MULTIV

Multivariate Exploratory Analysis, Randomization Testing and Bootstrap Resampling

User's Guide v. 2.4

Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil

Introduction

MULTIV is a computer application program for performing flexible exploratory analysis, randomization testing and bootstrap resampling with multivariate data (Figure 1). It can handle qualitative, quantitative and mixed data types, offering several options for data transformation, resemblance measures, ordination and clustering techniques. The methods of exploratory multivariate analysis are well described elsewhere (see, i.a., Orlóci 1978, Pielou 1984, Digby & Kempton 1987, Orlóci et al. 1987, Podani 2000, Legendre & Legendre 1998). The results are presented in text files and, when applicable, in graphs. The versions of MULTIV for Windows do not create graphics.

Integrated to the exploratory tools, MULTIV can do fast randomization testing in univariate or multivariate comparisons between groups of sampling units defined by one or more factors, the latter being especially useful in the analysis of variance due to factor and interaction effects in experimental and survey data. The technique is described in Pillar & Orlóci 1996, with improvements after Pillar 2004. In addition, MULTIV can perform the multiresponse permutation procedure (MRPP, Mielke & Berry 2001). Randomization tests are also available for pair-wise comparisons of variables (see Manly 1991).

MULTIV can do bootstrap resampling to generate empirical confidence limits useful in estimation, to evaluate group partition sharpness in cluster analysis (Pillar 1999a) and to evaluate the significance of ordination dimensions (Pillar 1999b). Bootstrap resampling is integrated with process sampling (Orlóci & Pillar 1989) for examining sampling sufficiency (Pillar 1998). These functions were originally implemented in program SAMPLER.

This manual gives a description of MULTIV, with examples showing analytical pathways. The methods implemented are briefly presented without entering in details on appropriateness or advantages under different circumstances, which can be found in the general references already mentioned or in more specific ones throughout the text.

Hardware requirements

MULTIV is released in two versions for Macintosh and two for Windows OS:

(1) *Multiv* is the full version for Macintosh. It can run under Mac OS version 8.5 or later, including Mac OS X.

(2) *MultivMinor* is a demo version of *Multiv* which does everything the full version does, except that a maximum data size of 400 data elements (sampling units x variables) is allowed.

(3) *Multiv.exe* is the Windows OS version equivalent to *Multiv*, but not showing graphics.

(4) *MultivMi.exe* is the Windows OS version equivalent to *MultivMinor*, but not showing graphics.

To install MULTIV, copy a version that fits your hardware in your working volume and folder. It is advised to place MULTIV with the data files (the application, not its alias). Otherwise you will have to specify the full path when entering data file names. In any case, the output file Prinda.txt will be placed in the same folder with the application.

Memory allocation is dynamic, thus there is no a priori upper limit for the number of variables and sampling units that can be handled (except in *MultivMinor*). In the Macintosh, under

systems previous to MacOS X, if MULTIV runs out of memory it will give a message and exit. To increase the memory allocated to the application, the Macintosh user should quit MULTIV, select its icon in the Finder, choose Get Info in the File menu, and type a larger, compatible minimum size in the "Memory Requirements" box.

MULTIV can run in the background, that is, the user can operate on other applications that do not use the CPU as intensively while computations proceed in MULTIV, which is especially useful in randomization testing and bootstrap resampling with a large number of iterations.

Output

The numerical results are stored in a file named Prinda.txt, which can be open with any text editor. Under Mac OS X, scatter diagrams and dendrograms can be captured and saved by using the utility named Grab, which comes with Mac OS X; under previous systems, graphics can be stored as picture files by pressing the keys Shift Command 3 altogether while the graph is on the screen. With graphs on the screen, the user must close the graph's window in order to continue running the program. The versions for Windows OS do not produce graphical output.

MULTIV keeps track of the analytical steps performed on a data set, saving the information on a file named by default Sessao.trk. The user can choose to save the session with a different name (the extension .trk will be added automatically). Intermediate results are saved automatically in the session's file when working with small data sets. The program informs whether the session is saved or not. If not, the user should explicitly choose the option S "save session" in the main menu before quitting the program. Upon reopening, MULTIV will always check in its folder for the file Sessao.trk; if the file is found, the session with the last data file being used can be continued or not; if the file Sessao.trk or the session file pointed therein is not found, a new session is started. The user can also at any time open a previously defined session and continue the analysis.

Upon reopening, MULTIV will always check in the same folder for an existing file Prinda.txt, which if present can be appended with the new results or not.

Data input

The input data file for MULTIV must be a **text only** file, containing observations of one or more variables in a set of sampling units. You can use any text editor to create the data file, provided you save it "as text only". The data can be arranged in matrix format, with variables as rows and sampling units as columns, or vice-versa, or in a free format, provided entries in the data file are separated by blank spaces, tabs, carriage returns, or a combination of these. Actually, MULTIV will read as many data entries as the number of variables times the number of sampling units that were informed, in the order indicated; anything written after these will be disregarded (it is wise to use the end of the file to write useful information about the data, such as data origin, names and order of variables and sampling units, how they are arranged, etc.).

The user interface is interactive. The user should follow options offered by a sequence of menus shown on text-based screens. The options available are in general the ones that are applicable to the type of data informed by the user and analytical step that has been reached. In the "Preferences" menu the user can choose the language Portuguese or English. Major data entry is from text files. Numbers and letters (case insensitive) specify menu options. Keyboard input is processed upon pressing the Return key. Some questions may require multiple keyboard input, in which case it is enough to separate the entries by a blank space, a tab or a carriage return.

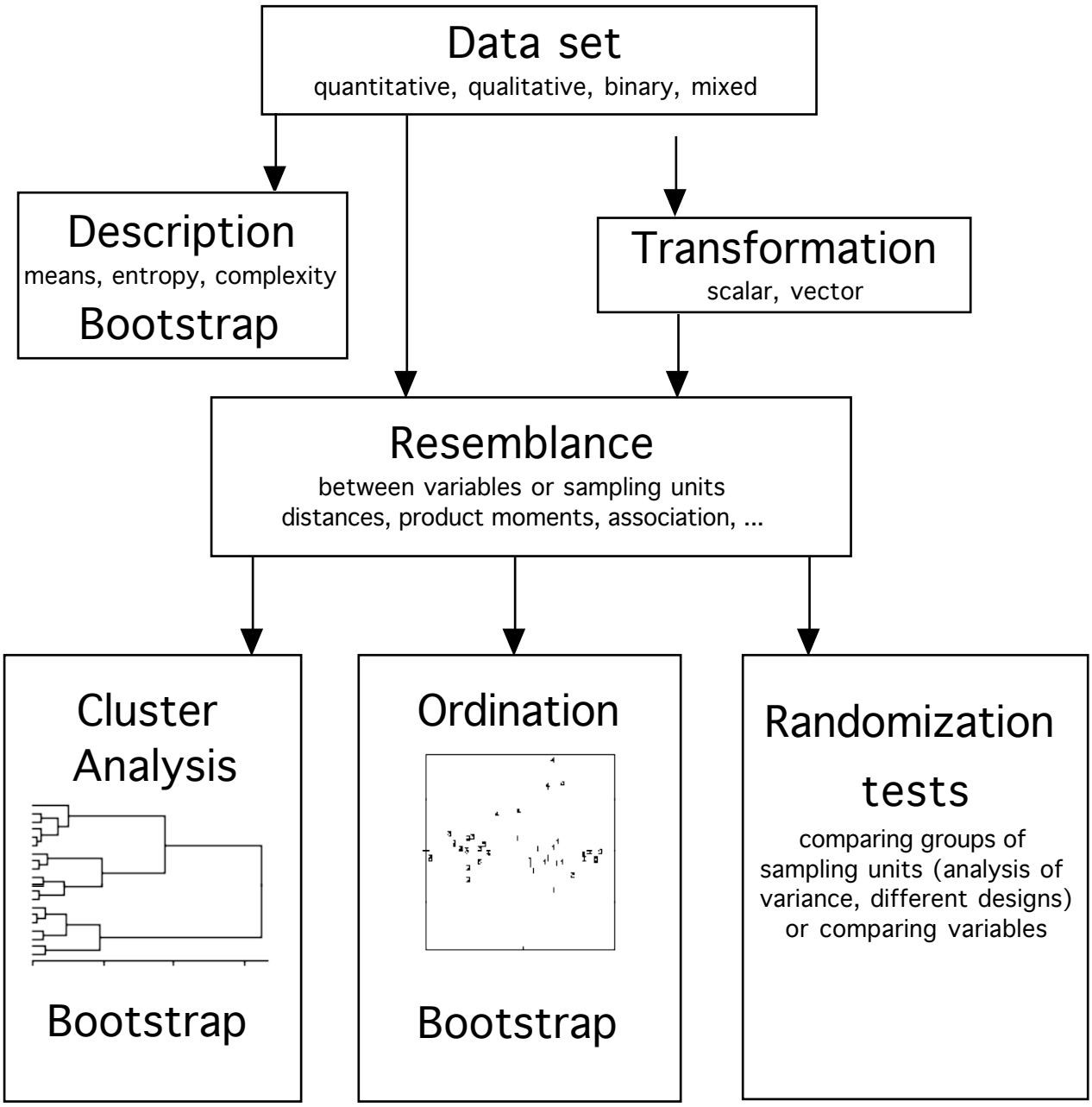


Figure 1. Flow diagram for exploratory analysis and randomization testing with MULTIV, showing options available at different analytical steps.

The options offered by the menus will be described in several sample runs that follow. Note that keyboard input is in *italics* and comments preceded by // are inserted.

```
MULTIV
for multivariate analysis, randomization tests and bootstrapping  version 2.4.0
------------------------------------------------------------------------------
…
------------------------------------------------------------------------------
Analysis status: Data not available.

------------------------------------------------------------------------------
MAIN MENU
    * N  specify data or open existing session
      V  descriptive attributes
      T  transform data
      R  resemblance measures
      G  specify groups of sampling units
      D  scatter diagrams
      O  ordination
      C  cluster analysis
      P  randomization tests comparing groups of sampling units
      A  randomization tests comparing variables
    * E  preferences
      S  save session
    * X  exit
------------------------------------------------------------------------------
Enter option (valid options with *): n

Wish to specify new data (n) or to open (o) an existing session? n/o n
Enter name of file containing data matrix: okop37.txt
```

// The data file must be in the same folder with MULTIV. Upon specifying the file, the program does not need to access it again afterwards, since the data matrix will be stored in the session's file. The session's file as well must be in the same folder with MULTIV. The small data set in file okop37.txt is from Orlóci et al. (1987, p.37):

26 29 29 29 30 35 39
28 31 31 33 27 38 36
18 14 13 13 19 15 15

1 2 3 4 5 6 7

Hkub Poch Pbre

OK&O p.37
3 variables (rows)
7 sampling units (columns)

```
Number of sampling units:    7
Number of variables: 3
Types of variables:
    (1) only quantitative variables, same measurement scale
    (2) only quantitative variables, different scales
    (3) only qualitative, non binary variables
```

```
    (4) only binary variables
    (5) several types
 Enter option: 1
```

// If option 5 is chosen, the type of each variable will have to be specified later.

```
Data ordered by (the rows of the matrix correspond to):
   (N) sampling units (u1v1,..., u1vp,..., unv1,..., unvp)
   (T) variables (u1v1,..., unv1,..., u1vp,..., unvp)
        Enter option: t
```

// The data is ordered by sampling units when in the file the first set of p entries refer to the observations in p variables in the first sampling unit, the second set of p entries to the second sampling unit, and so on up to the last set of p entries referring to the observations in the same p variables in the last sampling unit. If the data is arranged in a matrix, the rows correspond to the sampling units.
The data is ordered by variables when the first set of n entries refer to the observations in the first variable, the second set of n entries to the observations in the second variable, and so on up to the last set of n entries referring to the observations in the last variable. If the data is arranged in a matrix, the rows correspond to the variables. This is the case in file okop37.txt.

```
Labels for sampling units and variables:
    (1) given by default (1, 2, ...)
    (2) read labels after data on file okop37.txt
          (s. units first, labels with max. 5 alphanumeric symbols)
        Enter option: 2
Include data matrix in the output file? y/n y
```

// Labels for sampling units and variables are given by default (sequential numbers) or may be specified after the data on the same data file. The labels will be used in numerical results, dendrograms and scatter diagrams. The option to include the data matrix in the output file is to check whether the program read the data correctly.

// The current data file in the session is okop37.txt. If a new data is specified the information given about the previous data file will be lost. To avoid this problem you can save the session with a name of your choice before specifying a new data set, as in the following run; the session can be recovered afterwards.

```
Analysis status:
Data file name: okop37.txt
Dimensions: 7 sampling units, 3 variables
Data type: (1) quantitative, same measurement scales
Scalar transformation: (0)none
Vector transformation: (0)none
Session IS saved.

-------------------------------------------------------------------------------
MAIN MENU
    * N  specify data or open existing session
    * V  descriptive attributes
    * T  transform data
    * R  resemblance measures
    * G  specify groups of sampling units
    * D  scatter diagrams
      O  ordination
      C  cluster analysis
```

```
      P  randomization tests comparing groups of sampling units
      A  randomization tests comparing variables
    * E  preferences
    * S  save session
    * X  exit
------------------------------------------------------------------------------
Enter option (valid options with *): s
Enter name for the session: okop37_Session

Analysis status session okop37_Session:
Data file name: okop37.txt
Dimensions: 7 sampling units, 3 variables
Data type: (1) quantitative, same measurement scales
Scalar transformation: (0)none
Vector transformation: (0)none
Session IS saved.
```

// The following run will specify file *P&Op87.txt* with mixed data type:
1 0 1 1 1 1 1 1 1 1 1 1 1
0 1 0 0 0 0 0 0 0 0 0 1 1
2 1 2 1 2 2 2 2 2 1 1 1 2
1 3 1 1 1 1 1 2 2 1 1 1 1
3 6 2 1 1 4 5 5 4 3 4 4 2

Data from Pillar & Orloci (1993) p.87
5 variables, several types (rows) describing 13 sampling units (columns).
Variable types:
1 1 2 2 3

```
Analysis status session okop37_Session:
Data file name: okop37.txt
Dimensions: 7 sampling units, 3 variables
Data type: (1) quantitative, same measurement scales
Scalar transformation: (0)none
Vector transformation: (0)none
Session IS saved.

Analysis status:
Data file name: okop37.txt
Dimensions: 7 sampling units, 3 variables
Data type: (1) quantitative, same measurement scales
Scalar transformation: (0)none
Vector transformation: (0)none
Session IS saved.

------------------------------------------------------------------------------
MAIN MENU
    * N  specify data file or open existing session
    * V  descriptive attributes
    * T  transform data
    * R  resemblance measures
    * G  specify groups of sampling units
    * D  scatter diagrams
      O  ordination
      C  cluster analysis
      P  randomization tests comparing groups of sampling units
      A  randomization tests comparing variables
    * E  preferences
    * S  save session
```

```
    * X  exit
--------------------------------------------------------------------------------
Enter option (valid options with *): n

Wish to specify new data (n) or to open (o) an existing session? n/o n
Enter name of file containing data matrix: P&Op87.txt
Number of sampling units:    13
Number of variables: 5
Types of variables:
   (1) only quantitative variables, same measurement scales
   (2) only quantitative variables, different units
   (3) only qualitative, non binary variables
   (4) only binary variables
   (5) several types
        Enter option: 5
Enter variable type (1:binary,2:qualitative,3:quantitative):
Variable: 1  2  3  4  5
Type:     1  1  2  2  3
```

// Binary variables have two states, which must be zero or 1. Qualitative variables may have 2 or more states, which must be identified by integers. Quantitative variables are integers and/or real numbers.

```
Wish to correct? y/n n
Data ordered by (the rows of the matrix correspond to):
   (N) sampling units (u1v1,..., u1vp,..., unv1,..., unvp)
   (T) variables (u1v1,..., unv1,..., u1vp,..., unvp)
        Enter option: t
Labels for sampling units and variables:
    (1) given by default (1, 2, ...)
    (2) read labels after data on file P&Op87.txt
          (s. units first, labels with max. 5 alphanumeric symbols)
        Enter option: 1

Include data matrix in the output file? y/n y

See data on file Prinda.txt.
Wish to rename this session? y/n y
Enter name for the session: P&Op87_Session

Analysis status session P&Op87_Session:
Data file name: P&Op87.txt
Dimensions: 13 sampling units, 5 variables
Data type: (5) mixed
Variable: 1 2 3 4 5
Type:     1 1 2 2 3
Scalar transformation: (0)none
Vector transformation: (0)none
Session IS saved.
```

// Note that now the current session is P&Op87_Session, with a new data file (P&Op87.txt). The information given about the previous data file has been saved in session okop37_Session.

