

PERMDISP

Permutational analysis of multivariate dispersions

A computer program
by Marti J. Anderson



Centre for Research on
Ecological Impacts of Coastal Cities
University of Sydney (1999)



Department of Statistics
University of Auckland
(2004)

DISCLAIMER

This FORTRAN program is provided without any explicit or implicit warranty of correct functioning. It has been developed as part of a university-based research program. If, however, you should encounter problems with this program, the author will be happy to help solve them. Researchers can use this program for scientific and research purposes, but intellectual property and copyright for the source code and program remains the property of Marti Jane Anderson. The program relies on several canned routines in FORTRAN from Numerical Recipes (Press et al. 1992). Namely, for finding the eigenvalues of a real symmetric distance matrix, using a householder reduction, I used the routines “tred2” and “tqli.” Also, for sorting, I used a variation of the routine “piksr2.” To determine the number of possible permutations, I used the gamma function for calculating factorials (“gammln”), the function for calculating the logarithm of a factorial (“factln”), and a variation on the function for a binomial coefficient (“bico”).

Users of the program may refer to the present user’s guide as follows:

Anderson, M.J. (2004). PERMDISP: a FORTRAN computer program for permutatinoal analysis of multivariate dispersions (for any two-factor ANOVA design) using permutation tests. Department of Statistics, University of Auckland, New Zealand.

Author’s contact details:

Dr Marti J. Anderson
Department of Statistics
University of Auckland
Private Bag 92019
Auckland, New Zealand
Tel: 64-9-373-7599 ext 85052
Fax: 64-9-373-7018
Email: mja@stat.auckland.ac.nz
Website: <http://www.stat.auckland.ac.nz/~mja>

I. Introduction

PERMDISP is a computer program for comparing the multivariate dispersions among groups on the basis of any distance or dissimilarity measure of choice. More specifically, the test can be considered in two steps: (1) calculation of the distances from observations to their centroids and (2) comparison of the average of these distances among groups, using ANOVA. A P -value is then obtained using permutation of the observations. The approach is a multivariate analogue to Levene's test (Levene 1960). The analysis of these distances to centroids can be done for any two-factor design, just as in PERMANOVA. It is well-known that the test for differences in location among groups in multivariate space (such as PERMANOVA) is sensitive to differences in dispersion among the groups. Thus, rejection of the null hypothesis for PERMANOVA suggests that groups may differ because of their location, their relative dispersion, or both. The program is designed to be used as a companion to PERMANOVA, to assist in unraveling the possible reasons for rejection of the null hypothesis. It is also useful in its own right, however, to investigate differences in dispersions alone when hypotheses of this nature arise. For example, it has been suggested that variability in the structure of ecological assemblages may increase (Warwick and Clarke 1993) or decrease (e.g. Chapman et al. 1995) as a consequence of environmental stress or impact.

II. Description of the Test Statistic

Levene (1960) proposed doing an analysis of variance (ANOVA) on the absolute values of deviations of observations from their group mean. For Levene's test, let x_{ij} be the observation on the univariate response variable for the j th observation ($j = 1, \dots, n$) in the i th group ($i = 1, \dots, g$). One then does a traditional ANOVA comparing the values of $z_{ij} = |x_{ij} - \bar{x}_i|$ among the g groups, where $\bar{x}_i = \sum_{j=1}^n x_{ij}$. A multivariate analogue was described by Van Valen (1978) and given in Manly (1994), based on distances. As such, it opened up possibilities for ecological analysis based on distances or dissimilarities. Here, one does a traditional ANOVA on the Euclidean distances from individual observation vectors to their group centroid (Fig. 1). Thus, let \mathbf{x}_{ij} be a multivariate response vector for the j th observation ($j = 1, \dots, n$) in the i th group ($i = 1, \dots, g$), consisting of the responses x_{ijk} on each of $k = 1, \dots, p$ variables (e.g. in an ecological context, the vectors may be counts of abundances for each of p species or taxa). Calculate Euclidean distances as:

$$z_{ij} = \|\mathbf{x}_{ij} - \bar{\mathbf{x}}_i\| = \sqrt{\sum_{k=1}^p (x_{ijk} - \bar{x}_{ik})^2} \quad (1)$$

where $\bar{\mathbf{x}}_i$ is a vector of means $\bar{x}_{ik} = \sum_{j=1}^n x_{ijk}$ for each of the $k = 1, \dots, p$ variables for the i th group.

