

CAP

Canonical Analysis of Principal coordinates

A computer program
by Marti J. Anderson



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. This program relies on several canned FORTRAN routines from other sources, including:

- (a) routines “tred2” and “tqli” from Numerical Recipes (Press et al. 1992) are used to finding the eigenvalues of a real symmetric distance matrix, using householder reduction;
- (b) routine “gaussj” from Numerical Recipes (Press et al. 1992) is used to find the solution to linear equations (i.e. the inverse of a matrix) by Gauss-Jordan elimination;
- (c) routine “cholde” from Numerical Recipes (Press et al. 1992) is used to find the Cholesky decomposition of a positive-definite symmetric matrix;
- (d) routines “svd” and “pythag” from a translation of the Algol procedure (Wilkinson and Reinsch 1971) were used to find the singular value decomposition of a matrix.

Research publications that use this method should cite the following papers.

Anderson, M.J. and Robinson, J. 2003. Generalised discriminant analysis based on distances. *Australian and New Zealand Journal of Statistics* 45(3): 301-318.

Anderson, M.J. and Willis, T.J. 2003. Canonical analysis of principal coordinates: a useful method of constrained ordination for ecology. *Ecology* 84: 511-525.

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

Anderson, M.J. 2004. CAP: a FORTRAN computer program for canonical analysis of principal coordinates. Department of Statistics, University of Auckland, New Zealand.

Author’s contact details:

Dr Marti J. Anderson
 Department of Statistics
 Tamaki Campus
 University of Auckland
 Private Bag 92019
 Auckland, New Zealand
 Tel: 64-9-373-7599 ext 85052
 Fax: 64-9-373-7000
 Email: mja@stat.auckland.ac.nz
 Website: <http://www.stat.auckland.ac.nz/~mja>

I. Description

CAP is a computer program that calculates a canonical analysis on the principal coordinates based on any symmetric distance matrix, including a test by permutation, as described by Anderson and Willis (2003) and Anderson and Robinson (2003). Consider an $(N \times p)$ matrix of response variables \mathbf{Y} (N = the total number of observational units and p = the number of variables). Consider also an $(N \times q)$ matrix, \mathbf{X} , which is of interest for a multivariate hypothesis. The purpose of CAP is to perform a canonical analysis for the effect of \mathbf{X} , if any, on \mathbf{Y} on the basis of any distance measure of choice, using permutations of the observations. Note that \mathbf{X} may contain the codes of an ANOVA model (a design matrix), yielding a generalised discriminant analysis, or it may contain one or more explanatory (predictor) variables of interest (e.g. environmental variables), yielding a generalised canonical correlation analysis.

To do the analysis, the first step is to let $\mathbf{D} = (d_{ij})$ be an $(N \times N)$ distance matrix calculated from observation units of \mathbf{Y} , using some chosen appropriate distance measure. Let $\mathbf{A} = (a_{ij}) = (-\frac{1}{2}d_{ij}^2)$, then calculate Gower's (1966) centered matrix (\mathbf{G}) by centering the elements of \mathbf{A} , i.e.

$$\mathbf{G} = \left(\mathbf{I} - \frac{1}{N}\mathbf{1}\mathbf{1}'\right)\mathbf{A}\left(\mathbf{I} - \frac{1}{N}\mathbf{1}\mathbf{1}'\right)$$

where $\mathbf{1}$ is a column of 1's of length N and \mathbf{I} is an $(N \times N)$ identity matrix. Matrix \mathbf{G} is then decomposed into its component eigenvalues and corresponding orthonormal eigenvectors \mathbf{Q} . We then choose a subset of these eigenvectors (say m of them), as a matrix for the ensuing canonical analysis that we will call \mathbf{Q}_m .

The general idea is to include as much of the relevant information in matrix \mathbf{Q} (and thus \mathbf{G}) as is reasonable for further analysis. It is important to keep m relatively small compared to N , the total number of observations. The user can either choose m manually or allow the computer program to choose m on a non-arbitrary basis using diagnostic information on the appropriate dimensionality for the canonical analysis.