Descriptive attributes and bootstrap resampling

This option is for the computation of sample parameters such as means and standard deviations of variables (when variables are quantitative or binary) and entropy (when variables are qualitative or binary). Also, relevant if the data matrix is a contingency table, an option is offered to compute entropy and complexity measures for sampling units and for marginal totals in the

table. For methods see Orlóci (1991), Anand & Orlóci (1996) and references therein. A data set with all variables measured with the same unit is taken as a contingency table by MULTIV. The program uses Rényi entropy functions. One of them,

$$H^{\alpha} = \frac{\ln \sum_{i=1}^{s} p_i^{\alpha}}{1-\alpha}$$

measures entropy in a frequency distribution of n sampling units $\mathbf{P} = (p_1\ p_2\ ...\ p_s)$ such that $p_i = f_i/n$ and $p_1 + p_2 + ... + p_s = 1$ (since $n = f_1 + f_2 + ... + f_s$), where $f_1, f_2, ... f_s$ are counts in each of s classes in the variable being considered. The scale factor $\alpha \geq 0$ is specified by the user. If $\alpha=0$ is given, $H^0 = \ln s$ for any frequency distribution. If $\alpha=1$ is given, the program actually takes α lim 1 and uses Shannon's entropy

$$H = -\sum_{i=1}^{s} p_i \ln p_i$$

If the data matrix is taken as a contingency table, these functions are used to compute entropy with the column totals and with the row totals, that is, actually for the variable defining rows and for the variable defining columns. Mutual entropy and coherence coefficient between these two variables are also computed (see definitions in resemblance measures).

Kolmogorov complexity (Anand & Orlóci 1996) in each sampling unit is also computed when the data matrix is taken as a contingency table. Total complexity (L) is the total code length after Huffman coding of the frequency distribution in the sampling unit. Structural complexity is the difference between total complexity and Rényi entropy of a given order α.

The option for generating confidence limits by bootstrap resampling is offered in case means and standard deviations or entropy in variables are computed. The method generates confidence limits by the percentile method (Efron & Tibshirani 1993), combined with process sampling (Orlóci & Pillar 1989, Pillar 1998). It is based on data resampling of the sample with n sampling units, considered here as a pseudo sampling universe. At each of many iterations a sample (bootstrap sample) of size $n_k \leq n$ is extracted at random with replacement from the pseudo sampling universe and the chosen parameter (mean and standard deviation, or entropy) in the bootstrap sample are computed and stored. After performing B iterations the values are sorted from smallest to largest. For a given α probability the lower and upper confidence limits will be respectivelly the $B\alpha/2$ and $1+B(1-\alpha/2)$ –th values in the sorted list. The number of iterations is adjusted in order to have integers identifying the position of the limits. For instance, with 1000 iterations and $\alpha = 0.05$, the lower and upper confidence limits will be the values at the 25th and 976th positions respectivelly. Based on the confidence interval we can conclude, with an α probability of being wrong, that the true value of the parameter is in between the limits. Confidence intervals indicate whether a sample with size n_k is sufficient for the required precision. Process sampling involves generating confidence limits for increasing sample sizes $n_k \leq n$. By examining confidence limits over increasing sample sizes we can judge how well a better precision would compensate the effort of getting larger samples. The user can select the way process sampling is performed, as illustrated in the following example.

The run below illustrates the computation of means and standard deviations, followed by bootstrap confidence limits and process sampling:

```
Analysis status:
Data file name: EEASolo.txt
Dimensions: 15 sampling units, 8 variables
Data type: (2) quantitative, different units
Scalar transformation: (0)none
Vector transformation: (0)none
Session IS saved.

-------------------------------------------------------------------------------
MAIN MENU
    * N  specify data file or open existing session
    * V  descriptive attributes
    * T  transform data
    * R  resemblance measures
    * G  specify groups of sampling units
    * D  scatter diagrams
      O  ordination
      C  cluster analysis
      P  randomization tests comparing groups of sampling units
      A  randomization tests comparing variables
    * E  preferences
    * S  save session
    * X  exit
-------------------------------------------------------------------------------
Enter option (valid options with *): v


-------------------------------------------------------------------------------
DESCRIPTIVE ATTRIBUTES
Options:
    (1) means and standard deviations of variables
    (99)cancel and return to main menu
        Enter option: 1
Get confidence limits using bootstrap resampling? y/n y
```

// If the answer is yes confidence intervals for the mean and standard deviation in each variable will be generated and the information below will be requested. As specified in this example, process sampling will start with 10 sampling units, adding 2 sampling units at each step, until the total sample size is reached (15). At each sample size in process sampling, bootstrap samples will be taken to generate confidence limits for the mean and standard deviation at that sample size. A minimum of 1000 iterations for bootstrap resampling is recommended. By specifying the same number for initialization the analysis will reproduce identical results for the same data and options, but initialization of random numbers should normally be automatic.

```
Initial sample size: 10
Number of sampling units added at each sampling step: 2
Enter number of iterations in bootstrap resampling: 1000
Enter alpha probability for confidence limits: 0.05
Initialization of random number generation:
    (1) automatic
    (2) specify seed
    Enter option: 1
Save intermediate results? y/n  n
See results on file Prinda.txt.

Analysis status:
Data file name: EEASolo.txt
```

```
Dimensions: 15 sampling units, 8 variables
Data type: (2) quantitative, different units
Scalar transformation: (0)none
Vector transformation: (0)none
Session IS saved.
```

Data transformation

Table 1 indicates the relevance of transformations offered by MULTIV under different data types. Data transformation is available only when data is quantitative, binary or mixed with both. In scalar transformations the transformed value X'_{ij} in variable i, sampling unit j, will depend on the non-transformed value X_{ij} only, while in vector transformation X'_{ij} will depend on the set of observations in the same variable i (within variables), or in the same sampling unit j (within sampling units) or in both (double adjustment). Vector transformations within sampling units are only applicable when the variables are measured in the same scale (or standardized).

Scalar transformation

(1) *Presence/absence*. The transformed value $X'_{ij}=1$ if the original value $X_{ij}>0$ and $X'_{ij}=0$ if $X_{ij}=0$. After this transformation, data is treated as binary, to which will apply the relevant resemblance measures.

(2) *Square root*. The transformed value $X'_{ij}=\sqrt{|X_{ij}|}$.

(3) *Logarithm.* The transformed value $X'_{ij}=\log(|X_{ij}+1|)$

Vector transformation

The following options are available, either within variables or within sampling units:

(1) *Standardizing by marginal total*. Each observation X_{ij} is divided by the variable i total (within variables) or the sampling unit j total (within sampling units).

(2) *Centering and division by N-1* (implicit in covariance). See (3) below. This transformation does not involve any standardization.

(3) *Centering*. The variable i average (within variables) or sampling unit j average (within sampling units) is subtracted from each observation X_{ij}. Also, this transformation does not involve any standardization.

(4) *Normalization*. When it is within sampling units, the observations in each unit j are divided by $(\sum_{h=1}^{p} X_{hj}^2)^{1/2}$, where p is the number of variables. When normalization is within variables, the observations in each variable i are divided by $(\sum_{k=1}^{n} X_{ik}^2)^{1/2}$ where n is the number of sampling units.

(5) *Centering and (followed by) normalization* in the same direction. See (3) and (4) above.

(6) *Deviations from expected values based on marginal totals*. Assumes the input data matrix F is a contingency table in which each observation F_{jh} is a frequency count. The adjusted value $F'_{jh} = F_{jh}/ (F_{j.} F_{.h})^{1/2} - (F_{j.} F_{.h})^{1/2} / F_{..}$, where $F_{j.}$ is the variable j total , $F_{.h}$ is the sampling unit h total and $F_{..}$ is the grand total of the table.

(7) *Standardizing by the range*. Each observation X_{ij} is divided by the range in variable i (when within variables) or in sampling unit j (when within sampling units). The range is (SUP - INF), where SUP and INF are respectively the maximum and minimum values found in the data vector in consideration.

(8) *Rescaling into classes*. The range (see (7) above) observed in each variable i (within variables) or in each sampling unit j (within sampling units) is automatically equally divided in a specified number of classes. The transformed value X'_{ij} is the class (1, 2, ..., c classes) to which observation X_{ij} belongs.

(9) *Rescaling into ranks*. The observations are ranked from 1 to n sampling units (if within variables) or from 1 to p variables (if within sampling units). Tied ranks are adjusted according to Kendall and Gibbons 1990 p.40. Computing correlations between variables (sampling units) with the transformed data within variables (sampling units) will give the Spearman coefficient of rank correlation.

Table 1. Data transformations offered by MULTIV that are relevant under different data types.

	Quant. Same scales	Quant. Different scales	Binary	Qualit.	Mixed Quant.+ qualit.	Mixed Quant.+ binary	Mixed Qual.+b inary	All mixed
Scalar transformations	yes	yes	yes	no	no	yes	no	no
Vector transformation:								
within variables	yes	yes	yes	no	no	yes	no	no
within s. units	yes	no	yes	no	no	no	no	no

The following run performs data transformation in session okop37_Session, which has been previously saved:

```
…
Enter option (valid options with *): n

Wish to specify new data (n) or to open (o) an existing session? n/o o
Enter session name (without the extension .trk): okop37_Session

Analysis status session okop37_Session:
Data file name: okop37.txt
Dimensions: 7 sampling units, 3 variables
Data type: (1) quantitative, same measurement scales
Scalar transformation: (0)none
Vector transformation: (0)none
Session IS saved.
------------------------------------------------------------------------
MAIN MENU
    * N  specify data file or open existing session
    * V  descriptive attributes
```

```
    * T  transform data
    * R  resemblance measures
    * G  specify groups of sampling units
    * D  scatter diagrams
      O  ordination
      C  cluster analysis
      P  randomization tests comparing groups of sampling units
      A  randomization tests comparing variables
    * E  preferences
    * S  save session
    * X  exit
--------------------------------------------------------------------------------
Enter option (valid options with *): t

Analysis status session okop37_Session:
Data file name: okop37.txt
Dimensions: 7 sampling units, 3 variables
Data type: (1) quantitative, same measurement scales
Scalar transformation: (0)none
Vector transformation: (0)none
Session IS saved.
--------------------------------------------------------------------------------
DATA TRANSFORMATION
Options:
    (0)restore original data (no transformation)
    (1)specify transformations on original data
    (99)leave as is
        Enter option: 1

Scalar transformation:
    (0)none
    (1)presence/absence
    (2)square root (sqrt(|x|))
    (3)logarithm (log(|x+1|))
    (99)cancel and return to main menu
        Enter option: 0

Vector transformation:
    (0)none
    (1)within sampling units
    (2)within variables
        Enter option: 1

Vector transformation within sampling units:
 (includes double adjustment based on marginal totals)
    (0)none
    (1)standardizing by marginal total
    (2)centering and division by N-1 (implicit in covariance)
    (3)centering
    (4)normalization
    (5)centering and normalization (implicit in correlation)
    (6)deviations expected values via marginal totals (double adjustment)
    (7)standardizing by the range
    (8)rescaling into classes
    (9)rescaling into ranks
    (99)cancel and return to main menu
        Enter option: 4
Include data matrix in the output file? y/n y

See results on file Prinda.txt.
```

//When within variable vector transformation is selected first, a second vector transformation (only within sampling units) can be performed on the current transformed data, as shown in the following run:

```
...
Vector transformation:
        (0)none
        (1)within sampling units
        (2)within variables
        (0)none
                Enter option: 2

Vector transformation within variables:
 (includes double adjustment based on marginal totals)
        (0)none
        (1)standardizing by marginal total
        (2)centering and division by N-1 (implicit in covariance)
        (3)centering
        (4)normalization
        (5)centering and normalization (implicit in correlation)
        (6)deviations expected values via marginal totals (double adjustment)
        (7)standardizing by the range
        (8)rescaling into classes
        (9)rescaling into ranks
        (99)cancel and return to main menu
                Enter option: 4

Specify more transformations (only within sampling units)? y/n y

Vector transformation within sampling units:
 (include double adjustment based on marginal totals)
        (0)none
        (1)standardizing by marginal total
        (2)centering and division by N-1 (implicit in covariance)
        (3)centering
        (4)normalization
        (5)centering and normalization (implicit in correlation)
        (6)deviations expected values via marginal totals (double adjustment)
        (7)standardizing by the range
        (8)rescaling into classes
        (9)rescaling into ranks
        (99)cancel and return to main menu
                Enter option: 4
Include data matrix in the output file? y/n y
See results in file Prinda.txt.
```

Resemblance measures

The resemblance measure options available are the ones applicable to the type of data in hand (see Table 2). Note that some measures carry implicit transformations, which can also be performed explicitly in the previous menus. In the following descriptions, $X_{i\gamma}$ is the observation in variable i, sampling unit γ.