Hereafter, we shall use the notation $\|\mathbf{x}_1 - \mathbf{x}_2\|$ to denote the Euclidean distance between two specified vectors \mathbf{x}_1 and \mathbf{x}_2 . One then does a univariate ANOVA comparing the z_{ij} 's among the g groups. If two or more groups differ in their relative dispersions, then these distances (z_{ij} 's) will be different, on average, among the groups (e.g. Fig. 1).

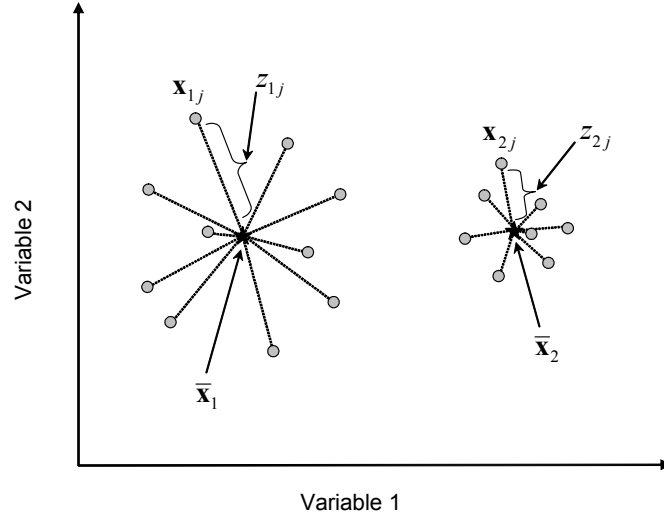


Fig. 1. Example of a multivariate systems with 2 variables (2 dimensions) and 10 observations in each of 2 groups. The values of the z 's will, on average, differ significantly if the dispersion of points in multivariate space differs for the two groups.

Now, rather than take the z 's and do a traditional univariate ANOVA on them, the program PERMDISP does a permutational ANOVA (Anderson 2001). Thus, the same F -statistics are calculated, but P -values are obtained by permutation. In this fashion, we make no particular assumptions about the nature of the distribution of the original variables, nor of the z 's. In other words, consider that after PERMDISP has calculated the values for the z 's (one for each observation), it inputs these into the PERMANOVA procedure to test the null hypothesis of no difference in the average value of the z 's among the groups (see the PERMANOVA user notes for further details about this procedure).

One special feature of the PERMDISP method is that it does not need to be based on Euclidean distances, as shown in Fig. 1. In fact, the z 's can be based on any dissimilarity measure of choice. If the measure chosen is not Euclidean, then the centroid in the space of observation points for a particular group is no longer going to be simply the arithmetic average of the original observations using the original variables. Instead, one must first calculate principal coordinates from the distance matrix chosen, then calculate appropriate centroids and z values from these (e.g. see McArdle and Anderson 2001). This and the use of permutations are two important ways in which the method used here differs from that described by Underwood and Chapman (1998).

III. Notes

Note 1: *If you have 2 observations per group*, then the distance to the centroid (i.e. the value of z) will be the same for the two observations within the same group. Thus, there will be no within-group variance when the sample size is only 2. For this reason, the program will not calculate the tests of dispersion for sample sizes less than $n = 3$.

Note 2: *If you have a nested hierarchical design* (i.e. factor B nested within factor A), there are two approaches to the test of dispersions. The first is using the individual observations as the units of dispersion. The second is in the test of factor A, one might also use the centroids of the levels of factor B within A as the units of dispersion. For the nested design, the program PERMDISP will ask whether you wish to calculate dispersion using the second approach after it does the calculations using the first approach.

Note 3: *Your choice of transformation will radically affect your results.* This is a well-known fact for univariate analysis, but is often forgotten in the context of multivariate analysis. Recall that many workers

use transformations to “fix” the problem of heterogeneity when dealing with univariate data sets. Be aware that when you transform data (using square roots, fourth roots, logs, etc.) you are also affecting the relative spread of the observations.

Note 4: *Your choice of distance or dissimilarity measure will radically affect your results.* Many dissimilarity measures have intrinsic row (observation) standardizations in them, such as Bray-Curtis, Chi-squared, CY dissimilarity, etc. As such, these change the relative dispersions within groups, especially for observations that have species with large abundances. Distance measures such as Euclidean, Manhattan and Gower’s measure do not have such intrinsic row standardizations.