Next, we calculate the "hat" or projection matrix $\mathbf{H} = \mathbf{X}[\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'$ (e.g. Johnson & Wichern 1992). Then, the canonical test statistic is the trace:

$$tr(\mathbf{Q}_m'\mathbf{H}\mathbf{Q}_m).$$

The eigenvalues of matrix $\mathbf{Q}^* = \mathbf{Q}_m'\mathbf{H}\mathbf{Q}_m$ are the squared canonical correlations $\delta_1^2, \dots, \delta_s^2$ (where $s = \min(m, q)$). In addition to the trace statistic, the program calculates the greatest root statistic δ_1^2 . A P -value for each of the trace and greatest root statistics is then obtained by recalculating each of them for a large number of random re-orderings of the observations (i.e. the rows and columns of \mathbf{Q}_m), while keeping \mathbf{X} (and \mathbf{H}) constant. There will be s canonical axes, for plotting, which are given by \mathbf{Q}^* , standardized by the appropriate canonical correlation, δ_i , $i = 1, \dots, s$.

Output from the program includes:

- (a) Eigenvalues and eigenvectors from the principal coordinate analysis. The latter are the PCO axes that can be used to plot an unconstrained (metric MDS) of the data.
- (b) Canonical correlations and squared canonical correlations
- (c) Canonical axes scores (position of multivariate points on the canonical axes to be used for plotting).
- (d) Correlations of each of the original variables with each of the canonical axes.
- (e) Correlations of each X variable with each of the canonical axes (if a canonical correlation is done).
- (f) Diagnostics used to determine the appropriate value for the choice of m . The criterion used is either the value of m resulting in the minimum misclassification error (in the case of groups) or the minimum residual sum of squares (in the case of \mathbf{X} containing one or more quantitative variables). Also, m must not exceed p or N and is chosen so that the proportion of the variability explained by the first m PCO axes is more than 60% and less than 100% of the total variability in the original dissimilarity matrix.

- (g) In the case of groups, a table of results for the “leave-one-out” classification of individual observations to groups is given, along with the misclassification error for the choice of m used.
- (h) If requested, the results of a permutation test using the two different test statistics, (trace and largest root).

II. What is the difference between CAP and PERMANOVA?

An important point is that the analysis described here takes into account the correlation structure among the variables. Although the test statistic described for the programs PERMANOVA (Anderson 2004, Anderson 2001) or DISTLM (McArdle and Anderson 2001) test a similar multivariate hypothesis for a linear model, they do not take into account the correlation structure among the variables. CAP does this by essentially following the more traditional multivariate discriminant (or canonical) analysis, but does this on the principal coordinates from the distance matrix. One might expect to get similar results using CAP and the permutation test given by PERMANOVA. Both have exact type I error for the same multivariate null hypothesis of no differences among *a priori* groups. However, there are some situations in which either CAP or PERMANOVA will tend to be more powerful than the other, just like the situation in traditional MANOVA, where different test statistics (e.g., Roy’s greatest root, Wilks’ lambda, Pillai’s trace) have different power for different alternative hypotheses.

For example, in ecology, there are often situations where there are several highly abundant species that do not change across treatment groups and are strongly correlated with one another. However, there may be less abundant or patchy species that are not correlated with the abundant species, and do indeed differ significantly across treatment groups being considered. In this case, CAP will find these real and significant differences among the assemblages more often than PERMANOVA (Anderson and Robinson 2003). On the other hand, PERMANOVA (or DISTLM, which is equivalent) may provide a more powerful test than the CAP analysis if the original variables are *not* strongly correlated with one another (Anderson and Robinson, unpublished results).

CAP provides a constrained ordination diagram, which is not output by these other routines¹. In situations such as that described above, an unconstrained ordination (such as metric or non-metric MDS or principal coordinate analysis) will not show any clear separation of groups, even though group differences do occur. This is because the group differences occur along another dimension. This dimension will be drawn by the canonical plot (output by CAP). In fact, the canonical analysis finds the axis (or axes) in the principal coordinate space that is best at discriminating among the *a priori* groups.