Quantitative and /or binary data

Product moment. The similarity of two sampling units α and β is $q_{\alpha\beta} = \Sigma (X_{i\alpha}X_{i\beta})$ for i = 1, ..., p variables. The similarity of two variables i and k is $s_{ik} = \Sigma (X_{i\gamma}X_{k\gamma})$ for γ = 1, ..., n sampling units. Note that no transformation is implicit. For other product moments, such as a centered product moment, covariance or correlation, previous data transformation can be specified. For instance, to compute centered product moments between variables, choose the option "centering" in vector transformation within variables; to compute covariances (between variables), choose the option "centering and division by n-1" in vector transformation within variables; to compute correlations between variables, choose the option "centering and normalization" in vector transformation within variables. Correlation can also be selected directly (see below).

Correlation. Instead of choosing product moment with previous transformation, correlation can also be selected directly, in which case centering and normalization in the proper direction is automatic.

Absolute value function, also known as Manhattan distance or city block distance. The distance between sampling units α and β is $a_{\alpha\beta} = \Sigma|X_{i\alpha} - X_{i\beta}|$ for i = 1, ..., p variables. The distance between variables i and k is $a_{ik} = \Sigma|X_{i\gamma} - X_{k\gamma}|$.

Euclidean distance. Between sampling units α and β the distance is $d_{\alpha\beta} = \{\Sigma(X_{i\alpha} - X_{i\beta})^2\}^{1/2}$ for i = 1, ..., p variables. Between variables i and k the distance is $d_{ik} = \{\Sigma(X_{i\gamma} - X_{k\gamma})^2\}^{1/2}$.

Chord distance. This distance between sampling units is equivalent to computing Euclidean distance with data normalized within sampling units. If the measure is between variables, data is normalized within variables. On choosing this option, transformation is automatic.

Legendre-Chodorowski index. This similarity coefficient (varying from 0 to 1) is for the comparison of sampling units only. It is implemented as described in Legendre & Legendre (1998, p. 267). The index is a composite of partial similarities computed for each variable, akin to Gower's index, but not applicable to qualitative data. A parameter (k) is requested.

Bray-Curtis dissimilarity or *percentage difference*. The dissimilarity (varying from 0 to 1) between sampling units α and β is $b_{\alpha\beta} = \Sigma|X_{i\alpha} - X_{i\beta}| / \Sigma|X_{i\alpha} + X_{i\beta}|$ for i = 1, ..., p variables. It does not hold Euclidean metric properties, thus may not be appropriate for ordination based on eigenanalysis nor to incremental sum of squares clustering.

Second eigenvalue of pairwise correlation. This is the function involved in hierarchical factorial clustering of variables (Denimal 2001, Camiz et al. 2006), and should be the resemblance choice for using that clustering method. See cluster analysis section later.

When the data file is a contingency table *Mutual information*. The data matrix must be a contingency table. To compare sampling units the function is

$$I_{\alpha\beta} = \sum_{i=1}^{p} \sum_{\substack{\gamma=\alpha \\ X_{i\gamma}>0}}^{\beta} X_{i\gamma} \ln \frac{X_{i\gamma} X_{..}}{X_{.\gamma} X_{i.}}$$

In this, p is the number of variables, $X_{.\gamma}$ is the marginal total for sampling unit α or β and $X_{i.}$ is the total of variable i in sampling units α and β, i.e., $X_{i\alpha} + X_{i\beta}$. The function measures the

mutual information between rows and columns classification criteria in the p variables x 2 sampling units contingency table extracted from the complete contingency table (see Feoli et al. 1984, Pillar & Orlóci 1993). To compare variables the function is also defined, but the contingency table is then 2 variables x n sampling units.

Table 2. Resemblance measures offered by MULTIV that are relevant under different data types for comparing sampling units.

	Quant. Same scales	Quant. Different scales	Binary	Qualit.	Mixed Quant.+ qualit.	Mixed Quant.+ binary	Mixed Qual.+b inary	All mixed
Resemblance measures between sampling units:								
Product moment	yes	no	yes	no	no	no	no	no
Absolute value function	yes	no	yes	no	no	no	no	no
Euclidean distance	yes	no	yes	no	no	no	no	no
Mutual information	yes	no	no	no	no	no	no	no
Gower index	yes	yes	yes	yes	yes	yes	yes	yes
MDIS	no	no	no	no	no	no	no	no
Mutual entropy	no	no	no	no	no	no	no	no
Rajski's metric	no	no	no	no	no	no	no	no
Coherence coefficient	no	no	no	no	no	no	no	no
Chi-square	no	no	no	no	no	no	no	no
Jaccard	no	no	yes	no	no	no	no	no
Simple matching	no	no	yes	yes	no	no	yes	no
Sokal-Michener	no	no	yes	no	no	no	no	no
Ochiai	no	no	yes	no	no	no	no	no
Sorensen	no	no	yes	no	no	no	no	no
Mean square contingency	no	no	yes	no	no	no	no	no
Correlation	yes	no	yes	no	no	no	no	no
Chord distance	yes	no	yes	no	no	no	no	no
Legendre&Chodorowski	yes	no	yes	no	no	no	no	no
Bray&Curtis	yes	no	yes	no	no	no	no	no

Table 3. Resemblance measures offered by MULTIV that are relevant under different data types for comparing variables.

	Quant. Same scales	Quant. Different scales	Binary	Qualit.	Mixed Quant.+ qualit.	Mixed Quant.+ binary	Mixed Qual.+b inary	All mixed
Resemblance measures between variables:								
Product moment	yes	no	yes	no	no	no	no	no
Absolute value function	yes	no	yes	no	no	no	no	no
Euclidean distance	yes	no	yes	no	no	no	no	no
Mutual information	yes	no	no	no	no	no	no	no
Gower index	no	no	no	no	no	no	no	no
MDIS	no	no	yes	yes	no	no	yes	no
Mutual entropy	no	no	yes	yes	no	no	yes	no
Rajski's metric	no	no	yes	yes	no	no	yes	no
Coherence coefficient	no	no	yes	yes	no	no	yes	no
Chi-square	no	no	yes	yes	no	no	yes	no
Jaccard	no	no	yes	no	no	no	no	no
Simple matching	no	no	yes	no	no	no	no	no
Sokal-Michener	no	no	yes	no	no	no	no	no
Ochiai	no	no	yes	no	no	no	no	no
Sorensen	no	no	yes	no	no	no	no	no
Mean square contingency	no	no	yes	no	no	no	no	no
Correlation	yes	yes	yes	no	no	yes	no	no
Chord distance	yes	yes	yes	no	no	yes	no	no
Legendre&Chodorowski	no	no	no	no	no	no	no	no
Bray&Curtis	no	no	no	no	no	no	no	no
Second eigenvalue pairwise correlation	yes	yes	yes	no	no	yes	no	no

Qualitative and/or binary data

Resemblance measures in this category are only applicable to comparisons between qualitative and binary variables. Most of the measures are based on information theory (see Orlóci 1991). For each pair-wise comparison a contingency table is computed containing joint frequencies of the variables' states. The contingency tables are saved in the file Tables.txt when the option to "Include data matrix in the output file?" is set to 'y'.

Mutual entropy. The Rényi information function of order α measures the entropy shared by two variables, given by

$$H_{ik}^{\alpha} = \frac{\ln \sum_{j=1}^{s_i} \sum_{h=1}^{s_k} \frac{p_{jh}^{\alpha}}{(p_{j\cdot} p_{\cdot h})^{\alpha-1}}}{\alpha - 1}$$

In this, f_{jh} is the joint frequency of states j and h, of variables i and k, s_i and s_k are the number of states in variables i and k, $p_{jh} = \frac{f_{jh}}{n}$, $p_{j\cdot} = \frac{f_{j\cdot}}{n}$, $p_{\cdot h} = \frac{f_{\cdot h}}{n}$, $f_{j\cdot} = \sum_{h=1}^{s_k} f_{jh}$ (frequency of state j, variable i), $f_{\cdot h} = \sum_{j=1}^{s_h} f_{jh}$ (frequency of state h, variable k) and α is a scaling factor, specified by the user, in the interval $0 \leq \alpha \leq \infty$ except $\alpha = 1$. When α approaches 1 ($\alpha = 1$ specified by the user), the function reduces to

$$H_{ik} = \sum_{j=1}^{s_i} \sum_{h=1}^{s_k} p_{jh} \ln \frac{p_{jh}}{p_{j.} p_{.h}}$$

In the resemblance matrix, each element in the diagonal (H_{ii}^{α}) is the entropy in the frequency distribution of the states of variable i; when α approaches 1, this is the Shannon-Weaver index. The coherence coefficient, Rajski's metric and MDIS are derived from the mutual entropy.

Coherence coefficient. Is a similarity index expressing mutual entropy in relative terms, varying from zero (no association) to 1 (strong association), defined by $\rho_{ik}^{\alpha} = \sqrt{1 - (d_{ik}^{\alpha})^2}$, where d_{ik}^{α} is the Rajski's metric (see below).

Rajski's metric. It is a dissimilarity, varying from zero (maximum association) to 1 (no association), defined as

$$d_{ik}^{\alpha} = \frac{H_{i+k}^{\alpha} - H_{ik}^{\alpha}}{H_{i+k}^{\alpha}}$$

In this, H_{i+k}^{α} is the joint entropy $H_{i+k}^{\alpha} = H_{ii}^{\alpha} + H_{kk}^{\alpha} - H_{ik}^{\alpha}$, where H_{ii}^{α} and H_{kk}^{α} are entropies in the frequency distributions of the states of variables i and k.

MDIS (Kulbach's Minimum Discriminant Information Statistics). The association between two qualitative variables i and k can be measured by

$$2I_{ik} = 2f_{..}H_{ik} = 2 \sum_{h=1}^{s_i} \sum_{j=1}^{s_k} [f_{hj} \ln (f_{hj}/f^{o}_{hj})]$$

where $f^{o}_{hj} = (f_{h.}f_{.j})/f_{..}$ is the expected value based on marginal totals, H_{ik} is the mutual entropy of order 1 and the other terms defined as for the mutual entropy. The function approximates the chi-square (below).

Chi-square. Measures the association between variables i and k as $\chi^2_{ik} = \sum_{h=1}^{s_i} \sum_{j=1}^{s_k} \frac{(f_{hj} - f^{o_{hj}})^2}{f^{o_{hj}}}$, being the terms as defined above.

Binary data

For each pair-wise comparison between variables or between sampling units a contingency table is computed containing joint frequencies of the variables' 0/1 states. As for qualitative data, the contingency tables are saved in the file Tables.txt when the option to "Include data matrix in

the output file?" is set to 'y'. When comparing variables (or sampling units), the terms are defined as:

		Variable k (sampling unit β)		Total
		1	0	
Variable i (sampling unit α)	1	a	b	a+b
	0	c	d	c+d
Total		a+c	b+d	q=a+b+c+d

When comparing variables, q is the total number of sampling units; when comparing sampling units, q is the total number of variables. The following similarity functions measure association, varying from zero (no association) to 1 (complete association). There is positive association between variables (or sampling units) if $a > (a+b)(a+c)/q$ and negative if $a < (a+b)(a+c)/q$.

Jaccard. S_{ik} (or $S_{\alpha\beta}$) $= \frac{a}{a+b+c}$

Simple matching. S_{ik} (or $S_{\alpha\beta}$) = (a+d)/q. Since a+d is the total number of matchings, this measure can also be used to compare sampling units on the basis of qualitative variables (Gower 1971), i.e., $S_{\alpha\beta}$ is the number of variables in which the units match divided by the number of variables.

Sokal-Michener. S_{ik} (or $S_{\alpha\beta}$) = a/q

Ochiai. S_{ik} (or $S_{\alpha\beta}$) $= \frac{a}{[(a+b)(a+c)]^{1/2}}$

Sorensen. S_{ik} (or $S_{\alpha\beta}$) $= \frac{2a}{2a+b+c}$

Mean square contingency. r^2_{ik} (or $r^2_{\alpha\beta}$) $= \frac{(ad-bc)^2}{[(a+b)(c+d)(a+c)(b+d)]}$

Note that r^2 can be derived from the previously described correlation coefficient (squared) and chi-square ($r^2 = \frac{\chi^2}{q}$).

Mixed data

Gower index. It is a similarity coefficient (varying from 0 to 1), only applicable to compare sampling units, defined as a weighted average of p partial similarities $S_{\alpha\beta h}$ between sampling units α and β (Gower 1971):

$$S_{\alpha\beta} = \frac{\sum_{h=1}^{p} t_{\alpha\beta h}\,\delta_{\alpha\beta h}}{\sum_{j=1}^{p} \delta_{\alpha\beta j}}$$

Where $t_{\alpha\beta h}$ and $\delta_{\alpha\beta h}$ are set according to the type of variable h:

(1) If variable h is binary, a similar rule as in the Jaccard coefficient prevails, i.e., $t_{\alpha\beta h} = 1$ and $\delta_{\alpha\beta h} = 1$ if h is present in both sampling units, $t_{\alpha\beta h} = 0$ and $\delta_{\alpha\beta h} = 1$ if the sampling units do not coincide in variable h, $t_{\alpha\beta h} = 0$ and $\delta_{\alpha\beta h} = 0$ if h is absent in both sampling units. MULTIV also offers the option for counting matching absences, a similar rule as in the Simple matching coefficient, that is, $t_{\alpha\beta h} = 1$ and $\delta_{\alpha\beta h} = 1$ if h is absent in both sampling units.

(2) If variable h is qualitative, $t_{\alpha\beta h} = 1$ if sampling units α, β coincide for variable h, and $t_{\alpha\beta h} = 0$ if they do not. In both cases $\delta_{\alpha\beta h} = 1$.

(3) If variable h is quantitative, $t_{\alpha\beta h} = 1 - \frac{|X_{\alpha h} - X_{\beta h}|}{(\max X_h - \min X_h)}$ and $\delta_{\alpha\beta h} = 1$, where $X_{\alpha h}$ is the value of variable h, sampling unit α and $X_{\beta h}$ is the value of the same variable in sampling unit β.