III. Running the Program

A. Parameters of the Program

The parameters of PERMDISP are currently set as follows:

maximum no. of levels of either factor (a or b) = 20

maximum sample size (no. of replicates, n) = 50

maximum no. of variables = 500

maximum total no. of observations or samples ($N = a \times b \times n$) = 500

If the program will not run on your machine, it could be that the parameter settings are too large for your computer’s memory, or your data matrix or design exceeds the parameter settings.

B. Format and Structure of the Input Data File

One can either input a file containing raw data, or input a file containing a symmetric matrix of distances or dissimilarities. In each case, the file must be saved in tab delimited (ASCII text, *.txt) or comma delimited (*.csv) format. If you are using a Macintosh, then save the file (for example, from Excel) as “Text (Windows)”. If you save it as a simple text file from Excel, it won’t run unless you open up the file again (for example, using a text editor such as BBEdit) and manually add a carriage return (“Enter”) to the end of the last line. This is a weird quirk of the way the Mac version of Excel saves text files that, for some reason unknown to me, causes my program to crash. You should not experience any such problem, however, if the file is saved as “Text (Windows)” or if the file was created using a different text editor.

The option of entering a symmetric distance matrix means that the user can base the analysis on any distance measure of choice. If none of the choices of distance measures offered in the PERMDISP program are suitable, then one can use a different distance function in another program and input this distance or dissimilarity matrix directly.

A file containing raw data can be organised in one of two ways. Either (i) observations are in rows and variables (species) are in columns or (ii) variables are in rows and observations are in columns. The organisation of the samples should follow the experimental design, with the first partitioning of the data according to factor 1, the second partitioning according to factor 2, etc. For example, let’s say the data are organised so that columns are variables and rows are samples. In the case of two species variables in a 2-factor crossed (orthogonal) experimental design, where each factor has 2 levels and the sample size is $n = 3$, one would have two columns of numbers (corresponding to the two species) with the following organisation:

Factor 1	Factor 2	Sp. 1	Sp. 2	
Level 1	Level 1	5	4	
		4	1	
		1	1	
	Level 2	Level 2	0	2
			3	6
			7	2
Level 2	Level 1	2	6	
		7	3	
		5	7	
	Level 2	Level 2	8	8
			3	9
			0	1

Data file
(no headers
on columns
or rows)

The matrix should be saved (preferably as an ASCII text file, commonly using the extension *.txt), with no headers on either the columns or the rows. This is the input file for the program. To avoid dealing with long file names and paths to locate files, place the file in the same location on your computer (i.e. the same directory) as the PERMDISP.exe (or PERMDISP.mac) file, for use with the program. Double-click on the “PERMDISP.exe” (or “PERMDISP.mac”) file to run the program.

C. Output File

The output file will contain the details of your experimental design, the choices you have made (in terms of the distance measure, any transformation or standardization you have used, the method and number of permutations chosen, etc.) and the ANOVA table of results. The *F*-ratio and associated *P*-value for each term in the analysis can be interpreted in the same way that one would interpret the result of a univariate analysis of variance, but it is the hypothesis of “no difference in multivariate dispersions among groups” for each factor that is being tested.

It is a good idea to name the output file something with the extension *.txt on the end. This means that when one is working in Windows, double-clicking on the output file brings it up automatically in Notepad (for example) for easy examination and printing, if desired.

D. Questions Asked by the Program

The questions asked by the program are best demonstrated by an example. The data analysed here are from an experiment concerning the effect of shade and the effect of proximity to the seafloor on assemblages of subtidal invertebrates and algae on hard surfaces near marinas (courtesy of Dr Tim Glasby). For further details, see Glasby (1999).

The experiment was a two-way crossed (orthogonal) design with $n = 4$ having the following structure:
 Factor 1 = Position (fixed, 2 levels: far or near to the seafloor)
 Factor 2 = Shade (fixed, 3 levels: shade, a procedural control in the form of a clear plexiglass shade, and no shade)

Organisms colonising subtidal 15cm x 15cm sandstone settlement plates were counted and a total of 46 taxa were included in analyses. (Organisms that occurred less than twice were not included). The data matrix (24 rows (samples) x 46 columns (taxa)) is contained in an ASCII text file called “Tim.txt.”

The questions asked by the program are given in Courier font, while responses given for this data set are in *Times bold italics*. Note that you can stop the program at any time by typing Ctrl-C (as for any DOS program) or, if you are a Mac user and this does not work, then try typing a period (a full stop) while holding down the apple key.

Type the name of the input file containing your data

Tim.txt

Type a name for the output file of results (*.txt)

Results.txt

Note: If the program crashes it may be because you already have a file in the directory with the same name as what you are typing here for the output file. If so, then you need to delete the previous file or type a new name. Unfortunately, the program is not smart enough to write over the top of an existing file.

