A USER GUIDE to MDA:
Morphospace-Disparity Analysis for Matlab, Version 1.2

Nicolas Navarro
UMR 5561-Biogéosciences, Centre des Sciences de la Terre, Université de Bourgogne, 6 bld Gabriel, F-21000 Dijon;
E-mail: nicolas.navarro@u-bourgogne.fr

This is MDA version 1.2. MDA is a program that runs under Matlab® with a simple text-user interface. It has been developed to perform disparity analyses on morphospace data. Four major modes of investigation are available, three of which are specifically designed for examining time series. Most morphospace occupation estimators found in the literature are incorporated in this version.

When used please cite author in the form:
Navarro, N. 2001. MDA: Morphospace-Disparity Analysis for Matlab, Version 1.2. Biogéosciences-Dijon Publisher.
Or cite the introductory paper of the program in Computers & Geosciences
Navarro, N. MDA: a Matlab-based program for Morphospace-Disparity Analysis. Computers & Geosciences

Installation, utilization and contents

1 Extract zip archive MDAvs1.2 to the directory C:\

=> Automatic directories are created: MDAvs1.2 and a help folder containing this user's guide and example files.

2 Add the MDAvs1.2 directory to the Matlab path.

The table of contents is then available in help windows. This table contains a description of the functions, descriptors and the input file format. Some of the information given in the table of contents is explained below.

Usage: **mda** in the matlab workspace.

Input data and user-defined parameters of analysis

Input file format: : (see Appendix 2)

Data must be bidimensional at least and must be tab-delimited. Input default filenames can be easily modified in MDA_prefs.m. Using the default filename saves user-choice time but files must be placed in the MDA directory. If there are no default files in the MDA directory, an input dialog box allows the required files to be opened in the directory chosen by the user.

Depending on the options chosen in the user-interface two, three or four input files are required. Files contain data only with no column or row headings (see MDAvs1.2/help/Ex_Input_files/*.txt). The first file contains the coordinates on PCs with rows equal to species and columns equal to PCs (default filename: data_matrix.txt). The second file contains a matrix of presence/absence by stratigraphic level (or by geographical area or whatever depending on the specific subject-matter) with rows equal to species and columns equal to levels (default filename: occurrence.txt). An optional third input file corresponds to another group file as the occurrence file allows multiple subdivision of the data file. The final input file contains PC eigenvalues (see MDA 1.2/specimen input files).

Disparity estimates available:

Most of the more recent or more widely used disparity estimates found in the literature have been implemented in this version: sum of univariates ranges and variances, N root of product of univariates ranges and variances, range as the maximal Euclidean pairwise distance, Area of convex hull, PCO volume, mean and median pairwise distance, mean distance to centroid, partial disparity metric (Foote, 1993), minimum and maximum of each component. Estimates for discrete characters have not been included (see Foote, 1999 and Ciampaglio et al. 2001 for some of these estimates). Foote (1999) used the convex hull volume to estimate overall occupation of morphospace (a range-based estimate). However, this metric is computer intensive. Because Matlab provides a simple and fast method for convex hull calculation in two dimensions, this metric has been introduced in MDA but is limited to the first two dimensions. Metrics based on Euclidean pairwise distance used the vectorized m-file (distance.m) of R. Bunschoten available from http://www.mathworks.com associated with the m-file (trilow.m) of R. E. Strauss which permits the extraction of the lower triangular portion of a squared matrix without the diagonal. This latter function is available from http://www.biol.ttu.edu/Strauss/Matlab/matlab.htm. This two codes are free and included in the main function MDA and so, did not required their download. These two files have been integrated because this method is more quick than the previous one used in earlier version of MDA, especially with large sample size.

Options available:

Rescaled variance on eigenvalues:

Depending on the data file, variances on axes may be rescaled using eigenvalues. This option is required with data derived from PCA with certain statistical package where all variances on axes are equal to unity. By rescaling variances on axes using the root square of their eigenvalues their

initial part of variance can be recovered. When the program is used with PCA from other program, PCO, correspondence analysis or partial warps, variances on axes are equal to their respective initial part of variance and so no rescaling is required.