The following run computes Euclidean distances between sampling units:

```
------------------------------------------------------------------------------
MAIN MENU
    * N  specify data file or open existing session
    * V  descriptive attributes
    * T  transform data
    * R  resemblance measures
    * G  specify groups of sampling units
    * D  scatter diagrams
      O  ordination
      C  cluster analysis
      P  randomization tests comparing groups of sampling units
      A  randomization tests comparing variables
    * E  preferences
    * S  save session
    * X  exit
------------------------------------------------------------------------------
Enter option (valid options with *): r

Analysis status session okop37_Session:
Data file name: okop37.txt
Dimensions: 7 sampling units, 3 variables
Data type: (1) quantitative, same measurement scales
Scalar transformation: (0)none
Vector transformation: (0)none
Session IS saved.
------------------------------------------------------------------------------
RESEMBLANCE MEASURES
Options:
        (1)between sampling units
        (2)between variables
        (99)cancel and return to main menu
                Enter option: 1
```

```
Type of resemblance measure (1)between sampling units:
        (1)product moment
        (2)absolute value function
        (3)Euclidean distance
        (4)mutual information
        (5)Gower index
        (17)correlation
        (18)chord distance
        (19)Legendre-Chodorowski index
        (20)Bray-Curtis dissimilarity
        (99)cancel and return to main menu
                Enter option: 3

Include data matrix in the output file? y/n y
See results in file Prinda.txt.
```

The following run reads the first 4 binary and qualitative variables from file *P&Op87.txt* and computes a coherence coefficient between variables:

```
-------------------------------------------------------------------------------
MAIN MENU
...
-------------------------------------------------------------------------------
Enter option (valid options with *): n

Wish to specify new data (n) or to open (o) an existing session? n/o n
Enter name of file containing data matrix: P&Op87.txt
 Number of sampling units:    13
Number of variables: 4
Types of variables:
    (1) only quantitative variables, same measurement scale
    (2) only quantitative variables, different scales
    (3) only qualitative, non binary variables
    (4) only binary variables
    (5) several types
        Enter option: 5
Enter variable type (1:binary,2:qualitative,3:quantitative):
Variable: 1  2  3  4
Type:     1  1  2  2
Wish to correct? y/n n
Data ordered by (the rows of the matrix correspond to):
    (N) sampling units (u1v1,..., u1vp,..., unv1,..., unvp)
    (T) variables (u1v1,..., unv1,..., u1vp,..., unvp)
        Enter option: t

Labels for sampling units and variables:
    (1) given by default (1, 2, ...)
    (2) read labels after data on file P&Op87.txt
           (s. units first, labels with max. 5 alphanumeric symbols)
        Enter option: 1

Include data matrix in the output file? y/n y
See data on file Prinda.txt.
Wish to rename this session? y/n y
Enter name for the session: P&Op87_4var_Session
```

//In this case, since the file is ordered by variables, we could read the first four variables, ignoring the last one.

```
Analysis status session P&Op87_4var_Session:
```

```
Data file name: P&Op87.txt
Dimensions: 13 sampling units, 4 variables
Data type: (5) mixed
Variable: 1 2 3 4
Type:     1 1 2 2
Scalar transformation: (0)none
Vector transformation: (0)none
Session IS saved.

------------------------------------------------------------------------------
MAIN MENU
    * N  specify data file or open existing session
    * V  descriptive attributes
      T  transform data
    * R  resemblance measures
    * G  specify groups of sampling units
    * D  scatter diagrams
      O  ordination
      C  cluster analysis
      P  randomization testing comparing groups of sampling units
      A  randomization tests comparing variables
    * E  preferences
    * S  save session
    * X  exit
------------------------------------------------------------------------------
Enter option (valid options with *): r
…
------------------------------------------------------------------------------
RESEMBLANCE MEASURES
Options:
    (1)between sampling units
    (2)between variables
    (99)cancel and return to main menu
        Enter option: 2

Type of resemblance measure (2)between variables:
    (6)MDIS
    (7)mutual entropy
    (8)Rajski's metric
    (9)coherence coefficient
    (10)chi-square
    (99)cancel and return to main menu
        Enter option: 9
Enter order of entropy measure (alpha in Reyni's general entropy): 12
Include resemblance matrix in the output file? y/n y

See results on file Prinda.txt.
See contingency tables in file Tables.txt.

Analysis status session P&Op87_4var_Session:
Data file name: P&Op87.txt
Dimensions: 13 sampling units, 4 variables
Data type: (5) mixed
Variable: 1 2 3 4
Type:     1 1 2 2
Scalar transformation: (0)none
Vector transformation: (0)none
Resemblance measure: (9)coherence coefficient, (2)between variables
Order of entropy measure (alpha in Reyni's general entropy): 12
Session IS saved.
```

//When the resemblance measure involves the construction of a contingency table for each pair-wise comparison, and 'Include data matrix in the output file' is set to y, the contingency tables are saved in a file named Tables.txt.

Ordination

Ordination is offered if a resemblance matrix is available. Three methods based on eigenanalysis (principal coordinates analysis, principal components analysis and correspondence analysis) are available in MULTIV.

If the selected method is principal coordinates analysis or principal components analysis, the user may choose the option for performing bootstrapped ordination to evaluate the significance of ordination axes. The algorithm uses bootstrap resampling and has been proposed by Pillar (1999b). If bootstrap samples with increasing sizes are selected, the results can be used to evaluate sampling sufficiency (Pillar 1998). The output file in Prinda.txt will contain probabilities $P(\theta_i^o \geq \theta_i^*)$. The probability $P(\theta_i^o \geq \theta_i^*)$ is an indicator of the strength of the structure in an ordination, as compared to the ordination of a null data set containing variables with the same observed distribution but zero expected association. Setting an α probability threshold will help the interpretation of $P(\theta_i^o \geq \theta_i^*)$. A small $P(\theta_i^o \geq \theta_i^*)$, that is, smaller than α, will indicate that the ordination dimension in consideration is significantly more stable than that would be expected for the same dimension in the ordination of a random data set. In this case we reject the null hypothesis and conclude, with a probability $P(\theta_i^o \geq \theta_i^*)$ of being wrong, that the given ordination dimension is nontrivial and worthy of interpretation. Otherwise, we accept the null hypothesis and consider the ordination dimension unstable and indistinguishable from a random data ordination. Alternatively, we may adopt a more flexible view on the significance, and consider the resulting $P(\theta_i^o \geq \theta_i^*)$ values as representative of a gradient of reliability.

The algorithm for bootstrapped ordination involves Procrustean adjustments. The number of dimensions involved in Procrustean adjustments is likely to affect the fitted scores and their correlation with the reference scores. The interpretation of the probabilities should start from the highest (the least relevant) to the first ordination dimension. In metric ordination the dimensions are uncorrelated and meaningfully ordered by relevance. If the second dimension is significant, it is expected that the first will also be significant. Likewise, if the third dimension is significant, the second and first ones will be as well, and so on. Once dimension i is deemed significant, the test may stop and all lower and more relevant $i - 1$ dimensions are also considered significant irrespective of the corresponding probabilities.

An insufficient sample size may cause type II error; that is, the test may not detect significance of dimensions that would be found significant if the sample were larger. Sample size sufficiency is indicated by stability of probabilities under increase of sample sizes (Pillar 1998). If sample size is deemed sufficient and $P(\theta_i^o \geq \theta_i^*)$ is larger than α, the ordination axis is truly irrelevant. Otherwise, we may say that a larger sample size would be needed to reach a more confident conclusion.

Principal coordinates analysis.

Available for any resemblance matrix between sampling units (ordination of sampling units) or variables (ordination of variables). If the resemblance matrix contains dissimilarities they

are squared; if they are similarities (s_{ik}), they are transformed into squared dissimilarities ($\delta_{ik} = s_{ii} + s_{kk} - 2\ s_{ik}$). The squared dissimilarities δ_{ik} are transformed into products by $q_{ik} = -\frac{1}{2}(\delta_{ik} - \delta_{i.}/n - \delta_{k.}/n + \delta_{..}/n^2)$, where n is the number of sampling units (variables). Matrix **Q** is subjected to eigenanalysis. If **Q** has no Euclidean metric properties, meaning that negative eigenvalues were found, a warning is shown in the results, based on the discrepancy between the total sum of the eigenvalues and the trace of **Q**. The scores y_{ik} of the sampling units (variables) in each ordination axis are in the eigenvectors of **Q** adjusted according to the magnitude of the corresponding eigenvalue λ_i, such that the sum of squared scores equals the eigenvalue, that is, $y_{ik} = b_{ik}\ (\ \lambda_i\ /\ \sum_{h}^{n}\ b_{ih}^2\ )^{1/2}$. Correlation coefficients between the original descriptors and each ordination axis are computed, allowing the interpretation of variation in the ordination space.

Principal components analysis

Can be performed only when the resemblance function is a product moment or a correlation, no restrictions regarding data transformation (a covariance matrix can be obtained by applying the appropriate transformation). If the resemblance is between variables, the ordination is of sampling units, and vice-versa. The resemblance matrix **S** is subjected to eigenanalysis. The contribution of each component is indicated by the percentage of the corresponding eigenvalue in relation to the total (trace of **S**). The normalized eigenvectors (**B**) are the component coefficients. The scores of sampling units (variables) are computed by **B'A**, where **A** is the data matrix (after vector transformations, implicit in the resemblance measure or not). The rows in **A** correspond to variables when **S** is between variables, and vice-versa. The results also present the correlation coefficients between the original variables (or sampling units if **S** is between sampling units) and each component. The contribution of variable (sampling unit) i in ordination component k is defined as $u_{ik} = (\ b_{ki}\sqrt{s_{ii}}\ )^2\ /\ \sum_{h=1}^{p}\ u_{hk}$ where p is the number of variables (sampling units).

The following run performs a principal coordinates analysis:

```
Analysis status session okop37_Session:
Data file name: okop37.txt
Dimensions: 7 sampling units, 3 variables
Data type: (1) quantitative, same measurement scales
Scalar transformation: (0)none
Vector transformation: (0)none
Resemblance measure: (3)Euclidean distance, (1)between sampling units
Session IS saved.
------------------------------------------------------------------------------
MAIN MENU
    * N  specify data file or open existing session
    * V  descriptive attributes
    * T  transform data
    * R  resemblance measures
    * G  specify groups of sampling units
    * D  scatter diagrams
    * O  ordination
    * C  cluster analysis
      P  randomization testing comparing groups of sampling units
```

```
      A  randomization tests comparing variables
    * E  preferences
    * S  save session
    * X  exit
------------------------------------------------------------------------------
Enter option: o

Ordination method:
    (1)principal coordinates analysis
    (2)principal components analysis
    (3)correspondence analysis
    (99)return to main menu
        Enter option: 1
Save intermediate results? y/n y
See results in file Prinda.txt.
```

//The option for bootstrapped ordination that follows is only offered if the ordination is of sampling units and the method is principal coordinates analysis or principal components analysis. The maximum number of ordination axes to be monitored is bounded by the rank evaluated above, but it sets the minimum initial sample size in process sampling. Process sampling will take bootstrap samples with sizes varying from the initial to the total number of sampling units. The number of steps in process sampling will be a function of these choices. If the initial sample size is equal to the total number of sampling units, there will be only one step in process sampling.

```
Evaluate significance of ordination axes using bootstrap resampling? y/n y

Enter the maximum number of ordination principal axes to be monitored: 3

Initial sample size: 4

Number of sampling units added at each sampling step: 1
```

//A minimum of 1000 iterations in bootstrap resampling is recommended. At each step in process sampling, this number of bootstrap samples will be taken with replacement and subjected to the chosen ordination method. By specifying the same number for initialization the analysis will reproduce identical results for the same data and options, but initialization of random numbers should normally be automatic.

```
Enter number of iterations in bootstrap resampling: 1000
Initialization of random number generation:
    (1) automatic
    (2) specify seed
    Enter option: 1
Save intermediate results? y/n  n
Results saved on file Prinda.txt
```

//Since the option for bootstrapped ordination was set, in the Macintosh version the profiles with the probabilities $P(\theta_i^o \geq \theta_i^*)$ at increasing sample size for each ordination axis, starting from the least important axis are shown in sequence on screen. Under Mac OS X you can use the utility Grab to capture and save each profile. You have to close the profile window (click in the upper left box) for the next graph, if any, to appear on screen.

Correspondence analysis

The method is applicable if the data set is a contingency table (or at least, in MULTIV, all variables are in the same units). The data table must be previously transformed to deviations from expected values based on marginal totals (see data transformation above) and the resemblance measure must be a product moment between variables (or between sampling units). Though the data set may be a contingency table, by convention the rows are variables when the resemblance is comparing variables or the rows are sampling units when the resemblance is comparing sampling units. The resemblance matrix (**S**) is subjected to eigenanalysis, in which the t eigenvalues are the squared canonical coefficients R_1^2 , ..., R_t^2 . The eigenvectors of **S** are the α scores, which after adjustment correspond to the scores for the variables (if **S** compares variables):

$$U_{hm} = \alpha_{hm}\sqrt{\frac{X_{..}}{X_{h.}}}$$

where U_{hm} is the score of variable h in ordination axis m and $X_{h.}$ and $X_{..}$ are respectively the row h total and the grand total in the non-transformed contingency table. The scores for sampling units are computed by:

$$V_{jm} = \sum_{h=1}^{p} \frac{X_{hj}U_{hm}}{X_{.j}R_m}$$

where V_{jm} is the score of sampling unit j in the ordination axis m. If **S** is comparing sampling units, the U scores will correspond to sampling units and the V scores to variables.

Scatter diagrams

Two dimensional scatter diagrams can be produced on screen using as axes original variables, transformed variables or available ordination scores. Scatter diagrams use the same scale in the horizontal and vertical axes. The axes are not drawn through the origin. When the ordination is of sampling units, biplots can be produced using correlations between variables and ordination axes (see Podani 2000, p. 227). In this case the plot scale is defined by the sampling unit scores, while the correlations are rescaled to the maximum unit score.