Nature of the data in the input file:

- 1) raw data (n x p)
- 2) distance matrix (n x n)

1

Structure of the input file:

- 1) rows are samples and columns are variables
- 2) columns are samples and rows are variables

1

How many variables (columns) are there?

46

Choice of transformation:

- 1) none
- 2) square-root
- 3) fourth-root
- 4) ln(x)
- 5) ln(x+1)
- 6) log10(x)
- 7) log10(x+1)
- 5) presence/absence

3

Choice of standardization:

- 1) none
- 2) standardize by row (sample) sums
- 3) standardize by column (variable) sums
- 4) double standardize by row and column sums
- 5) standardize each variable to z-scores (normalize)
- 6) standardize each variable by dividing by its s.d.
- 7) standardize each variable by dividing by its range

1

Choice of distance measure:

- 1) Bray-Curtis dissimilarity
- 2) square root of Bray-Curtis
- 3) Euclidean distance
- 4) Orloci's Chord distance
- 5) Chi-square metric
- 6) Chi-square distance
- 7) Hellinger distance
- 8) Gower dissimilarity
- 9) Gower, excluding double zeros
- 10) Canberra distance
- 11) square root of Canberra distance
- 12) CY dissimilarity
- 13) Deviance based on the binomial
- 14) Deviance per observation (scale invariant)
- 15) Kulczynski dissimilarity
- 16) McArdle-Gower dissimilarity
- 17) Manhattan (city block) distance
- 18) Jaccard dissimilarity

1

Note: The same options were used here for the analysis as were chosen for the PERMANOVA of these data (see the PERMANOVA user notes). This was done so that the null hypothesis of no differences in

dispersion could be examined for these data as a companion to the PERMANOVA test for location differences that might be sensitive to this issue.

Do you want to output the distance matrix?

- 1) Yes
- 2) No

2

This is usually not necessary unless you want to see particular distance values.

ANOVA Experimental design:

- 1) One-way
- 2) Two-way nested
- 3) Two-way crossed (i.e. factorial or orthogonal)

3

Experimental design of two-way crossed analysis:

- 1) Fixed effects - both factors are fixed
- 2) Random effects - both factors are random
- 3) Mixed model - factor 1 is fixed, 2 is random
- 4) Mixed model - factor 1 is random, 2 is fixed

1

What is the name of factor 1?

Position

Type the number of levels for factor 1

2

What is the name of factor 2?

Shade

Type the number of levels for factor 2

3

What is the number of replicates?

4

How many permutations do you want for the tests?
(i.e. 99, 499, 999, 4999, etc.)

9999

Which general method of permutation do you want?

- 1) Permutation of raw data
- 2) Permutation of residuals (reduced model)
- 3) Permutation of residuals (full model)

1

Type an integer to be used as the seed
for the random permutations

5

The program will then print the results of the permutational tests of dispersions. This is written simultaneously to the screen and to the output file. In this case, the output looks like this:

--- Results ---

Permutational Test of Multivariate Dispersion:
tests for heterogeneity in the average dissimilarities
of points from the central location of their group.

Source	df	SS	MS	F	P	Possible No.perm.	Denom. MS
Posit	1	0.7265	0.7265	0.0777	0.7870	0.135E+07	Res
Shade	2	8.6154	4.3077	0.4609	0.6377	0.158E+10	Res

PoxSh	2	28.3288	14.1644	1.5154	0.2532	>1.0E+10	Res
Resid	18	168.2509	9.3473				
Total	23	205.9217					

Data were transformed to fourth root
 No standardization
 Analysis based on Bray-Curtis dissimilarities
 Permutation of raw data
 No. of permutations used = 9999

Once you have investigated the results, you may wish to analyse particular terms further by *a posteriori* tests. In the analysis of Tim's data, there were no significant differences in multivariate dispersion among the groups. Thus, there is no need to investigate pair-wise comparisons. The next question of the program is:

Would you like pair-wise *a posteriori* tests?
 (results are printed directly to the output file)
 1) Yes
 2) No
2

These results have been sent to the output file.
 End of the program.
 Press q to quit.
q

After completing PERMDISP, open the output file to see all of the results, including any *a posteriori* comparisons you may have done. The output file also contains the details of the experimental design. In the case of Tim's data, the file "Results.txt" looks like this:

```
Program PERMDISP
by M.J. Anderson
Department of Statistics, University of Auckland (2004)
```