Number of PCs retains:
Analysis can be performed with a user-limited number of axes. Some results (minimum and maximum on each component) are truncated if this number exceeds 10 axes. Because major events of occupational patterns induced major events of dispersion and since major components alone summarize the essential part of dispersion and so display the more interesting patterns (minoring noise), such truncation is not really a problem.

Analyses available:
Four main analyses are available. Three for examining time series but which can be used for examining geographical series, different taxonomic groups or other factors.

SGA

performs a single group analysis: data are subdivided by temporal level and disparity estimators on each subspace are estimated. This option requires two files (data and occurrence files).

MGA

runs in the same way but on more than one group such as geographical, taxonomic group. This option requires three files (data, occurrence and group files).

PDA

performs a partial disparity analysis (Foote, 1993). As with MGA, a third argument (group file) is required. In this case, overall disparity is partitioned into the different groups. This analysis corresponds to a decomposition of variance and so, uses a different metric from the previous two.

BTailTest

Observed disparity estimates of some groups are compared with the bootstrap distribution of disparity estimate of reference group (first column of group file except if Partial Disparity is selected, in which case, the reference group corresponds to the complete data set). This option requires two input files (data and group files). For reasons of memory limitations with the number of bootstrap resampling operations (random resampling with replacement [Efron and Tibshirani, 1993]), this option is limited to one disparity estimate (user's choice). Moreover, some estimates (range-based) are unavailable because their unbiased estimation requires a rarefaction procedure (see below). Three estimations of probabilities (corresponding to the two-tailed and two one-tailed tests) are computed by counting the number of bootstrapped values of the reference group exceeding the observed value of the analyzed group and then dividing by the number of bootstrap resampling operations.

Bootstrap and rarefaction:
- Bootstrap estimation:

Because the observed values of disparity metrics are sample-dependent, an estimator and the associated standard error is computed using bootstrapping (Efron and Tibshirani, 1993): data are resampled randomly with replacement, and the mean and standard deviation then calculated. Standard deviation then provides an estimate of standard error, which, in paleobiological studies, essentially reflects an estimate of analytical error (Foote 1993; Eble 2000). A confidence interval with a user-defined percentile can be also calculated (see below). The number of bootstrap resampling operations is chosen by the user. When using mean and standard deviation as estimators, 200 replicates may be a "good" minimum value (standard deviation tends to stabilize at this value but this is dependent on the data), this

number may be increased if CI is chosen. For the bootstrap tail test, the number of replicates must be increased to at least 1000.

- Rarefaction:

  Disparity estimators can be grouped into two categories: range-based and variance-based. Estimators from the first category and the two location parameters (minimum and maximum on each component) are sensitive to sample size (Foote, 1992). For correct sample size dependence, a rarefaction procedure is used (see Foote, 1992; Wills et al. 1994; Eble, 2000). For the majority, rarefaction is carried out by bootstrapping (Efron and Tibshirani, 1993). The bootstrapping procedure used for rarefaction yields the same statistical structure of error for all indices whether rarefied or not (Eble 2000). For PCO volume (Ciampaglio et al., 2001), rarefaction is performed using standardization by sample size following the initial procedure of Ciampaglio et al. (2001). The results are similar to those obtained using rarefaction by bootstrapping. It is easy to modify the code to use bootstrapping rarefaction instead of sample size standardization for the PCO volume. Users may use rarefaction or not and may choose between a user sample size for rarefaction or a predefined sample size corresponding to minimum observed sample size.

Confindence Interval:

A user-input choice allows a confidence interval to be calculated for each estimate. This choice is made in two stages: first it must be decided whether or not to use CI (CI requires a large number of replicates and so is time consuming); next the level of confidence interval may be chosen. This confidence interval is based on a non-parametric bootstrap resampling and uses the percentile method.

Output files and figures

Figures:

For SGA and MGA, two main figures are obtained. The first contains an error bar plot of the ten disparity estimates throughout the temporal level (on the x-axis) or others depending on the type of data in the occurrence file. Similarly, the second graph contains an error bar plot of the minimum and maximum on each axis (limited to 10). The error bar corresponds to ± one standard deviation if the CI option is chosen, or otherwise to CI. In MGA, both figures are replicated in line with the number of groups in the group file. For PDA, temporal series of metrics for each group are plotted. For the bootstrap tail test, a histogram of bootstrap replicates of the chosen disparity metric is displayed. Estimates of probabilities are displayed in workspace if the "save results" option is chosen (see below).