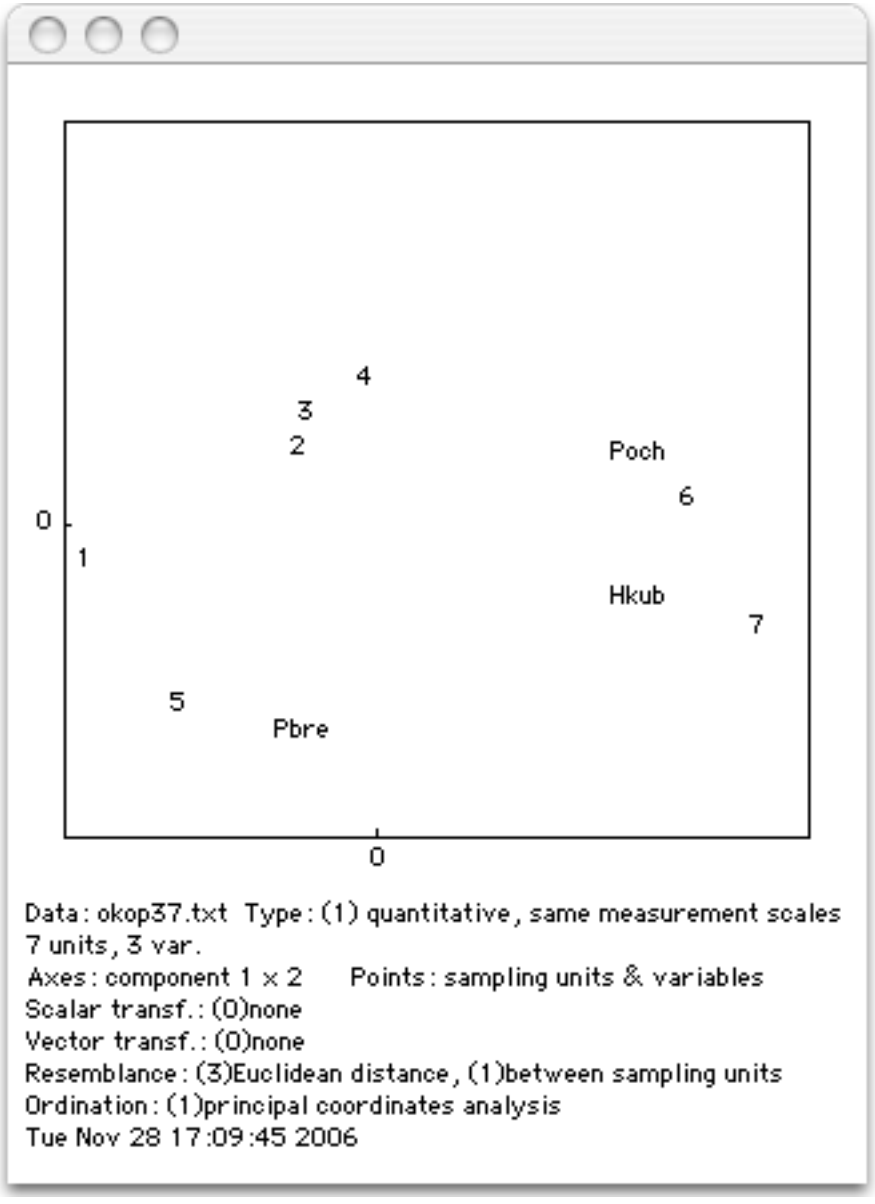
```
------------------------------------------------------------------------------
MAIN MENU
...
------------------------------------------------------------------------------
Enter option (valid options with *): d

Analysis status session okop37_Session:
Data file name: okop37.txt
Dimensions: 7 sampling units, 3 variables
Data type: (1) quantitative, same measurement scales
Scalar transformation: (0)none
Vector transformation: (0)none
Resemblance measure: (3)Euclidean distance, (1)between sampling units
Ordination scores available: (1)principal coordinates analysis
Session IS saved.
Use as coordinates of the scatter diagram:
        (1)original variables
        (2)original variables transformed
```

```
        (3)ordination scores
        (99)return to main menu
                Enter option: 3
Type of scatter diagram:
   1  plot sampling units only
   2  biplot of sampling units (scores) and variables (correlations with axes)
Type option no.: 2
Labels to use in the scatter:
   0  none (only squares)
   1  sampling unit labels
Type option no.: 1
Connect sampling unit points in scatter?  y/n n
Enter component for the horizontal axis: 1
Enter component for the vertical axis: 2
```

//The option to connect sampling unit points is relevant when they are spatially or temporally ordered

//The following window is shown on screen (only in the versions for Macintosh systems):



//Under Mac OS X the graph can be captured by using the system utility Grab. To close the window, click in the upper left box, after which the main menu will appear.

//Under Windows OS the diagram will not be produced, but based on file ScatterData.txt, stored in the same folder with Prinda.txt, a similar diagram can be drawn using other tools (if you use Excel, make sure the decimal separator is consistent with your system configuration).

Cluster analysis and dendrograms

The methods implemented are single linkage, complete linkage, average clustering (UPGMA and WPGMA), incremental sum of squares and a topological constrained version of it, and hierarchical factorial clustering (the latter only for clustering variables). The analysis uses the resemblance matrix available (between variables or between sampling units). The algorithm is agglomerative, at the beginning all groups are formed by one object (a variable or a sampling unit):

(1) Each agglomerative step joins the two most similar groups in the matrix.

(2) The resemblance matrix is redefined according to the new group that was formed in step 1.

(3) The process is repeated until all objects are joined in one group. There will be n-1 clustering steps, where n is the number of objects (variables or sampling units).

The clustering methods implemented differ in the criterion for redefinition of inter-group resemblance (step 2 above). The clustering process is automatically represented in a dendrogram on screen (Macintosh version only), which can be saved as explained for scatter diagrams. The dendrogram presents several partition possibilities; a decision which is left to the user. The numerical results in file Prinda.txt will also contain the possible partitions indicated by the analysis.

The user may choose the option for performing analysis of group partition sharpness using bootstrap resampling. The method is described in Pillar (1999a). If bootstrap samples with increasing sizes are selected, the results can be used to evaluate sampling sufficiency (Pillar 1998). The output file in Prinda.txt will contain probabilities $P(G^\circ \leq G^*)$ for each partition level, from 2 groups up to the maximum specified, and for each sample size. The probability $P(G^\circ \leq G^*)$ is the proportion of bootstrap iterations in which G° is found smaller than or equal to G^*. If $P(G^\circ \leq G^*)$ is not larger than a specified threshold α, we conclude, with a probability $P(G^\circ \leq G^*)$ of being wrong, that the k groups in the partition are not sharp enough to consistently reappear in resampling. That is, we reject the null hypothesis and conclude that the groups are fuzzy. If, instead, we accept the null hypothesis, we conclude that there is not enough evidence to refute that the groups are sharp.

Sample size influences the power of the group partition stability test. A small sample size will likely increase the error type II, the probability of accepting the null hypothesis when it is actually false. That is, a $P(G^\circ \leq G^*)$ larger than a chosen threshold α may either indicate the group structure is sharp or that the sample size is too small. Sample size sufficiency can be evaluated by examining the stability of probabilities across a series of sample sizes (process sampling). The user chooses the initial sample size and the number of sampling units added at each sampling step. For this, probabilities $P(G_k^0 \leq G_k^*)$ will be found by bootstrap resampling for each sample size $n_k \leq n$ taken from the observed sample with n sampling units (Pillar 1998). If the null hypothesis is accepted ($P(G^\circ \leq G^*) > \alpha$), the sample is considered sufficient if the probabilities reach stability within the range of sample sizes evaluated. If, instead, the null hypothesis is rejected ($P(G^\circ \leq G^*) \leq$

α), and the probabilities are decreasing and not yet stable, it is likely that the conclusion will not change with larger samples.

Single linkage (nearest neighbour)

When the resemblance matrix contains dissimilarities, the dissimilarity between groups P and Q is defined as:

d_{PQ} = INF [d_{jk}, for j=1, ..., n-1 and k=j+1, ..., n objects, provided j is in P and k in Q]

where d_{jk} is an element in the resemblance matrix and INF is the lowest value in the set indicated between brackets. When the resemblance matrix contains similarities, the similarity between groups is defined as above, but SUP is used instead.

Complete linkage

When the resemblance matrix contains dissimilarities, the dissimilarity between groups P and Q is defined as:

d_{PQ} = SUP [d_{jk}, for j=1, ..., n-1 e k=j+1, ..., n objects, provided j is in P and k in Q]

with terms as defined for the single linkage criterion. When the resemblance matrix contains similarities, INF is used instead.

Average clustering

The resemblance (similarity or dissimilarity) between groups P and Q is the arithmetic average of the resemblances between the objects in one group and the ones in the other group. In unweighted arithmetic average clustering (UPGMA), the resemblances are not weighted by group sizes. In weighted arithmetic average clustering (WPGMA), the resemblances are weighted by the group sizes. The implemented algorithms followed Legendre & Legendre (1998).

Incremental sum of squares

The method has been described by Ward (1963) and Orlóci (1967). It works on a dissimilarity matrix. If similarities are available, they are transformed into squared dissimilarities ($d_{ik}^2 = s_{ii} + s_{kk} - 2\, s_{ik}$). The clustering criterion minimizes

$$Q_{PQ} = \frac{n_p n_q}{(n_p+n_q)}\ d_{PQ}^2$$

where d_{PQ}^2 is the squared distance between centroids of groups P and Q, and n_p, n_q are the number of elements in groups p and q. Q_{PQ} can be computed by difference from the total and within sum of squares:

$$Q_{PQ} = Q_{P+Q} - Q_P - Q_Q$$

where:

$Q_{P+Q} = \frac{1}{n_p + n_q} \sum_h \sum_i d_{hi}^2$ for h=1, ..., n-1 and i=h+1, ..., n objects, provided h is in group P and i in group Q.

$Q_P = \frac{1}{n_p} \sum_h \sum_i d_{hi}^2$ for h=1, ..., n-1 and i=h+1, ..., n objects, provided h and i are in group P

$Q_Q = \frac{1}{n_q} \sum_h \sum_i d_{hi}^2$ for h=1, ..., n-1 and i=h+1, ..., n objects, provided h and i are in group Q

Q_{PQ} expresses the increment in the variance within group P+Q comparing to the variance existing within groups P and Q if isolated.

The option for performing a topological constrained clustering (Grimm 1987) with this method is also available. The user may indicate the topology is given by the neighborhoods in the sampling unit order in the data set or specify an upper half matrix with the topology (using 1 for adjacent pairs and 0 for otherwise). The option for evaluating group sharpness by bootstrap resampling is not available under topological constrained clustering.

Hierarchical factorial clustering of variables

The method, initially proposed by Denimal (2001, see Camiz et al. 2006), is based on a sequence of PCA-like eigenanalyses so that, at each step, a principal plane is created where we can represent both the variables belonging to the group and the units as seen by these variables only. In addition, the first principal component of each PCA is adopted as a representative variable of the corresponding group, a role similar to the centroid in the clustering of units.

The following run performs cluster analysis:

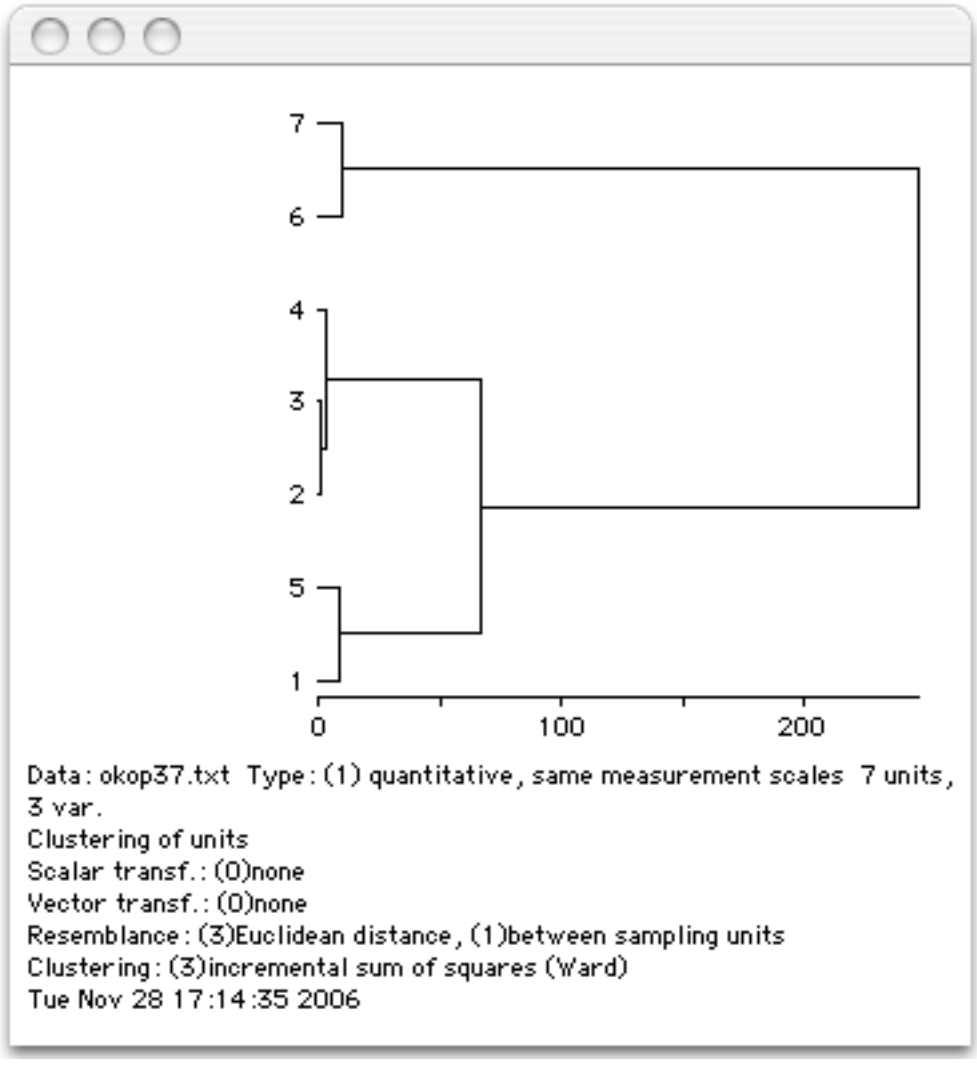
```
--------------------------------------------------------------------------------
MAIN MENU
    * N  specify data file or open existing session
    * V  descriptive attributes
    * T  transform data
    * R  resemblance measures
    * G  specify groups of sampling units
    * D  scatter diagrams
    * O  ordination
    * C  cluster analysis
      P  randomization testing comparing groups of sampling units
      A  randomization tests comparing variables
    * E  preferences
    * S  save session
    * X  exit
--------------------------------------------------------------------------------
Enter option (valid options with *): c

Analysis status session okop37_Session:
Data file name: okop37.txt
Dimensions: 7 sampling units, 3 variables
Data type: (1) quantitative, same measurement scales
Scalar transformation: (0)none
```

```
Vector transformation: (0)none
Resemblance measure: (3)Euclidean distance, (1)between sampling units
Ordination scores available: (1)principal coordinates analysis
Session IS saved.
Clustering method:
    (1)simple linkage (nearest neighbor)
    (2)complete linkage
    (3)incremental sum of squares (Ward)
    (4)average linkage (UPGMA)
    (5)average linkage (WPGMA)
    (6)hierarchical factorial (Denimal & Camiz)
    (7)topologically constrained incremental sum of squares
    (99)return to main menu
                Enter option: 3

See results in file Prinda.txt.
```

//The following window is shown on screen (only for Macintosh versions):



//Under Mac OS X the graph can be captured by using the system utility Grab. To close the window, click in the upper left box, after which the main menu will appear.
//Under Windows OS the graph will not be produced, but all possible partitions based on the dendrogram are presented in Prinda.txt.

