COENOSES 3 (2): 107-109, 1988
© 1988 by Edizioni LINT Trieste

INTRODUCTION TO DATA ANALYSIS: A COMPREHENSIVE PROGRAM PACKAGE FOR PERSONAL COMPUTERS

N.C. Kenkel, Department of Botany, The University of Manitoba, Winnipeg, Manitoba R3T 2N2 Canada

Keywords: Data analysis, Programs, BASIC, Community, Population, Ecology

Abstract: A comprehensive program package written in BASIC and available for use on Apple //, Apple Macintosh, and IBM PC personal computers is described. The programs were written for the analysis of data from population and community ecology studies, but should be of use in a number of fields where exploratory multivariate data analysis plays an important role.

Introduction

Personal computers have a number of potential advantages over mainframe configurations: computational costs are minimal or nonexistent, programs are easily transferred between institutions, programming is generally more flexible, high-resolution graphics routines are readily accessible, and the system is more 'user-friendly'. The limited amount of available memory, which until recently has been a major limitation of personal computer systems, is no longer a problem given the new generation of machines such as Apple's Macintosh plus.

While personal computers are now widely available in teaching and workplace environments, corresponding software for specialized applications is often lacking. The monograph and associated program package 'Introduction to Data Analysis' by Orlóci, Kenkel and Orlóci (1988) was written to aid in the teaching of data analysis methods to students of population and community ecology. The programs should be useful both for educational (university undergraduate level courses in introductory data analysis and multivariate analysis in community studies) and research purposes. The package includes programs for univariate statistical computations, probability integral calculation for common statistical distributions, fitting discrete and continuous distributions, the analysis of variance, regression analysis, distance and similarity calculations, univariate and multivariate hypothesis testing, character weighting, cluster analysis, ordination, identification, canonical analysis, and nested hierarchical character structure analysis. All programs are interactive and emphasize flexibility and ease of use. A modular approach is taken, so that more than one program may have to be run to complete a specific task. The package also includes a number of programs for text file creation and manipulation.

Description of programs

The following is a brief overview of programs in the package. A few options or programs not discussed in Orlóci, Kenkel and Orlóci (1988) are also outlined. Unless otherwise stated, those programs which analyze matrices assume that the p rows are variables and the n columns individuals. The three exceptions are programs TESTCOV/2, EMVO and GENDIST (see discussion below). Some of the programs do not require an input text file. These include: EXPONENTIAL, RANDOM, ENTROPY/DIST, PROBS, CHIPROBS, TPROBS, FPROBS, and COMBINATORIAL. The following programs require input which is not in 'true' matrix form: DISTRIBUTIONS, NORMAL, ANOVA, and FANOVA.

CORRELATION: Cross-products (correlation, covariance, product moment, correlation dual, and covariance dual) are calculated between variables.

METRICS: Pairwise metric distances (Manhattan metric, Euclidean distance, and their normalized forms) are calculated between individuals.

EXPONENTIAL: This program simulates exponential population growth.

ENTROPY: For each of p variables, Shannon and Simpson entropy values are calculated. For each pairwise combination of variables, joint and mutual information, Rajski's metric, and the coherence coefficient are computed.

MOMENTS: This calculates, for each variable of a matrix, the mean, third and fourth moments. In addition, the product-moment, covariance, and correlation matrices between variables are calculated.

DISTRIBUTIONS: The Poisson or Bernoulli (Binomial) distribution is fit to discrete (integer) data. Input file represents s frequency values for each category (0, 1, 2, ..., s-1).

NORMAL: The normal distribution is fit to frequency data in s categories (integers). The mean, second, third, and fourth moments, skewness and kurtosis values, and goodness of fit are calculated.

RANDOM: This generates a set of random numbers within a range specified by the user, or a distribution of random values.

ESTIMATE H: This program calculates an unbiased estimate of population entropy based on sample data using the method of Pielou (1966).

ENTROPY/DIST: Random (Monte Carlo) simulation is used to generate a distribution of mutual information values, given j states of the first variable and k of the second.

PROBS: Calculates the alpha-probability values for a given Z variate (standard normal deviate), or Z for a given α-value.

CHIPROBS: Calculates the alpha-probability value for a given chi-squared variate at specified degrees of freedom, or χ^2 for a given α-value.