Output files: (see Appendix 2)

A file heading is associated with the output files corresponding to the analysis summary (also displayed in workspace) with the user-choices. Output files are formatted with a column heading. The first columns correspond to sample-size, rarefaction size, interval, and group (which is one when using SGA). For the bootstrap tail test, the first columns correspond to the group analyzed and the observed value of the disparity estimate chosen (see MDA 1.2/help/Ex_Output_files). An output dialog box is used for selecting the directory and filename. If the filename exists, the results are appended to the prior results.

MDA with one-dimensional data

The function MDA_1D contains SGA and BTailTest routines for one-dimensional or multiple one-dimensional data (max. 10). For SGA, mean, minimum, maximum, variance and range of each variable are given. The bootstrap tail test is performed on the variance of each variable.

References

Ciampaglio, C. N., M. Kemp, and D. W. McShea. 2001. Detecting changes in morphospace occupation patterns in the fossil record: characterization and analysis of measures of disparity. Paleobiology 27: 695-715.

Eble, G. J. 2000. Contrasting evolutionary flexibility in sister groups: Disparity and diversity in Mesozoic atelostomate echinoids. Paleobiology 26: 56-79.

Efron, B., and R. J. Tibshirani. 1993. An Introduction to the Bootstrap. Chapman and Hall, New York.

Foote, M. 1992. Rarefaction analysis of morphological and taxonomic diversity. Paleobiology 18: 1-16.

Foote, M. 1993. Contributions of individual taxa to overall morphological disparity. Paleobiology 19: 403-419.

Foote, M. 1999. Morphological diversity in the evolutionary radiation of Paleozoic and post-Paleozoic crinoids. Paleobiology 25: 1-115.

Wills, M. A., D. E. G. Briggs, and R. A. Fortey. 1994. Disparity as an evolutionary index. A comparison of Cambrian and Recent arthropods. Paleobiology 20: 93-110.

Appendix 1: Text User-Interface

```
PCs need rescaling to eigenvalues: y/n
Help:
If you use data from certain statistical package, all variances on axes can have been
scale to unity. In this case the rescaling option (using the root squared of the
eigenvalue)is used to obtain variance on PCs equal to their eigenvalue and so equal
to their initial proportion of variance

Rescaling on eigenvalue: X

number of PCs selected: X

Four analyses are available:

(1) SGA: unique group analysis
(2) MGA: several group analysis
(3) PDA: Partial Disparity Analysis
(4) BTailTest: Bootstrap Tail Test

choice: X

Rarefaction Menu:

  Help:
  Some estimates of morphospace occupation are sensitive to sample size:
    -Total range,
    -N Root of Product of ranges,
    -Range,
    -Area of convex hull,
    -Min & Max on PC
  Variance-based estimates are not sensitive to sample size and their analyses
  are not based on rarefied samples
  The rarefaction procedure is used to eliminated this link (Foote 1992 Paleobiology)
  The rarefaction is performed using bootstrapping (Eble 2000 Paleobiology).

(1) no rarefaction
(2) rarefaction with minimum observed sample size = X
(3) rarefaction with a user-choice sample size

choice:  X

sample size of rarefaction: X

Number of bootstrap resamplings: X

>>> studied sample n°1 on 14
for one sample, rarefaction size is upper to sample size:
analysis is performed with sample size !!

>>> studied sample n°2 on 14
>>> studied sample n°3 on 14
>>> studied sample n°4 on 14
...
Do you want a confidence interval (e.g. 95%) on the bootstrap y/n
case n: error bar = +/- 1 std
choice: X

input level of confidence interval (e.g., 95): X

==================================
Save numerical results?
-> data are appended if file name already exists
(y/n): X

Analysis Summary:
==================================
Analysis performed: X
Number of PCs selected: X
```

```
   PCs rescaled to eigenvalues: X
   Rarefaction: X: rarefaction size used: X
   Number of bootstrap resamplings: X
   Upper-lower values: X

   Start again analysis(y) or exit (x)?: X
```
If Bootstrap Tail Test is chosen:

```
   Bootstrap Tail Test Help:

   >> The first group in the group file is the one use for comparison
   This group is resampled with replacement (bootstrap) [except for one metric see
   below]
   and the observed values of other groups are compared with the distribution obtained.
   Because disparity estimates may increase or decrease, three cases arise:
     - one-tailed test with either upper tail or lower tail
     - two-tailed test.

   Range-based disparity estimates are sensitive to sample size, and unavailable here.

   1000 bootstrap resamplings is the minimum value for accuracy.

   Choosing one disparity estimate saves time, allowing analysis with a higher number of
   bootstrap resamplings.
   Disparity estimates available:
     (1) Sum of variances
     (2) N-root product of variances
     (3) Mean pairwise Euclidean distance
     (4) Median pairwise Euclidean distance
     (5) Mean Euclidean distance to group centroid
     (6) Partial Disparity (resampled group is the complete data set)

   choice: X

   Number of bootstrap resamplings: X
   number of resamplings is low: results may be inaccurate

   >>> bootstrap iteration n°1 on R
   >>> bootstrap iteration n°1/10*R on R
   ...
   >>> studied group n°1 on G
   ...
   ==================================
   Save numerical results?
   -> data are appended if file name already exists
   (y/n): X

   Analysis Summary:
   ==================================
   Analysis performed: BTailTest
   Number of PCs selected: X
   PCs rescaled to eigenvalues: X
   Rarefaction: none
   Number of bootstrap resamplings: X
   Upper-lower values: none
   Disparity metric used: X

   ----------------------------------
Group  Observed value      p two-tail   p upper one-tail    p lower one-tail

   Start again analysis(y) or exit (x)?: X
```