A potential disadvantage of using CAP is that it will ignore some of the variability contained in matrix **G**. This will probably be irrelevant noise, but the choice of m is crucial to the analysis. Also, m must not be too large, or the canonical analysis may give results that are not meaningful (see Anderson and Willis 2003 for more details). No such decision needs to be made when using PERMANOVA or DISTLM, as either of these approaches will simply use all of the information in the distance matrix.

III. Input file(s)

The program allows the user to input a distance matrix **D** directly or a raw multivariate data matrix of response variables **Y**. In either case, the file should be saved as ASCII *.txt, with no column or row headings. 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.

¹ If the eigenvectors from a decomposition of the **HGH** matrix were output by either DISTLM or PERMANOVA, then this would be an ordination called a distance-based redundancy analysis (dbRDA).

This is the first input file for the program. If a file with a raw data matrix is input, either the rows or the columns may correspond to the variables for the analysis (the user will be given the option to choose one of these). Then, the user has several options for transformation, standardization and choice of distance measure.

If the hypothesis of interest involves differences among groups, then the user should choose “Discriminant Analysis” as the type of analysis to be done. The program will then ask the user relevant questions regarding the number of groups and the number of observations per group (see section IV below). Thus, no other data files are needed.

On the other hand, if the hypothesis of interest is the relationship of the species data with some other quantitative variables (e.g. environmental data), then the user should choose “Canonical Correlation Analysis” and follow the relevant instructions. In this case, another input file is required, consisting of an \mathbf{X} matrix containing one or more variables of interest for canonical analysis. For the \mathbf{X} input matrix, variables must be columns and observation units must be rows. The file must be saved in ASCII *.txt format with no headers or labels of any kind. The user will specify the number of columns, while the number of rows must be equal to the number of observational units already specified for \mathbf{Y} (or \mathbf{D}).

To avoid dealing with long file names and paths to locate files, place the relevant input file(s) in the same location on your computer (i.e. the same directory) as the CAP.exe (or CAP.mac) file, for use with the program. Double-click on the “CAP.exe” (or “CAP.mac”) file to run the program.

The program uses dynamic memory allocation, and so (theoretically) does not have any limits on the sizes of matrices (numbers of rows or columns) that may be used for the input files. If you are using a Macintosh and you get an error that reads:

```
BUFFER allocation failed
REWIND(UNIT=*,...
```

then you need to increase the memory allocated to the program. To do this, click on the program’s icon and type “i” while holding the apple key, then choose “Memory” (or choose “Get Info > Memory” from the File menu). Depending on the size of your matrices, you might even have to increase this value to something obnoxious, like 50000 or so. However, if you have a large input file and cannot seem to get the program to work even after doing this, then please contact the author.

IV. Questions Asked by the Program

The questions asked by the program are best demonstrated by an example. The data analysed here are from an underwater visual census of fishes at three different times at the Poor Knights Islands, New Zealand (courtesy of Trevor Willis and Chris Denny). There were 47 fish species variables recorded as divers swam along 25m long transects. Counts from nine such transects at the location were pooled to constitute a single observational unit. Data were collected in September 1998 ($n_1 = 15$), March 1999 ($n_2 = 21$) and September 1999 ($n_3 = 20$). Thus, $N = 56$ and $p = 47$. The hypothesis of interest (examined below) was to test for a significant difference in the assemblages of fishes recorded at the Poor Knights at each of the three times. For more details, see Willis and Denny (2000), Anderson and Robinson (2003) and Anderson and Willis (2003).

The response variable matrix \mathbf{Y} (56 rows x 47 columns) was contained in an ASCII text file called “PK.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.

CAP

Canonical Analysis of Principal coordinates

A program for
generalised discriminant analysis or
canonical correlation analysis
on the basis of any distance measure

by M.J. Anderson
Department of Statistics
University of Auckland (2004)

Type the name of the input file containing your data
(e.g. species variables or a distance matrix).

PK.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?

47

How many observations (rows) are there?

56

Choice of transformation for Y data:

- 1) none
- 2) square-root
- 3) fourth-root
- 4) $\ln(x)$
- 5) $\ln(x+1)$
- 6) $\log_{10}(x)$
- 7) $\log_{10}(x+1)$
- 8) presence/absence

5

Choice of standardisation:

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

Choose one of the following types of analysis:

- 1) Discriminant Analysis
(i.e. find axes that maximise separation among groups)
- 2) Canonical Correlation Analysis
(i.e. find axes that maximise correlation with a set of quantitative variables in a second matrix, X)

1

How many groups are there?