//The option that follows, for evaluating group partition sharpness using bootstrap resampling, is only offered if the clustering is of sampling units. The maximum number of groups evaluated is bounded by the number of sampling units minus one and sets the minimum initial sample size in process sampling. Process sampling will take bootstrap samples with sizes varying from the initial to the total number of sampling units. The number of steps in process sampling will be a function of these choices. If the initial sample size is equal to the total number of sampling units, there will be only one step in process sampling.

```
Evaluate group partition sharpness using bootstrap resampling? y/n y
Enter the maximum number of groups (group partition level) to be evaluated: 3
Initial sample size: 4
Number of sampling units added at each sampling step: 1
```

//A minimum of 1000 iterations in bootstrap resampling is recommended. At each step in process sampling, this number of bootstrap samples will be taken with replacement and subjected to the chosen clustering method. By specifying the same number for initialization the analysis will reproduce identical results for the same data and options, but initialization of random numbers should normally be automatic.

```
Enter number of iterations in bootstrap resampling: 1000
Initialization of random number generation:
    (1) automatic
    (2) specify seed
    Enter option: 1
Save intermediate results? y/n  n
Results saved on file Prinda.txt
```

//Since the option for bootstrapped resampling was set, in the Macintosh version the profiles with the probabilities P($G° \leq G*$) at increasing sample size for each group partition level are shown in sequence on screen. Under Mac OS X you can use the utility Grab to capture and save each profile. You have to close the profile window (click in the upper left box) for the next graph, if any, to appear on screen.

Randomization tests: comparing variables

This option is offered when the available resemblance is between variables. The test criterion is any pair-wise resemblance function already defined and pertinent to the type of variables, e.g., correlation coefficient, chi-square, coherence coefficient, Jaccard, Euclidean distance. If the Null Hypothesis (Ho) of no association (correlation) between the variables pair-wise is true, the state observed in a variable in a given sampling unit is independent from the states observed in the other variables. For a data set with t variables and n sampling units, the complete reference set will contain $(n!)^{t-1}$ permutations. See Manly (1991) for details. A random data set member of the reference set is generated by shuffling the observations in each variable among the sampling units. At each random iteration a random member of the reference set is generated, the resemblance matrix is computed and each resemblance value for a variable pair (r_{ik}^o) is compared to the corresponding one found in the observed data set (r_{ik}). After many iterations, if $P(r_{ik}^o \geq r)$ is small (smaller than α), Ho is rejected and we conclude that there is significant association (correlation) between variables i and k. It is recommended to use at least 5000 iterations when $\alpha = 0.05$. The results present probabilities for all variable pairs.

The following run implements a randomization test to evaluate the significance of the coherence coefficient between qualitative variables:

```
Analysis status session P&Op87_4var_Session:
Data file name: P&Op87.txt
Dimensions: 13 sampling units, 4 variables
Data type: (5) mixed
Variable: 1 2 3 4
Type:     1 1 2 2
Scalar transformation: (0)none
Vector transformation: (0)none
Resemblance measure: (9)coherence coefficient, (2)between variables
Order of entropy measure (alpha in Reyni's general entropy): 12
Session IS saved.

------------------------------------------------------------------------------
HYPOTHESIS TESTING VIA RANDOMIZATION
Initialization of random numbers:
    (1) automatically
    (2) wish to specify a seed
        Enter option: 1
Number of iterations for randomization test: 1000
Save random data and results of each iteration? y/n n

See results on file Prinda.txt.
```

The results in file Prinda.txt are:

```
------------------------------------------------------------------------------
RANDOMIZATION TEST
------------------------------------------------------------------------------
Fri Feb 13 17:57:16 2004
Elapsed time: 1 seconds

Analysis status session P&Op87_4var_Session:
Data file name: P&Op87.txt
Dimensions: 13 sampling units, 4 variables
Data type: (5) mixed
Variable: 1 2 3 4
Type:     1 1 2 2
Scalar transformation: (0)none
Vector transformation: (0)none
Resemblance measure: (9)coherence coefficient, (2)between variables
Order of entropy measure (alpha in Reyni's general entropy): 12
Session IS saved.
Number of iterations: 1000
Random number generation initializer: 1076695032
Test criterion (Lambda):
      Resemblance measure: (9)coherence coefficient, (2)between variables
Results:
      (Resemblance matrix below the main diagonal and
 corresponding probabilities P(|Lambda random| >= |Lambda observed|) above the
main diagonal)
      0       0.234       0.375       0.227
0.71909           0       0.342       0.575
0.54919     0.61006           0           1
0.71909     0.39375           0           0
```

//The matrix presents each resemblance value below the main diagonal in row *i* column *k* and the corresponding probability above the main diagonal in row *k* column *i*. In this case, with a

4x4 matrix comparing the variables 1 and 3, the probability of finding a coherence coefficient larger than 0.54919 if Ho were true is 0.375.

Randomization tests: comparing groups of sampling units

This option is offered when the available resemblance is between sampling units. The procedures are identical for univariate or multivariate comparisons. Groups of sampling units must also be specified (see at the end of this section). Different test statistics can be chosen. When the sum of squares between groups or the pseudo F-ratio is selected as test statistic, the results are interpreted similarly to the ones in an analysis of variance table; the method is described in Pillar & Orlóci (1996), with improvements after Pillar (2004). MULTIV also implements MRPP (Mielke & Berry 2001). Pillar (2004) evaluated by data simulation the accuracy and power of the different methods described in the sequel.

The methods are based on two sets of information. The first is the matrix of p variables describing n units and the second is one or more factors with discrete states defining k groups of units. An n-by-n dissimilarity matrix obtained from the first matrix is needed for computing the chosen test statistic according to the k groups. MULTIV can perform the analysis with any type of resemblance matrix, irrespective of bearing or not Euclidean metric properties.

The test criterion may be the sum of squares between groups (Q_b) computed according to Pillar and Orlóci (1996):

$$Q_b = Q_t - Q_w$$

where

$$Q_t = \frac{1}{n}\sum_{h=1}^{n-1}\sum_{i=h+1}^{n} d_{hi}^2$$

is the total sum of squares of $n(n\text{-}1)/2$ pair-wise squared dissimilarities between n sampling units and

$$Q_w = \sum_{c=1}^{k} Q_{wc}$$

is the sum of squares within k groups, such that

$$Q_{wc} = \frac{1}{n_c}\sum_{h=1}^{n_c-1}\sum_{i=h+1}^{n_c} d_{hi|c}^2$$

where $d_{hi|c}^2$ is comparing units belonging to group c, which contains n_c sampling units.

In one-factor analysis, the k groups are given by the states of the factor. In two-factor analysis, as explained later, Q_b will be partitioned and the groups will be the ones defined by each factor in isolation, for the main effects, or their joint states for the interaction.

This partitioning of the total sum of squares based on the distance matrix is not constrained by the type of variables describing the units (e.g., means are not defined for qualitative variables), provided an appropriate dissimilarity measure is found that meets the requirements for the partitioning, allowing the use of randomization testing with qualitative and mixed data types.

The test criterion may as well be the pseudo F-ratio

$$F = Q_b / Q_w$$

In randomization testing, Q_b/Q_w is equivalent to the pseudo F-ratio used by Edgington (1987) and Anderson & ter Braak (2003), i.e. for the same permutations both statistics will give identical probabilities. There is no need to divide the sum of squares by the degrees of freedom in the numerator and denominator, since the degrees of freedom are constant over all permutation iterations. Furthermore, the F-ratio is equivalent to Q_b in randomization testing for one-factor multivariate analysis of variance; in two-factor designs they are also equivalent when the test is exact (this will be further explained in the next sections).

Furthermore, the chosen test criterion may be the average dissimilarity within groups, as described for the multiresponse permutation procedure (MRPP, Mielke & Berry 2001). In this the test criterion is

$$\delta = \sum_{c=1}^{k} \frac{1}{n_c} \xi_c$$

is a weighted average dissimilarity within k groups, such that

$$\xi_c = \frac{1}{n_c(n_c-1)/2} \sum_{h=1}^{n_c-1} \sum_{i=h+1}^{n_c} d_{hi|c}$$

The only option implemented for MRPP in MULTIV is this MRPP statistic using the non-squared dissimilarities (v = 1 in Mielke & Berry 2001). Another choice would be using the squared distances (v = 2, Mielke & Berry 2001) and $C_c = (n_c-1)/(n-k)$ as group weight for each group c with n_c sampling units, where k is the number of groups, but this statistic is equivalent to Q_b, that is, will give identical probabilities in a randomization test with the same set of permutations.

Random permutations in one-factor designs

If the null hypothesis (Ho) is true, the observation vector in a given unit is independent from the group to which the unit belongs. The observed data set is seen as one of the possible permutations of observation vectors among the units and their groups in the data matrix. An example is given in the following table:

Factor groups	1	2	3	4	1	2	3	4
Observation vector identities	1	2	3	4	5	6	7	8
One permutation:								
Factor groups	1	2	3	4	1	2	3	4
Observation vector identities	6	8	1	4	5	7	2	3

The observation vectors contain p variables and are permuted intact, preserving the correlation structure in the data. In some cases with more than one factor, random permutations of the vectors may be restricted within the levels of one factor or units belonging to the same factor combination may be permuted as blocks (more details will be given later). Over permutations, Q_t remains constant; it is only redistributed among Q_b and Q_w. Therefore, the test is exact (Anderson & ter Braak 2003). The same is valid for the MRPP.

Thus, the basis of randomization testing is to randomly permute the data according to Ho and for each of these permutations compute the test criterion (e.g. Q_b^o), comparing it to the value of Q_b found in the observed data. The probability $P(Q_b^o \geq Q_b)$ will be given by the proportion of permutations in which $Q_b^o \geq Q_b$. For the MRPP, since δ measures within group dispersion, the probability $P(\delta^o \leq \delta)$ is the relevant one, which will be given by the proportion of permutations in which the test statistic under H_0 is smaller than or equal to the observed test statistic ($\delta^o \leq \delta$).

The collection of all possible permutations, the reference set, can be generated systematically, but the number of permutations may be too many to be all considered in computations. A random but still large sample of this reference set is sufficient for generating the probabilities (Hope 1968, Edgington 1987). A minimum of 1000 and 5000 iterations for tests at the 5% and 1% significance levels is recommended to obtain reliable results, that is, P-values close to the exact ones that would be obtained in complete systematic data permutation (Edgington 1987, Manly 1991). By convention, the observed data set is included as one of these permutations, thus determining a minimum probability of 1/*B*, where *B* is the number of random permutations.

Ho will be rejected if $P(Q_b^o \geq Q_b) \leq \alpha$, where α is the significance level chosen for the test. If Ho is rejected, we conclude that the groups differ.

Contrasts

When there are more than two groups, contrasts are used to find which groups are different from each other. Contrasts are as well evaluated by randomization testing (Pillar & Orlóci 1996). The computation of sum of squares for evaluating contrasts is done similarly as in the previous explanations. Contrast coefficients help indicating which groups are involved in the contrast. For instance, with three groups, these contrasts could be defined as:

c_d: 0 1 -1 -> compares 2 vs. 3

c_e: 2 -1 -1 -> compares 1 vs. 2 and 3 taken with equal weight

For any contrast *i* there will be a total sum of squares involving the dissimilarities of units belonging to the groups in the contrast (Q_{ti}), a sum of squares within groups with positive contrast coefficients (Q_{wi+}), and a sum of squares within groups with negative coefficients (Q_{wi-}). The between groups sum of squares for the contrast *i* is

$$Q_{bi} = Q_{ti} - (Q_{wi+} + Q_{wi-})$$

For $u = k$-1 independent contrasts,

$$Q_b = \sum_{i=1}^{u} Q_{bi}$$

(To be noted is that two contrasts are independent when the product of the coefficient vectors is zero). Therefore, for each contrast *i* there is a test criterion Q_{ti} for which a probability will be found in randomization.

The permutations of observation vectors are done similarly as for the evaluation of the general Ho, but only for the vectors and units belonging to the groups in the contrast. In versions of MULTIV previous to version 2.3 permutations involved all vectors irrespective of belonging or not to the contrast and therefore some contrast probabilities, depending on the data, may have been inaccurate or overestimated. For instance, power may have been lower in contrasts having smaller differences when other contrasts in the same set had larger differences.

Random permutations in two-factor designs

In two-factor designs the *k* groups are defined by the joint states of the two factors, i.e., $k = l_A \text{ x } l_B$, where each factor is individually defining l groups (levels). The sum of squares between these *k* groups can be partitioned as

$$Q_b = Q_{b|A} + Q_{b|B} + Q_{b|AB}$$

where $Q_{b|A}$ is the sum of squares between l_A groups according to factor A disregarding factor B, $Q_{b|B}$ is the sum of squares between l_B groups according to factor B disregarding factor A, and $Q_{b|AB}$ is the sum of squares of the interaction AB, obtained by difference.

In two-factor designs, there is no general agreement on how permutations should be done (Edgington 1987, Manly 1991, Anderson & ter Braak 2003). Permutations may be unrestricted, similarly as for one-factor designs, but with the groups defined by the joint states of the two factors. For each permutation, the test criteria for each factor and for the interaction are computed and compared to the observed ones. In case the permutations are unrestricted, the denominator used in the F-ratio computation should be Q_w. An example is given in the following table:

Groups factor A	1 1 1 1 2 2 2 2
Groups factor B	1 1 2 2 1 1 2 2
Observation vector identities	**1 2 3 4 5 6 7 8**
One permutation:	
Groups factor A	1 1 1 1 2 2 2 2
Groups factor B	1 1 2 2 1 1 2 2
Observation vector identities	**6 8 1 4 5 7 2 3**

Unrestricted permutations will lead the test not being exact, for Q_t will be redistributed between more than two terms, i.e., $Q_{b|A}$, $Q_{b|B}$ and Q_w (Anderson & ter Braak 2003). However, depending on the data, the probabilities when Ho is true (type I error) may be acceptably close to the nominal significance level. No general solution has been found for an exact test of the interaction (Anderson & ter Braak 2003).

In two-factor crossed (not nested) designs, for testing one factor, permutations may be restricted to occur within the levels of the other factor (Edgington 1987). In this case, permutations will have to be done separately for finding the probabilities for the main effect of each factor. In the absence of interaction, such a restriction will give an exact test, for when testing for factor A, $Q_{b|B}$ will be kept constant over permutations and hence Q_t will be redistributed among $Q_{b|A}$ and Q_w, and vice-versa (Anderson & ter Braak 2003). The F-ratio is calculated using Q_w in the denominator. The following table gives an example of restricted permutation within the levels of factor A:

Groups factor A	1 1 1 1 2 2 2 2
Groups factor B	1 1 2 2 1 1 2 2
Observation vector identities	**1 2 3 4 5 6 7 8**

One permutation:	
Groups factor A	1 1 1 1 2 2 2 2
Groups factor B	1 1 2 2 1 1 2 2
Observation vector identities	**2 4 3 1 5 7 8 6**

In two-factor nested designs, for testing the nested factor, permutations should be restricted to occur within the levels of the nesting factor (Edgington 1987). This will give an exact test. For testing the nesting factor, however, observation vectors belonging to the same combination of the two factors would be permuted as blocks, which as well will give an exact test.

