Subroutine

Although most of the overall size variance may be accounted for in PC I, subsequent principal components may be partially correlated with size (i.e., the first PC). An optional SHEAR analysis may be executed on PC II, III, and IV. Presumably, the shear analysis eliminates size (=PC I)

influences on each of these components producing size-free shape components. In this analysis, variables are adjusted by their mean value per group, e.g., the average head length HL in Group 1 is subtracted from each HL observation in Group 1, the average HL value in Group 2 is subtracted from each HL observation in Group 2, and so on, to produce mean-corrected observations (see Green 1976, p.15). The first component in a PC analysis of the variance-covariance matrix of the mean-corrected data represents a within-group size factor S and is a better estimate of within-group size than, say, fork length (Humphries et al. 1981). The factor S is then used in a series of calculations to complete the SHEAR analysis.

OUTPUT from SHEAR Subroutine - to printer

1. eigenstructure of mean-adjusted data set.
2. component scores on first four eigenvectors for each case.
   (N.B. PC I = within-group size factor S.)
3. mean and standard deviation values by group for initial PC scores.
4. correlations of size S with mean-adjusted PC II, III, and IV.
5. sheared component scores for PC II - IV and their associated sheared eigenvectors.

OUTPUT from PCA/SHEAR (when SANAL=Y) - to disk

1. case identifying information (e.g., POPNO,FISH NO.).
2. miscellaneous data - WEIGHT, EYE DIAMETER, etc.
3. scores on PC I - IV.
4. scores on sheared PC II - IV.

Procedure for a PCA/SHEAR run:

1. Create FILE 1 containing run information, e.g., PCA/START/__.
2. Check FILE 2, the raw data file: is it sorted by a group criterion?
3. Select a name for FILE 3 in which results will be stored, e.g., PCA/RESULTS/__.
4. Insert files names 1, 2, and 3 into respective locations in the WFL file EXAMINE/PCA/RESULTS file.
5. Make other necessary changes in the EXAMINE/RESULTS file, e.g., number of cases, group names, and so on. Remember that output FILE 3 from PCA/SHEAR is the input FILE 3 for EXAMINE/RESULTS.
6. Execute the entire procedure by "START EXAMINE/RESULTS".

Programming Notes:

1. The array dimensions are now set for 20 groups, 40 variables, and 800 cases.
2. Dimension for VCV, calculated as [NVAR*(NVAR+1)]/2, is set at 820 (for 40 variables).
3. Missing values are reset to -0.0 for the Burroughs B6800.
4. The user can enter any set of distance variables in a total data set by means of the format statement.
5. Weight (WT) and fork length values are needed to calculate condition factor index. Therefore, the READ statement must be changed to include the variable WT. Assuming fork length values are in the raw data file, the user must identify the position of this variable in the data, e.g., if fork length is the tenth distance variable for each specimen, NFL=010 (see below).

FILE 1

First card:

col 1-4	I4	No. of cases	NFISH
col 5-6	I2	Total no. of variables	NVAR
col 7-8	I2	No. of groups	NGRP

Next "NGRP" cards:

col 1-3	I3	No. of cases in Group 1	NINGRP(1)
col 4-7	A4	Group 1 name	GRPNAM(1)

Next card:

col 1-3	I3	No. of cases in Group 2	NINGRP(2)
col 4-7	A4	Group 2 name	GRPNAM(2)

.

.

.

Card NGRP + 2:

col 1	A1	SHEAR analysis of PC II-IV? Enter "Y" or "N"	SANAL

Next card:

col 1	A1	Calculate condition index? Enter "Y" or "N".	NEWVAR
col 2-4	I3	Enter variable subscript of fork length for NEWVAR calculation	NFL

Next card:

col 1-6	F6.2	missing data code	MISS

Next card:	col 1	I1	transformation code for input data where 0 = none	ITRAN
			1 = IxI	
			2 = x	
			3 = (x + 1/2)	
			4 = x^2	
			5 + $\log_{10}x$	
			6 = $\log_{10}(x + 1)$	
			7 = arcsin (x) expressed in degrees, where x is between 0-1	
			8 = arcsine (x/100) expressed in degrees, where x is a % between 0 and 100	
			9 = 1/x	
Last card:	col 1-80	(13A6,A2)	format of input data in File 2; enclosed in parentheses	FMT

FILE 2 - data file: three identifying variables plus the raw data; sorted by group. For each specimen, give values for three identifying variables, such as drainage basin number, river system number, and specimen number, followed by the raw data.

OUTPUT FILE

FILE 3 - output file with specimen identifying information, several metric measurements, PC scores, and optionally, CF values, and sheared PC values.

The WFL file EXAMINE/PCA/RESULTS executes PCA/SHEAR and two BMD routines. These subroutines produce scattergrams and histograms of PC and SPC values for specimens analyzed in PCA/SHEAR. In order to make additional plots and histograms, lines 5100-5400 may be deleted once FILE 3 (PCA/RESULTS) is created. In the example below the data were categorized and presorted into two groups (as coded in the second variable GRPNO).