3

Type the sample sizes per group, separated by spaces.

15 21 20

Choose whether you wish to:

- 1) choose m manually
 - 2) let the computer program choose m
- Note: m is the number of PCO axes to be used in the canonical analysis.

2

Note: not knowing how many axes would be appropriate *a priori*, it is usually the case that one will let the computer program choose m , at least in the first instance.

Do you wish to output the principal coordinate (PCO) axes?
(i.e. these are the unconstrained metric MDS axes)

- 1) yes
- 2) no

1

How many PCO axes do you want in the output?

That is, enter the no. of dimensions you wish to plot, such as 2 or 3.

Type 999 if you want all of them up to and including m

2

Do you wish to do a test by permutation?

- 1) yes
- 2) no

1

Type the number of random permutations for the test.
(e.g. 99, 499, 999, 4999, 9999, etc.)

9999

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

12

Note: Any random integer seed will do here. If you wish to re-calculate the exact same permutation test at a later date (i.e. with the exact same set of re-orderings), make a note of this seed and use it again.

Please wait while I do some diagnostic calculations...

```
Choice of m          = 1
Choice of m          = 2
Choice of m          = 3
...
```

Note: Here, the program sequentially chooses increasing values for m and gets the diagnostic information (proportion of variability explained by the first m PCO axes, residual sum of squares, squared canonical correlations, and misclassification error in the case of groups) all to be output later. If you manually choose an integer value for m yourself, then you will also have the option of skipping these diagnostic calculations or doing them up to your choice for m . It is advised that you choose to do the diagnostic calculations in any event, as they are useful for making a decision about the appropriate dimensionality of the problem.

Now doing the canonical analysis...

Please wait while I do the permutation test...

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

q

When the program has completed the analysis, it will print the results to the named output file. In this case, the output file looks like this:

```

                                CAP
-----
Canonical Analysis of Principal coordinates
-----

A program for generalised
discriminant analysis or
canonical correlation analysis
on the basis of any distance measure

by M.J. Anderson
Department of Statistics
University of Auckland (2003)

Input file name for Y matrix: PK.txt
No. of variables in Y = 47
No. of observations = 56
No. of groups = 3
Sample sizes per group = 15 21 20
Choice of m = 7
Data were transformed to ln(x+1)
No standardisation
Analysis based on Bray-Curtis dissimilarities

-----
Results:  PRINCIPAL COORDINATE ANALYSIS
-----

Percentage of variation explained by the first m axes
            individual    cumulative
Axis  1    20.718%      20.718%
Axis  2    12.370%      33.088%
```

Axis	3	10.803%	43.891%
Axis	4	8.123%	52.014%
Axis	5	7.566%	59.580%
Axis	6	6.650%	66.230%
Axis	7	5.149%	71.379%

Principal Coordinate Axes (unconstrained)
Axes

Sample	1	2
1	11.0405	5.8615
2	-12.4226	13.4440
3	14.3011	-7.2383
4	15.6437	-3.3236
5	-15.9353	-12.2421
6	13.4904	-16.2449
7	8.4434	-12.7538
8	-0.7456	-8.6552
9	8.8673	-27.8706
10	-10.8420	-4.8348
11	13.2139	-11.0566
12	-16.5884	-10.5761
13	-9.6554	-21.3002
14	-27.7365	-22.1786
15	7.1447	-4.1638
16	21.9808	12.8256
17	-0.6240	27.1937
18	15.8028	11.9198
19	15.1660	6.3021
20	-16.0009	3.9222
21	5.6123	-14.6809
22	-3.7562	-1.4499
23	12.9124	-4.1513
24	4.8083	-4.4977
25	5.0281	-9.3237
26	-20.6706	1.1072
27	9.6389	5.9739
28	-16.2207	0.3786
29	11.7666	2.4592
30	-34.8437	10.3376
31	-22.8211	15.4325
32	13.6935	12.8670
33	-22.3225	13.9081
34	19.4266	5.9280
35	-9.7893	-3.1453
36	8.1799	10.8847
37	12.6228	13.6111
38	-3.4121	4.5885
39	23.3624	-9.0480
40	8.5090	5.8891
41	6.2119	-19.3504
42	4.5253	-1.2623
43	-17.3171	1.0019
44	5.5746	3.6311
45	0.8865	-3.5720
46	-18.3375	10.5159
47	18.1686	18.0574
48	20.0227	-0.3844
49	14.8003	-3.3752
50	-17.1280	-5.3576
51	-9.0583	-14.9106
52	-12.9000	-0.5811
53	-27.5960	9.1026
54	7.1389	18.4649
55	-10.7199	2.9735
56	-0.5405	8.9470