Block designs are nested designs in that the reference set is defined by restricting random allocations to within the blocks. The analysis is similar to a design with two factors (blocks, treatments), but in this case the sum of squares for the interaction term is not defined and the probability for the block factor is not computed.

Use of residuals

The use of observation vectors in which the effects of one or of both factors were removed has been suggested to overcome the impossibility of exact tests (e.g., for interactions) or when the available exact tests are too restrictive regarding power (McArdle & Anderson 2001, Anderson & ter Braak 2003). For this, residuals are computed in the data before obtaining the dissimilarity matrix. For testing the interaction in two-factor, crossed designs with multivariate data. the residuals are computed as

$$z_{hijk} = y_{hijk} - \bar{y}_{hi..} - \bar{y}_{h.j.} + \bar{y}_{h...}$$

In this, y_{hijk} is the observation of variable h in unit k, belonging to group i in factor A and to group j in factor B; $\bar{y}_{hi..}$ is the mean for variable h in factor A group i; $\bar{y}_{h.j.}$ is the mean for variable h in factor B group j; and $\bar{y}_{h...}$ is the overall mean for variable h in the data set. For testing factor A, the residuals were given by $z_{hijk} = y_{hijk} - \bar{y}_{h.j.}$ and for testing factor B they were $z_{hijk} = y_{hijk} - \bar{y}_{hi..}$.

General guidelines

In one-factor designs sum of squares and F-ratio will give identical probabilities (for the same random number generation initializer). The simulation results in Pillar (2004) indicated that for evaluating interactions in randomization testing in two-factor, crossed designs, the use of the F-ratio ($F = Q_b / Q_w$), with residuals, can improve accuracy and power. For testing factor main effects, however, the exact test, involving restricted permutations, with Q_b or F-ratio, is simpler and gave equivalent accuracy and power compared to using residuals and the F-ratio.

The accuracy and power in testing for factor effects is influenced by whether interaction is present or not. Therefore, in judging the tests for factor effects, it is better to interpret the probabilities for factors only when the interaction is detected as not significant. In fact, if the interaction is significant, it is recommended splitting the data according to the levels of one factor and then testing for effects of the other factor in each data set separately.

MULTIV offers the choice of using default options, which automatically will select for each case the combination of procedures that had been found by Pillar (2004) the most accurate and with higher power in data simulation. Users more acquainted with the intricacies of randomization testing and analysis of variance may decide to follow the advanced options.

If the groups are defined by one factor only, the default option will use the sum of squares as test statistics. If there are two factors, the first one explicitly defining blocks, the default option will as well use the sum of squares as test criterion, performing restricted permutations within blocks. For the probabilities in these cases it is irrelevant whether Q_b or F-ratio is used. For contrasts as well, the test criterion is irrelevant.

If the groups are defined by more than one factor, the default option will use different procedures and separate randomization tests for interactions and for main effects (the results, however, will be presented altogether in one table). For double interactions, residuals removing both factors involved in the interaction and the F-ratio as test criterion will be used for generating the probabilities (the observed sum of squares and F-ratio presented in the results, though, will be based on the raw data). For each main factor, restricted permutations (raw data, not residuals) within the groups defined by the combination of the other factors will be used. For comparing the groups defined by the joint states of all factors, unrestricted permutations with raw data will be used. In all cases, except for interactions as explained above, Q_b will be the test criterion.

Guidelines for testing factors and interactions under other multi-factor designs are found in Anderson & ter Braak (2003). In multifactor designs with more than 2 factors, MULTIV will be restricted to evaluate main effects and double interactions.

Specifying groups of sampling units

Groups must be specified before randomization comparing groups of sampling units. The user must inform the number and names of factors defining groups. If the design is a block design, the first factor always is blocks. The following run exemplifies the procedure (data from Orlóci et al. 1987, p.106):

```
Analysis status:
Data file name: okop106.txt
Dimensions: 12 sampling units, 1 variables
Data type: (1) quantitative, same measurement scales
Scalar transformation: (0)none
Vector transformation: (0)none

--------------------------------------------------------------------------
MAIN MENU
    * N  specify data file or open existing session
    * V  descriptive attributes
    * T  transform data
```

```
    * R  resemblance measures
    * G  specify groups of sampling units
    * D  scatter diagrams
    * O  ordination
    * C  cluster analysis
      P  randomization testing comparing groups of sampling units
      A  randomization tests comparing variables
    * E  preferences
    * S  save session
    * X  exit
------------------------------------------------------------------------------
Enter option (valid options with *): g
```

//Data set okop106.txt (from Orlóci et al. 1997, p.106) contains one variable describing 12 experimental units (the fact the data set is univariate will not change the procedure):

14 16 18 17 19 22 12 16 17 15 16 18

```
------------------------------------------------------------------------------
SPECIFY GROUPS
File: okop106.txt
Enter the number of factors that are defining groups: 2
Block design? y/n y
Enter the name of each factor:
Factor 1: Blocks
Factor 2: Diet
Read group memberships from file (f) or from the keyboard (k)? f/k k
Enter the group of each sampling unit:
Sampling unit:    1 2 3 4 5 6 7 8 9 10 11 12
Factor Blocks:
Groups:           1 1 1 2 2 2 3 3 3 4 4 4

Factor Diet:
Groups:           1 2 3 1 2 3 1 2 3 1 2 3
```

//If you choose option f the program will ask the name of the file containing the group memberships in the same order as if they were entered via keyboard.

//If option G is selected again, this partition will be shown on screen and the user may choose to specify new factors and partitions or leave as is.

The following run implements a randomization testing of treatment effect in data obtained in a block design, using default options:

```
Analysis status:
Data file name: okop106.txt
Dimensions: 12 sampling units, 1 variables
Data type: (1) quantitative, same measurement scales
Scalar transformation: (0)none
Vector transformation: (0)none
Resemblance measure: (3)Euclidean distance, (1)between sampling units
Session IS saved.

------------------------------------------------------------------------------
MAIN MENU
    * N  specify data or open existing session
    * V  descriptive attributes
    * T  transform data
    * R  resemblance measures
    * G  specify groups of sampling units
    * D  scatter diagrams
```

```
    * O  ordination
    * C  cluster analysis
    * P  randomization testing comparing groups of sampling units
      A  randomization tests comparing variables
    * E  preferences
    * S  save session
    * X  exit
------------------------------------------------------------------------------
Enter option (valid options with *): p
```

//Irrespective of having already specified the groups, the program will present the current groups partition for confirmation.

```
------------------------------------------------------------------------------
SPECIFYING GROUPS
File: p106.txt
Current group partition:
Sampling units:         1   2   3   4   5   6   7   8   9   10  11  12
Factor Blocks:
Groups:                 1   1   1   2   2   2   3   3   3   4   4   4
Factor Diet:
Groups:                 1   2   3   1   2   3   1   2   3   1   2   3
Change? y/n n

Analysis status:
Data file name: okop106.txt
Dimensions: 12 sampling units, 1 variables
Data type: (1) quantitative, same measurement scales
Scalar transformation: (0)none
Vector transformation: (0)none
Resemblance measure: (3)Euclidean distance, (1)between sampling units
Session IS saved.

------------------------------------------------------------------------------
HYPOTHESIS TESTING VIA RANDOMIZATION
Initialization of random numbers:
    (1) automatically
    (2) wish to specify a seed
        Enter option: 1
Use default (d) or wish to specify advanced (a) options? d/a d
Contrasts of groups:
    (0)no contrasts
    (1)perform all pairwise group contrasts
    (2)specify contrasts
    (99)return to main menu
        Enter option: 2

Enter the number of contrasts for factor Blocks (max. 3): 0
Enter the number of contrasts for factor Diet (max. 2): 2
Specify the coefficients in each contrast:
Groups:          1 2 3
Contrast 1:      1 -1 0
Contrast 2:      1 1 -2

Number of iterations for randomization test: 1000
Save random data and results at each iteration? y/n n
See results on file Prinda.txt.
```

The results presented on file Prinda.txt were:

```
------------------------------------------------------------------------------
```

```
RANDOMIZATION TEST
-------------------------------------------------------------------------------
Fri Feb 13 19:59:32 2004
Elapsed time: 1 seconds
Number of iterations (random permutations): 1000
Random number generation initializer: 1076702346
Group partition of sampling units:
Sampling units:    1 2 3 4 5 6 7 8 9 10 11 12
Factor Blocks:
Groups:              1 1 1 2 2 2 3 3 3 4 4 4
Order of groups in contrasts: 1 2 3 4

Factor Diet:
Groups:              1 2 3 1 2 3 1 2 3 1 2 3
Order of groups in contrasts: 1 2 3

Permuted data were vectors of raw data.
Randomization restrictions:
Random permutation within nesting groups:  1 1 1 2 2 2 3 3 3 4 4 4

Source of variation     Sum of squares(Q)        P(QbNULL>=Qb)
-------------------------------------------------------------------------------
Blocks:
Between groups                          31.333
-------------------------------------------------------------------------------
Diet:
Between groups                          36.167        0.007
   Contrasts:
 1 -1 0                                 10.125              0.123
 1 1 -2                                 26.042              0.029
Within groups                           3.1667
-------------------------------------------------------------------------------
Total                                   70.667

Mean vectors of each group:
Factor Blocks:
Group 1 (n=3):    16
Group 2 (n=3):    19.333
Group 3 (n=3):    15
Group 4 (n=3):    16.333
Factor Diet:
Group 1 (n=4):    14.5
Group 2 (n=4):    16.75
Group 3 (n=4):    18.75

Interaction factors Blocks x Diet:
Group 1 x 1 (n=1):      14
Group 1 x 2 (n=1):      16
Group 1 x 3 (n=1):      18
Group 2 x 1 (n=1):      17
Group 2 x 2 (n=1):      19
Group 2 x 3 (n=1):      22
Group 3 x 1 (n=1):      12
Group 3 x 2 (n=1):      16
Group 3 x 3 (n=1):      17
Group 4 x 1 (n=1):      15
Group 4 x 2 (n=1):      16
Group 4 x 3 (n=1):      18
```

//The probabilities indicate the effect of treatment was significant (P = 0.007) and that treatments 1 and 2 differed significantly (P = 0.029) from treatment 3.

The next run illustrates the analysis of a two-factor design (data from Orlóci et al. 1987, p.109), using default options:

```
Analysis status:
Data file name: okop109.txt
Dimensions: 24 sampling units, 1 variables
Data type: (1) quantitative, same measurement scales
Scalar transformation: (0)none
Vector transformation: (0)none
Resemblance measure: (3)Euclidean distance, (1)between sampling units
Session IS saved.

--------------------------------------------------------------------------------
MAIN MENU
    * N  specify data or open existing session
    * V  descriptive attributes
    * T  transform data
    * R  resemblance measures
    * G  specify groups of sampling units
    * D  scatter diagrams
    * O  ordination
    * C  cluster analysis
    * P  randomization testing comparing groups of sampling units
      A  randomization tests comparing variables
    * E  preferences
    * S  save session
    * X  exit
--------------------------------------------------------------------------------
Enter option (valid options with *): p
```

//In the data matrix in file okop109.txt 1 variable is describing 24 experimental units (the fact the data set is univariate will not change the procedure):

40 32 32 35 38 44 36 34 31 33 40 42 38
35 33 32 42 44 38 35 36 36 40 42

```
--------------------------------------------------------------------------------
SPECIFYING GROUPS
File: okop109.txt
Current group partition:
Sampling units:     1   2   3   4   5   6   7   8   9   10  11  12  13  14  15
16  17  18  19  20  21  22  23  24
Factor Lime:
Groups:                 1   1   1   2   2   2   1   1   1   2   2   2   1   1
1   2   2   2   1   1   1   2   2   2

Factor Nitrogen:
Groups:                 1   2   3   1   2   3   1   2   3   1   2   3   1   2
3   1   2   3   1   2   3   1   2   3
Change? y/n n

Analysis status:
Data file name: okop109.txt
Dimensions: 24 sampling units, 1 variables
Data type: (1) quantitative, same measurement scales
Scalar transformation: (0)none
Vector transformation: (0)none
Resemblance measure: (3)Euclidean distance, (1)between sampling units
Session IS saved.

--------------------------------------------------------------------------------
HYPOTHESIS TESTING VIA RANDOMIZATION
```

```
Initialization of random numbers:
    (1) automatically
    (2) wish to specify a seed
        Enter option: 1
Use default (d) or wish to specify advanced (a) options? d/a d
Contrasts of groups:
    (0)no contrasts
    (1)perform all pairwise group contrasts
    (2)specify contrasts
    (99)return to main menu
        Enter option: 1
Number of iterations for randomization test: 1000
Save random data and results at each iteration? y/n n

See results on file Prinda.txt.
```

The results were:

```
--------------------------------------------------------------------------------
RANDOMIZATION TEST
--------------------------------------------------------------------------------
Tue Nov 28 17:37:11 2006
Elapsed time: 0.561544 seconds
Number of random permutations plus observed data set: 1000
Random number generation initializer: 1164735406
Group partition of sampling units:
Sampling units:    1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
24
Factor Lime:
Groups:            1 1 1 2 2 2 1 1 1 2 2 2 1 1 1 2 2 2 1 1 1 2 2 2
Order of groups in contrasts: 1 2

Factor Nitrogen:
Groups:            1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3
Order of groups in contrasts: 1 2 3

Vectors of raw data were used to generate probabilities for the main effects
while vectors of residuals were used for double interactions (both factor
effects removed).
For each main effect permutations were restricted within the groups defined by
the combination of the other factors.
For the all-factor combination effect permutations were unrestricted.

(*) Probabilities P generated for sum of squares (Qb), except for interactions,
where F=Qb/Qw was used as test criterion.


Source of variation     Sum of squares(Q)        P(QbNULL>=Qb)*
--------------------------------------------------------------------------------
Factor Lime:
Between groups                              96    0.017
   Contrasts:
 1 -1                                       96          0.02
--------------------------------------------------------------------------------
Factor Nitrogen:
Between groups                              16    0.51
   Contrasts:
 1 -1 0                                      4          0.608
 1 0 -1                                     16          0.362
 0 1 -1                                      4          0.387
--------------------------------------------------------------------------------
Lime        x Nitrogen                     208    0.001
--------------------------------------------------------------------------------
```

```
Between groups                          320    0.001
Within groups                            50
------------------------------------------------------------------------------
Total                                   370

Mean vectors of each group:
Factor Lime:
Group 1 (n=12):   35
Group 2 (n=12):   39
Factor Nitrogen:
Group 1 (n=8):    36
Group 2 (n=8):    37
Group 3 (n=8):    38

Interaction factors Lime x Nitrogen:
Group 1 x 1 (n=4):      38
Group 1 x 2 (n=4):      34
Group 1 x 3 (n=4):      33
Group 2 x 1 (n=4):      34
Group 2 x 2 (n=4):      40
Group 2 x 3 (n=4):      43
```

//Since the interaction between the two factors was significant, for examining main factor effects the data set should be splitted so that, e.g, the effect of nitrogen is tested within each level of lime, or vice versa.