EXAMINE/PCA/RESULTS/EXAMPLE (03/12/84)

```
 100    ?BEGIN JOB EXAMINE;
 200    CLASS=20;                                                  Subroutine
 300    SUBROUTINE PLOT;                                          --Scatterplot
 400    BEGIN
 500    ?RUN BMD/P6D;
 600     FILE FILE3(TITLE=PCA/RESULTS/GARSPR,
 650          FILETYPE=7,KIND=DISK);
 700     FILE FILE5(KIND=READER,TITLE=IOCR,FILETYPE=7);
 800     FILE FILE6(KIND=PRINTER,TITLE=IOLP,MAXRECSIZE=22);
 900    ?DATA IOCR;
1000    /BMDP6D
1100    /PROBLEM TITLE IS'SHEAR ANALYSIS RESULTS--PLOTS'.
1200    /INPUT UNIT = 3.
1250            VARIABLE = 11.
1300            CASES    = 105.                                   --No. of specimens
1400            FORMAT   =*(F1.0,F2.0,F5.0,5F10.5/3F10.5)*.
1500    /VARIABLE NAMES ARE ID1, GRPNO, FSNO,                     --3 ID variables
1600                        S,PC1,PC2,PC3,PC4,SPC2,SPC3,SPC4.
1700         MISSING = 11*-0.0.
1800         LABELS ARE FSNO.                  --Variable #2 is the grouping variable, with
1900         GROUPING IS GRPNO.                  values of 15 and 16 and names
1910    /GROUP CODES(2)=15,16.                   GROUP 1 and GROUP 2
1920            NAMES(2)=GROUP1, GROUP2.
2200    /PLOT YVAR=PC2,SPC2.                      --These are the pair-wise plots
2300          XVAR=PC1,PC1.                       --
2400              STATISTICS. SIZE=100,50.
2500                GROUPS=1,2.                   --Data from groups 1 and 2 are plotted
2600    /END
2700    ?END;
2800    SUBROUTINE HISTO;                                          Subroutine
2900    BEGIN                                                     --Histogram
3000    ?RUN BMD/P7D;
3100     FILE FILE3(TITLE=PCA/RESULTS/GARSPR,
3110          KIND=DISK,FILETYPE=7);
3200     FILE FILE5(KIND=READER,TITLE=IOCR,FILETYPE=7);
3300     FILE FILE6(KIND=PRINTER,TITLE=IOLP,MAXRECSIZE=22);
3400    ?DATA IOCR;
3500    /BMDP7D
3600    /PROBLEM TITLE ='HISTOGRAMS OF SHEARED PCS'.
3700    /INPUT UNIT = 3.
3800            VARIABLES=11.
3810            CASES=105.                                        --No. of specimens
3900            FORMAT=*(F1.0,F2.0,F5.0,5F10.5/3F10.5)*.
4000    /VARIABLE NAMES ARE ID1, GRPNO, FSNO,                     --3 ID variables
4100                        S,PC1, PC2, PC3, PC4, SPC2, SPC3, SPC4.
4200         LABELS ARE FSNO.
4300         MISSING = 11*-0.0.                --Data are grouped by the second ID
4400         GROUPING IS GRPNO.                  variable GRPNO; a specimen is from
4500    /GROUP CODES(2) = 15,16.                 either group 15 or 16
4600            NAMES(2) = GROUP1, GROUP2.
4700    /HISTOGRAM VARIABLE = PC2, SPC2.
4800                GROUPING = GRPNO.
4900    /END
5000    ?END;                                     Files
5100    RUN OBJECT/PCA/SHEAR;
5200    FILE 1=PCA/START/GARSPR;                  --Start-up information
5300    FILE 2=CONVERT/DATA/GARSPR/INFART;        --Raw input data file
5400    FILE 3=PCA/RESULTS/GARSPR;                --PCA results (see lines 600 and 3100)
5500    PLOT;
5600    HISTO;
```

REFERENCES CITED

Dixon, W. J. and M. B. Brown. 1977. BMDP-77: Biomedical computer programs P-series. Univ. of Calif. Press, Berkeley, Calif. 880 p.

Green, P. E. 1976. (with contributions by J. D. Carroll) Mathematical tools for applied multivariate analysis. Academic Press, New York. 376 p.

Humphries, J. M., F. L. Bookstein, B. Chernoff, G. R. Smith, R. L. Elder, and S. G. Poss. 1981. Multivariate discrimination by shape in relation to size. Syst. Zool. 30: 291-308.

IMSL Library Reference Manual. 1977. Ed. 7. Houston, Tex.

Morrison, D. F. 1976. Multivariate statistical methods. McGraw-Hill, New York, New York. 415 p.