TPROBS: Calculates the alpha-probability value for a given Student's t variate at specified degrees of freedom, or t for a given α-value.

FPROBS: Calculates the alpha-probability value for a given F variate at specified numerator and denominator degrees of freedom, or F for a given α-value.

PINDEX: A simplified version of the Goodall (1966) probability index is calculated between individuals.

CALHOUN: This calculates the Calhoun set theoretical distance (Bartels *et al.* 1970) between individuals.

LIMITS: Tests the hypothesis of equality of a sample mean vector and a standard (population) mean vector.

TESTCOV/1: Tests the hypothesis of equality of a sample covariance matrix and a standard (population) covariance matrix.

ANOVA: Performs univariate, single factor analysis of variance based on one of three designs: (a) complete randomized, (b) randomized block, and (c) latin square. The input data set for complete randomized assumes that replicates are ordered by treatment affinity. Block designs require that values be ordered as treatments x blocks. For latin square, two files are required: (a) rows as treatments, columns as column affinities, (b) row affinities for each treatment. For all options, an ANOVA table is computed and multiple comparisons are performed using Scheffe's method.

FANOVA: Factorial analysis of variance (up to 4 factors) is performed, assuming an equal number of replicates/treatment. Two designs are available: (a) complete randomized, and (b) randomized block. Input matrix is treatments x replicates. Treatments represent all combinations of all factors at all levels, in lexicographical order and with columns as replicates. For block designs, the ordering of the replicates must reflect the blocking (as in program ANOVA). Note: all main effects and higher order interactions are calculated and tested. The significance of main effects must be interpreted with great caution if higher order interactions are present. Scheffé multiple comparisons (all pairwise, for main effects only) are also calculated. Again, these must be interpreted with great caution in the presence of significant interaction.

TESTCOV/2: This tests for the equality of two or more sample covariance matrices. Input raw data matrix is ordered as rows = individuals, columns = variables. Rows must be ordered according to sample (group) affinity.

EMVO: Tests for the equality of two or more sample mean vectors. The input data is as in program TESTCOV/2. Three options are available: (a) comparison of mean vectors (multivariate analysis of variance). All pairwise comparisons for each variable are also calculated, (b) Canonical groups analysis: analysis as above, but in addition performs a multiple discriminant (canonical) analysis, (c) Profile analysis: this variant of MANOVA is described by Morrison (1976; pages 153-160, 205-216).

CANCOR: Performs canonical correlation analysis. Two input matrices are required: p x n, where p is the number of variables in the first set, and q x n, where q is the number of variables in a second set, measured on the same set of individuals. Values calculated include canonical correlations and a test of their significance, variable weights, canonical scores, structure correlations, and redundancies for each set.

REGRESSION: Performs regression analysis. The available options are: (a) single independent variable (simple linear regression), (b) multiple independent variables (multiple linear regression), and (c) polynomial regression. For all options, an ANOVA table, regression coefficients, fitted Y-values and residuals, confidence limits for the regression coefficients, and Y-values and fitted Y-values, are computed.

WEIGHTING/SCP: Weights variables using one of four options: (a) common variance, (b) specific variance, (c) multiple correlation, and (d) sum of squares.

WEIGHTING/INF: This program weights variables based on one of five information options: (a) joint information, (b) mutual information, (c) equivocation information, (d) multiple information, and (e) multiple joint information.

PCAR: Principal components analysis of a product moment, covariance, or correlation matrix is performed. Output includes the calculated cross-products matrix, eigenvalues and vectors, and component scores.

PLOT: Produces a two-dimensional scattergram on the high-resolution graphics screen. The input file is produced by one of the following programs: PCAR, CANCOR, EMVO, MDSCAL, CONCENTRATION. The program can plot different sets of ordination axes.

STEREO: Produces a three-dimensional stereogram. Input as in PLOT.

MDSCAL: Performs nonmetric multidimensional scaling (program modified from Brambilla and Salzano 1981). Input is a distance matrix between individuals calculated by program METRICS. The user must specify the dimensionality of the solution. The starting configuration may be random or user-defined. The internal

distance function may be Euclidean or a gaussian function (see Fewster and Orlóci 1983).