In the next run, advanced options were set for the analysis with the same data set okop109.txt and groups defined by two factors. The analysis used permutation of residuals removing the effects of both factors:

```
Analysis status:
Data file name: okop109.txt
Dimensions: 24 sampling units, 1 variables
Data type: (1) quantitative, same measurement scales
Scalar transformation: (0)none
Vector transformation: (0)none
Resemblance measure: (3)Euclidean distance, (1)between sampling units
Session IS saved.

------------------------------------------------------------------------------
HYPOTHESIS TESTING VIA RANDOMIZATION
Initialization of random numbers:
    (1) automatically
    (2) wish to specify a seed
        Enter option: 1
Use default (d) or wish to specify advanced (a) options? d/a a
Choose the test statistic:
  (1) Sum of squares between groups (Qb, Pillar & Orloci 1996)
  (2) Pseudo F-ratio (Anderson 2001)
  (3) Average dissimilarity within groups (MRPP for v=1, Mielke & Berry 2001)
      Enter option: 2

Use residuals in the analysis?
  (0) No
  (1) Remove the effect of one factor
  (2) Remove the effect of two factors
      Enter option: 2

Specify the factors whose effects are to be removed ( 1.Lime 2.Nitrogen ):  1 2
```

```
Randomization restrictions:
Nesting (like a block design or a nested design):
  (0) None
  (1) Permutations restricted within groups of one factor already specified
  (2) Permutations restricted within groups to be specified
      Enter option: 0

Additional randomization restrictions:
Permutation among unit bundles (e.g., main plots in a nested design or when
subsampling is involved):
  (0) None
  (1) Bundles according to one factor already specified
  (2) Bundles to be specified
      Enter option: 0

Contrasts of groups:
    (0)no contrasts
    (1)perform all pairwise group contrasts
    (2)specify contrasts
    (99)return to main menu
        Enter option: 0
Number of iterations for randomization test: 10000
Save random data and results at each iteration? y/n n

See results on file Prinda.txt.
```

The results were:

```
--------------------------------------------------------------------------------
RANDOMIZATION TEST
--------------------------------------------------------------------------------
Tue Nov 28 17:47:28 2006
Elapsed time: 0.577719 seconds
Number of random permutations plus observed data set: 10000
Random number generation initializer: 1164736011
Group partition of sampling units:
Sampling units:    1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
24
Factor Lime:
Groups:            1 1 1 2 2 2 1 1 1 2 2 2 1 1 1 2 2 2 1 1 1 2 2 2

Factor Nitrogen:
Groups:            1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3

Permuted data were vectors of residuals (effects of factors Lime and Nitrogen
were removed).
Randomization restrictions: none


Source of variation     Sum of squares(Q)            F=Qb/Qw       P(FNULL>=F)
--------------------------------------------------------------------------------
Factor Lime:
Between groups                    0                   0          1
--------------------------------------------------------------------------------
Factor Nitrogen:
Between groups                    0                   0          1
--------------------------------------------------------------------------------
Lime        x Nitrogen          208                4.16          0.0001
--------------------------------------------------------------------------------
Between groups                  208                4.16          0.0001
Within groups                    50
```

```
------------------------------------------------------------------------------
Total                                    258

Mean vectors of each group:
Factor Lime:
Group 1 (n=12):   35
Group 2 (n=12):   39
Factor Nitrogen:
Group 1 (n=8):    36
Group 2 (n=8):    37
Group 3 (n=8):    38

Interaction factors Lime x Nitrogen:
Group 1 x 1 (n=4):      38
Group 1 x 2 (n=4):      34
Group 1 x 3 (n=4):      33
Group 2 x 1 (n=4):      34
Group 2 x 2 (n=4):      40
Group 2 x 3 (n=4):      43

Analysis status:
Data file name: okop109.txt
Dimensions: 24 sampling units, 1 variables
Data type: (1) quantitative, same measurement scales
Scalar transformation: (0)none
Vector transformation: (0)none
Resemblance measure: (3)Euclidean distance, (1)between sampling units
Session IS saved.
```

//The sum of squares for the factors Lime and Nitrogen are zero since the effects of both factors were removed in the observed data.

In the next analysis advanced options were set to perform the test with random permutations restricted within the levels of the factor Lime:

```
Analysis status:
Data file name: okop109.txt
Dimensions: 24 sampling units, 1 variables
Data type: (1) quantitative, same measurement scales
Scalar transformation: (0)none
Vector transformation: (0)none
Resemblance measure: (3)Euclidean distance, (1)between sampling units
Session IS saved.

------------------------------------------------------------------------------
HYPOTHESIS TESTING VIA RANDOMIZATION
Initialization of random numbers:
    (1) automatically
    (2) wish to specify a seed
        Enter option: 1
Use default (d) or wish to specify advanced (a) options? d/a a
Use as statistics sum of squares between groups (s) or an F-ratio (f)? s/f s

Use residuals in the analysis?
  (0) No
  (1) Remove the effect of one factor
  (2) Remove the effect of two factors
      Enter option: 0

Randomization restrictions (for testing factor effects and interactions):
```

```
Nesting:
  (0) None
  (1) Permutations restricted within levels of one of the factors
  (2) Permutations restricted within levels to be specified
      Enter option: 1

Specify which is the nesting factor ( 1.Lime 2.Nitrogen ):  1

Blocking (sets of units permuted as blocks):
  (0) None
  (1) Levels of one factor forming blocks
  (2) Blocks to be specified
      Enter option: 0
Contrasts of groups:
    (0)no contrasts
    (1)perform all pairwise group contrasts
    (2)specify contrasts
    (99)return to main menu
        Enter option: 1
Number of iterations for randomization test: 10000
Save random data and results at each iteration? y/n n

See results on file Prinda.txt.
```

The results were:

```
------------------------------------------------------------------------------
RANDOMIZATION TEST
------------------------------------------------------------------------------
Tue Nov 28 17:51:26 2006
Elapsed time: 1.72367 seconds
Number of random permutations plus observed data set: 10000
Random number generation initializer: 1164736245
Group partition of sampling units:
Sampling units:    1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
24
Factor Lime:
Groups:            1 1 1 2 2 2 1 1 1 2 2 2 1 1 1 2 2 2 1 1 1 2 2 2
Order of groups in contrasts: 1 2

Factor Nitrogen:
Groups:            1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3
Order of groups in contrasts: 1 2 3

Permuted data were vectors of raw data.
Randomization restrictions:
Random permutation within nesting groups:  1 1 1 2 2 2 1 1 1 2 2 2 1 1 1 2 2 2
1 1 1 2 2 2

Source of variation     Sum of squares(Q)        P(QbNULL>=Qb)
------------------------------------------------------------------------------
Factor Lime:
Between groups                              96    1
   Contrasts:
 1 -1                                       96          1
------------------------------------------------------------------------------
Factor Nitrogen:
Between groups                              16    0.5385
   Contrasts:
 1 -1 0                                      4          0.5915
 1 0 -1                                     16          0.3875
 0 1 -1                                      4          0.3734
------------------------------------------------------------------------------
```

```
Lime      x Nitrogen                    208    0.0001
-------------------------------------------------------------------------
Between groups                          320    0.0001
Within groups                            50
-------------------------------------------------------------------------
Total                                   370

Mean vectors of each group:
Factor Lime:
Group 1 (n=12):   35
Group 2 (n=12):   39
Factor Nitrogen:
Group 1 (n=8):    36
Group 2 (n=8):    37
Group 3 (n=8):    38

Interaction factors Lime x Nitrogen:
Group 1 x 1 (n=4):      38
Group 1 x 2 (n=4):      34
Group 1 x 3 (n=4):      33
Group 2 x 1 (n=4):      34
Group 2 x 2 (n=4):      40
Group 2 x 3 (n=4):      43
```

//In this case random permutations were restricted within factor Lime and thus the corresponding probabilities for factor Lime are all equal to one.

Setting preferences

The user can change the default settings for the language of the interface and results (English or Portuguese), the sizes of dendrograms and scatter diagrams, and the number of significant digits in outputs. The default settings are used until anything different is selected in the Preferences menu, after which a file named MULTIV.prf is placed in the same folder with the program. The new settings are then used as long as this file is kept with the program. When no MULTIV.prf file is found, the default settings are used.

```
-------------------------------------------------------------------------
MAIN MENU
    * N  specify data file or open existing session
    * V  descriptive attributes
    * T  transform data
    * R  resemblance measures
    * G  specify groups of sampling units
    * D  scatter diagrams
    * O  ordination
    * C  cluster analysis
    * P  randomization tests
      A  randomization tests comparing variables
    * E  preferences
    * S  save session
    * X  exit
-------------------------------------------------------------------------
Enter option (valid options with *): e

PREFERENCES
Set preferences of:
   L  language (change to Portuguese/mudar para português)
   S  scatter diagram
```

```
   D  dendrograms
   F  output
   C  return to main menu
Type option: s

-------------------------------------------------------------------------------
Dimensions of the scattergram in mm (zero for default  height 88 x width 88
mm):
 height (max. 183): 100
Scattergram height and width are the same.

PREFERENCES
Set preferences of:
   L  language (change to Portuguese/mudar para português)
   S  scatter diagram
   D  dendrograms
   F  output
   C  return to main menu
Type option: d

-------------------------------------------------------------------------------
Dimensions of the dendrograms in mm (zero for default  height 73 x width 73
mm):
 height (max. 187): 100
 width (max. 240):  100

PREFERENCES
Set preferences of:
   L  language (change to Portuguese/mudar para português)
   S  scatter diagram
   D  dendrograms
   F  output
   C  return to main menu
Type option: f

-------------------------------------------------------------------------------
Number of significant digits on printouts (enter zero for default):
10

PREFERENCES
Set preferences of:
   L  language (change to Portuguese/mudar para português)
   S  scatter diagram
   D  dendrograms
   F  output
   C  return to main menu
Type option: c
```

References

Anand, M. & L. Orlóci. 1996. Complexity in plant communities: the notion and quantification. *Journal of Theoretical Biology* 179: 179-186.

Anderson, M.J. & C. ter Braak. 2003. Permutation tests for multi-factorial analysis of variance. *Journal of Statistical Computations and Simulations* 73:85-113.

Camiz, S., J.-J. Denimal, & V.D. Pillar. 2006. Hierarchical factor classification of variables in ecology. *Community Ecology* 7: 165-179.

Denimal, J.J. 2001. Hierarchical Factorial Analysis, Actes du 10th International Symposium on Applied Stochastic Models and Data Analysis. Compiègne, 12-15 Juin 2001.

Edgington, E. S. 1987. Randomization Tests. Marcel Dekker, New York.

Efron, B., and R. Tibshirani. 1993. An Introduction to the Bootstrap. Chapman & Hall, London.

Digby, P. G. N. & R. A. Kempton. 1987. *Multivariate Analysis of Ecological Communities*. London, Chapman & Hall. 206 p.

Feoli, E., M. Lagonegro & L. Orlóci. 1984. *Information Analysis of Vegetation Data***.** Junk, The Hague.

Gower, J. C. 1971. A general coefficient of similarity and some of its properties. *Biometrics* 27: 857-874

Grimm, E.C. 1987. CONISS: A FORTRAN 77 program for stratigraphically constrained cluster analysis by the method of incremental sum of squares. *Computers and Geosciences* 13: 13-35.

Kendall, M. & J. D. Gibbons. 1990. Rank Correlation Methods 5th ed. Edward Arnold, London.

Legendre, P. & L. Legendre. 1998. Numerical Ecology 2nd English Edition. Elsevier, Amsterdam. 853 p.

McArdle, B.H. & M.J. Anderson. 2001. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology* 82:290-297.

Manly, B. F. J. 1991. *Randomization and Monte Carlo Methods in Biology*. London, Chapman & Hall. 281 p.

Mielke, P. W., and K. J. Berry. 2001. Permutation Methods: A Distance Approach. Springer-Verlag, New York, USA.

Orlóci, L. 1967. An agglomerative method for classification of plant communities. *Journal of Ecology* 55: 193-205.

Orlóci, L. 1978. *Multivariate Analysis in Vegetation Research*. 2.ed. The Hague, W. Junk. 451 p.

Orlóci, L. & V. D. Pillar. 1991. On sample optimality in ecosystem survey. In: Feoli, E. & L. Orlóci (eds.) *Computer Assisted Vegetation Analysis*. p.41-46. Kluwer, Dordrecht. 498p.

Orlóci, L. & V. D. Pillar. 1989. On sample size optimality in ecosystem survey. *Biometrie-Praximetrie* 29: 173-184.

Orlóci, L. 1991. *Entropy and Information*. SPB Academic Publishing, The Hague.

Orlóci, L., N. C. Kenkel & M. Orlóci. 1987. *Data Analysis in Population and Community Ecology*. University of Hawaii, Honolulu / New Mexico State University, Las Cruces. 211p.

Pielou, E. C. 1984. *The Interpretation of Ecological Data; a Primer on Classification and Ordination*. New York, J. Wiley. 263 p.

Pillar, V. D. & L. Orlóci. 1993. *Character-Based Community Analysis; the Theory and an Application Program*. SPB Academic Publishing, The Hague.

Pillar, V. D. & L. Orlóci. 1996. On randomization testing in vegetation science: multifactor comparisons of relevé groups. *Journal of Vegetation Science* 7: 585-592.

Pillar, V. D. 1998. Sampling sufficiency in ecological surveys. *Abstracta Botanica* 22: 37-48.

Pillar, V. D. 1999a. How sharp are classifications? *Ecology* 80: 2508-2516.

Pillar, V. D. 1999b. The bootstrapped ordination reexamined. *Journal of Vegetation Science* 10: 895-902.

Pillar, V.D. 2006. How accurate and powerful are randomization tests in multivariate analysis of variance with ecological data? *Ecology* (submitted).

Podani, J. 2000. *Introduction to the Exploration of Multivariate Biological Data*. Leiden, Backuys Publishers. 407 p.

Ward, J.H. 1963. Hierarchical grouping to optimize and objective function. *J. Amer. Stat. Assoc*. 58: 236-244.