Input file:Tim.txt

```
--- Experimental Design ---
Two-way crossed (orthogonal) ANOVA
Factor 1 is Position with 2 levels
Factor 2 is Shade with 3 levels
Factors 1 and 2 are fixed
The no. of replicates = 4
The no. of variables = 46
```

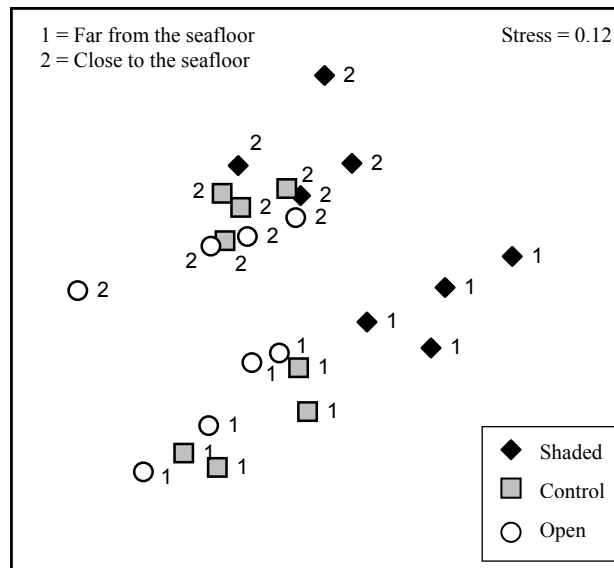
```
--- Results ---
Permutational Test of Multivariate Dispersion:
tests for heterogeneity in the average dissimilarities
of points from the central location of their group.
```

Source	df	SS	MS	F	P	Possible No.perm.	Denom. MS
Posit	1	0.7265	0.7265	0.0777	0.7870	0.135E+07	Res
Shade	2	8.6154	4.3077	0.4609	0.6377	0.158E+10	Res
PoxSh	2	28.3288	14.1644	1.5154	0.2532	>1.0E+10	Res
Resid	18	168.2509	9.3473				
Total	23	205.9217					

Data were transformed to fourth root
 No standardization
 Analysis based on Bray-Curtis dissimilarities
 Permutation of raw data
 No. of permutations used = 9999

The interpretation of the analysis is that there were no significant differences in multivariate dispersions for either of the factors: Position or Shade. The PERMANOVA (see the notes for the PERMANOVA program) did, however, find significant Position and Shade effects for these data on the basis of Bray-Curtis dissimilarities calculated on fourth-root transformed abundances. Thus, the effect of different positions and the effect of different shading treatments was to cause a shift in the assemblage structure (i.e. a location effect), not to make the assemblages either more or less variable.

These results are supported by a visual assessment of patterns in a non-metric MDS plot of double-root transformed data using Bray-Curtis distances, as shown below:



IV. References

- Anderson, M.J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology* 26: 32-46.
- Chapman, M.G., Underwood, A.J. and Skilleter, G.A. (1995). Variability at different spatial scales between a subtidal assemblage exposed to the discharge of sewage and two control locations. *Journal of Experimental Marine Biology and Ecology* 189: 103-122.
- Glasby, T.M. (1999). Interactive effects of shading and proximity to the seafloor on the development of subtidal epibiotic assemblages. *Marine Ecology Progress Series* 190: 113-124.
- Levene, H. (1960). Robust tests for equality of variances. Pages 278-292 in I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, and H. B. Mann, editors. Contributions to probability and statistics. Stanford University Press, Stanford, California.
- Manly, B.F.J. (1994). Multivariate statistical methods: a primer, 2nd edition. Chapman and Hall/CRC, Boca Raton, Florida.
- McArdle, B.H. and Anderson, M.J. (2001). Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology* 82(1): 290-297.
- Press, W.H., Teuklosky, S.A., Vetterling, W.T. and Flannery, B.P. (1992). *Numerical Recipes in FORTRAN, 2nd edition*. Cambridge University Press, Cambridge.
- Underwood, A.J. and Chapman, M.G. (1998). A method for analysing spatial scales of variation in composition of assemblages. *Oecologia* 117: 570-578.
- Van Valen, L. (1978). The statistics of variation. *Evolutionary Theory* 4: 33-43 (Erratum *Evolutionary Theory* 4: 202).

Warwick, R.M. and Clarke, K.R. (1993). Increased variability as a symptom of stress in marine communities. *Journal of Experimental Marine Biology and Ecology* 172: 215-226.