Appendix 2: Example of Input and Output Files

Input Files

data_Matrix.txt

variables	*1*	*j*	*P*
observations			
1	*-0.851*	*2.228*	*-2.424*
	-0.848	*7.017*	*-6.161*
i	*2.515*	*8.329*	*-6.562*
	-0.879	*11.004*	*-2.699*
	-2.793	*11.970*	*0.088*
N	*-2.173*	*14.089*	*-7.431*

occurrence.txt

levels	*1*	*j*	*L*
1	*0*	*0*	*0*
1	*0*	*0*	*0*
0	*1*	*1*	*1*
0	*1*	*1*	*0*
1	*1*	*1*	*1*
0	*0*	*1*	*1*

group.txt

group	*1*	*j*	*G*
1	*0*	*0*	
1	*0*	*0*	
0	*1*	*0*	
0	*1*	*0*	
0	*0*	*1*	
0	*0*	*1*	

eigenvalues.txt

eigenv.	*1*	*j*	*P*
	2.889	*1.768*	*0.146*

Output Files

A

```
Analysis Summary:
==================================
Analysis performs: SGA
Number of PCs retained: 2
Rescaling PCs on eigenvalues: n
Rarefaction: y: rarefaction size used: 4
Number of bootstrap resampling: 200
Upper-lower values: +/- 1 stdev
```

Sample Size	Rarefaction Size	Interval	Group	Sum of Ranges	Stdev SR	upper value SR	lower value SR
2	2	1	1	1.48	1.54	3.03	-0.06
4	4	2	1	9.35	2.66	12.01	6.68
13	4	3	1	11.83	4.80	16.64	7.03
10	4	4	1	15.53	5.54	21.08	9.99
13	4	5	1	22.06	7.41	29.47	14.65
14	4	6	1	22.18	8.91	31.09	13.27

B

```
Analysis Summary:
==================================
Analysis performs: BTailTest
Number of PCs retained: 2
Rescaling PCs on eigenvalues: n
Rarefaction: none
Number of bootstrap resampling: 10000
Upper-lower values: none
Disparity metric used: Partial Disparity
```

Group	Observed	p two-tail	p upper one-tail	p lower one-tail
1	327.09	0.9295	0.4649	0.5351
2	288.09	0.0178	0.9911	0.0089
3	199.28	0.000000	1	0.000000