Results: CANONICAL ANALYSIS

Eigenvalues (Correlations)
0.78101 0.69142

Squared Correlations (delta^2)

0.60998 0.47807

Canonical Axes (constrained)
Axes

Sample	1	2
1	-0.0093	0.0152
2	-0.0821	-0.0383
3	-0.1435	0.0296
4	0.0085	-0.0119
5	-0.1058	0.0082
6	-0.2029	0.0897
7	-0.2120	-0.0094
8	-0.0320	0.1058
9	-0.1303	0.2121
10	-0.1042	0.0462
11	-0.1250	0.0993
12	-0.1205	0.1361
13	-0.1839	0.0744
14	-0.1560	0.1736
15	-0.0465	-0.0106
16	0.1258	0.1548
17	0.1043	-0.0106
18	0.0866	-0.0608
19	0.1648	0.0581
20	0.0534	-0.0193
21	0.0049	-0.0012
22	0.2120	0.0531
23	0.0593	0.1688
24	-0.0138	0.0270
25	0.0292	0.0088
26	0.1192	0.0563
27	0.1016	-0.0244
28	0.0637	0.0389
29	0.1195	0.0741
30	0.2127	0.0570
31	0.0404	0.0633
32	0.0497	0.0032
33	0.0991	0.0001
34	0.0630	0.0107
35	0.1050	0.0533
36	0.1761	0.0631
37	0.1036	-0.0347
38	-0.1114	-0.0828
39	-0.0559	-0.1983
40	-0.0540	-0.2074
41	-0.2032	0.0409
42	-0.0457	-0.0245
43	-0.0570	-0.2420
44	0.0098	-0.0287
45	-0.0428	-0.1127
46	0.0082	0.0058
47	0.1193	-0.0378
48	0.0111	-0.1255
49	-0.0258	-0.0734
50	0.0202	-0.1064
51	0.0863	-0.0728
52	-0.0155	0.0194
53	-0.0297	-0.1376
54	-0.0160	-0.0396
55	0.0105	-0.1585
56	-0.0432	-0.0778

Correlations of Canonical Axes (Q*) with Original Variables (Y)
Axes

Var	1	2
1	0.3138	0.1391
2	0.1998	0.2430
3	0.3118	0.1140
4	0.2105	0.1220
5	-0.1435	-0.0779
6	0.0078	0.3033
7	-0.0165	-0.0276
8	-0.1477	0.1338
9	0.1226	-0.1656

10	0.0136	-0.2090
11	-0.1864	-0.0266
12	0.1099	-0.1331
13	0.1121	0.1241
14	0.1147	0.3073
15	0.0540	0.0888
16	0.4184	0.0796
17	0.4596	0.2383
18	0.3106	0.2265
19	0.3842	0.2452
20	0.2515	0.1743
21	0.1348	0.2432
22	-0.1296	0.0610
23	-0.3677	0.3113
24	-0.1892	-0.1290
25	-0.0554	-0.2402
26	0.2537	0.0658
27	0.0268	0.0739
28	0.3219	-0.1067
29	0.1052	-0.0409
30	0.0965	0.3064
31	-0.2959	-0.1358
32	0.1800	-0.0208
33	-0.2777	0.5117
34	0.7451	0.1841
35	-0.0323	-0.0749
36	-0.2252	0.2998
37	0.0234	0.0650
38	0.1913	0.1019
39	0.5656	-0.3508
40	-0.1299	0.3767
41	-0.1716	0.3172
42	0.1497	-0.1477
43	0.1939	-0.2596
44	0.3698	-0.0142
45	0.0012	-0.3048
46	0.1047	-0.1829
47	0.0494	0.0535