SLC: Performs single link agglomerative cluster analysis. Input is a similarity (program CORRELATION, option 3) or distance (METRICS, any option) matrix. Results are stored in a file which serves as input for producing a dendrogram (program TREE).

TREE: Produces a dendrogram on the high-resolution graphics screen, based on input from SLC, ALC, or SSA.

ALC: Performs centroid linkage agglomerative cluster analysis (algorithm UPGMC of Sneath and Sokal 1973). Input is a similarity (program CORRELATION, option 3 only) or distance (METRICS, any option) matrix.

SSA: Performs sum of squares agglomerative cluster analysis. Input must be a Euclidean distance matrix (program METRICS, options 3 or 4 only).

ADJUST: Performs an adjustment of the elements of a contingency table to equal block size as a prelude to analysis of concentration (program CONCENTRATION, options 1-3). Two p x n input files are required: (a) contingency table file, and (b) a file containing block sizes.

CONCENTRATION: Performs concentration analysis (dual scaling). Four options are available: (a) the first three are weighting variants of concentration analysis, as described in Orlóci (1978; pages 163-166). For these three options, run program ADJUST beforehand. (b) Option 4 performs correspondence analysis (Hill 1974), also known as reciprocal averaging.

LATTICE: Performs an additive partitioning of a contingency table using input from program CONCENTRATION.

GENDIST: Performs identification by calculating generalized distances between a priori groups and one or more individuals, measured on the same set of variables. The input data is as in program TESTCOV/2.

DISC: Identification by discriminant analysis is performed. Two external groups are assumed, and one individual requiring assignment.

INVERT/DET: Computes the inverse and determinant of a square matrix.

QEIGEN: Performs eigenanalysis of a square (symmetric or asymmetric) matrix.

COMBINATORIAL: This interactive program calculates the combinatorial formula.

CSA.RND/1: This program, which chains to two others, performs nested hierarchical character structure analysis as described in Orlóci *et al.* (1986). This analysis is available only with the Apple Macintosh version of the package.

Availability

The Orlóci, Kenkel and Orlóci (1988) manual describes methods and provides program documentation, listings and sample runs. The program package is available for the Apple] [series of personal computers (including the] [, //+, //c, and //e) in Applesoft BASIC, the Apple MacIntosh (Microsoft BASIC), and the IBM PC-XT (IBM Advanced BASIC). Manual and program package can be ordered through:

SCADA Associates,
P.O. Box 8059,
Substation 41,
London, Ontario N6G 2B0 Canada.

The exact price depends on the version requested and shipping charges.

REFERENCES

BARTELS, P.H., P.H. BAHR, D.W. CALHOUN, and G.L. WEID. 1970. Cell recognition by neighbourhood grouping techniques in Ticas. Acta Cytologia 14: 313-324.

BRAMBILLA, C. and G. SALZANO. 1981. A non-metric multidimensional scaling method for non-linear dimension reduction. Theory and computer program. Series III, Number 121, Istituto per le Applicazioni del Calcolo "Mauro Picone" Consiglio Nazionale delle Ricerche, Rome, Italy.

FEWSTER, P.H. and L. ORLÓCI. 1983. On choosing a resemblance measure for non-linear predictive ordination. Vegetatio 54: 27-35.

GOODALL, D.W. 1966. A new similarity index based on probability. Biometrics 22: 883-907.

HILL, M.O. 1974. Correspondence analysis: a neglected multivariate method. Appl. Stat. 23: 340-354.

MORRISON, D.F. 1976. *Multivariate Statistical Methods*. Second edition. McGraw-Hill, London.

ORLÓCI, L. 1978. *Multivariate Analysis in Vegetation Research*. Second edition. Dr. W. Junk, The Hague.

ORLÓCI, L., E. FEOLI, D. LAUSI and P.L. NIMIS. 1986. Estimation of character structure convergence (divergence) in plant communities: a nested hierarchical model. Coenoses 1: 11-20.

ORLÓCI, L. N.C. KENKEL and M. ORLÓCI. 1988. *Data Analysis in Population and Community Ecology*. U.W.O., London, Ontario. (Mimeographed).

PIELOU, E.C. 1966. The measurement of diversity in different types of biological collections. J. Theoret. Biol. 13: 131-144.

SNEATH, P.H.A. and R.R. SOKAL. 1973. *Numerical Taxonomy*. Freeman, San Francisco.