Results: DIAGNOSTICS

m	prop.G	ssres	d_1^2	d_2^2	%correct
1	0.20718	2.09795	0.00015	0.00000	5.357%
2	0.33088	1.84901	0.27131	0.00014	55.357%
3	0.43891	1.87486	0.33280	0.01279	57.143%
4	0.52014	1.90285	0.33289	0.08603	48.214%
5	0.59580	1.96072	0.35139	0.15689	46.429%
6	0.66230	1.48968	0.60976	0.22268	66.071%
7	0.71379	1.24578	0.60998	0.47807	73.214%
8	0.76010	1.14889	0.61529	0.50993	73.214%
9	0.80200	1.19283	0.61672	0.51686	71.429%
10	0.83996	1.21972	0.62080	0.53536	67.857%
11	0.87360	1.16798	0.68756	0.53851	69.643%
12	0.90610	1.17302	0.68781	0.54218	67.857%
13	0.93365	1.11178	0.73063	0.55217	67.857%
14	0.96002	1.14394	0.73149	0.55271	66.071%
15	0.98338	1.15323	0.73313	0.56195	67.857%

Note: m cannot include axes that cause the total SS to exceed the original total SS.

Results: CROSS-VALIDATION

Leave-one-out Allocation of Observations to Groups
(for the choice of m = 7)

Original Group		Classified into groups			Total	%correct
		1	2	3		
1	11	0	4	15	73.333%	

Group	2	1	16	4	21	76.190%
Group	3	2	4	14	20	70.000%

Total correct = 41/ 56 = 73.214%
 Mis-classification error = 26.786%

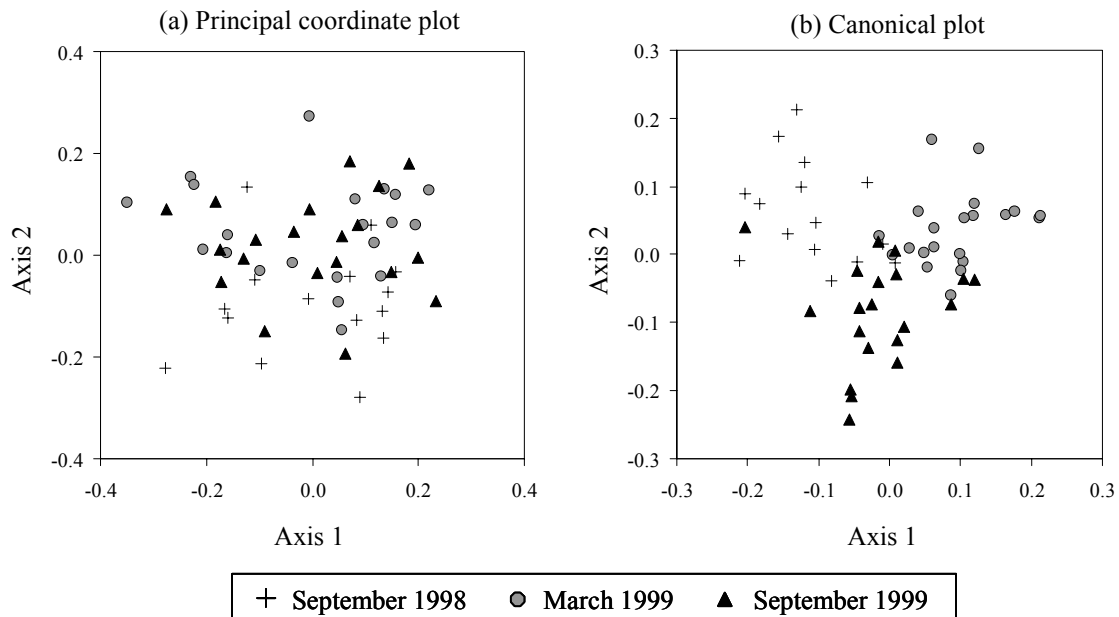
 Results: PERMUTATION TEST

trace statistic = (tr(Q_m'HQ_m))
 first squared canonical correlation = (delta_1^2)

 tr(Q_m'HQ_m) = 1.088049 P = 0.000100
 delta_1^2 = 0.609984 P = 0.000100

No. of permutations used = 9999
 Integer for the random seed = 12

These results indicate that there is a significant difference in the composition and relative abundance of fish species at the three different times of observation. Below are two ordination plots. The first is the (unconstrained) metric multi-dimensional scaling plot of the Bray-Curtis distances (produced using the first two principal coordinate axes from the output above). The second is the plot of the first two canonical axes produced by CAP (the canonical axes in the output above). The group differences are not apparent on the principal coordinate plot, while the canonical plot shows the differences quite clearly. This is because the axis of real group differences that occurred in multivariate space was not in the same direction as the maximum variation, which is why the differences do not appear in the unconstrained plot. Clearly, it is useful to have both the unconstrained and the constrained ordination to examine multivariate patterns in this data set.



V. Frequently Asked Questions

1. What happens if I type the name of the input file and the program shuts down without running?

Chances are, one of the following problems has occurred:

- (a) You have not typed in the correct name. It is useful to have an explicit extension on the file (such as *.txt), to have the computer set so that all filename extensions are shown (so that you can see them and you know exactly what the name of the file is).
- (b) You have not saved the file as an ASCII (*.txt) or a comma-separated file (*.csv). If you are using excel, it is important to explicitly choose one of these in the “File Type” category. Simply naming a file “something.txt” doesn’t automatically cause Excel to save it in this format!

2. What happens if I type the number of variables and the number of observations and the program shuts down without running?

Chances are, one of the following problems has occurred:

- (a) You have not typed the correct numbers. Double check this by examining the number of columns and the number of rows in the input data file before trying the program again.
- (b) You have given the program incorrect information as to whether variables are stored as columns versus stored in rows. Check this and try again.
- (c) You have used a program for saving the file that does not put a line-ending character at the end of the last line of the file. This problem can be remedied by opening up the file in any handy text editor (such as Notepad for PCs or BBEdit for Macintosh) and adding a return (“Enter”) at the very bottom of the file.
- (d) You have forgotten to save the file without any headers on the rows or columns. The program will not work if you have alphabetic characters in the file (such as names of variables, etc.). Make sure you save the file without any headers on the columns OR rows before proceeding.
- (e) You have a value in your data matrix which is not a number. Perhaps you have a cell in the matrix with nothing in it, or which contains one or more spaces or characters that are not numbers. Check this and try again.

3. What happens if the program stops when it starts doing diagnostic calculations? (i.e. the program appears to stop while showing “Choice of m = 1” on the screen).

This generally occurs because there are a large number of observations and the program has to work VERY HARD to do the leave-one-out allocation success (diagnostic calculations). The program does a separate canonical analysis after leaving out *each* observation one at a time for *each* choice of m separately. This is going to take some serious time if you have a hundred or more observations (depending on your processor’s speed). So, actually, there is nothing wrong with the program – it’s just going to take quite a while to do the diagnostics!! My suggestion if this is the case is to proceed as follows:

- (a) Run a principal coordinate analysis alone on the data (e.g. using the PCO.exe or PCO.mac routine on MJA’s website).
- (b) Look at the % of explained variation with increases in the number of PCO axes.
- (c) Find the number of axes that include about 80% of the variation in the original data.
- (d) Re-run the CAP program and choose m to be this value instead of having the computer program choose m .

Although this does not give you an optimal result (which could be obtained by letting the computer program choose m), you nevertheless will get a reasonable idea of what the canonical analysis is likely to show. Also, running the PCO alone using a separate program will tell you what the maximum value of m would be for the diagnostic analyses that would be done by the CAP routine. This will be the number of PCO axes that explain up to (but not exceeding) 100% of the original data. The value of m will also be no greater than the number of original variables.

Meanwhile, you can run the complete program (including the diagnostics) on your large data set overnight, for example, and see how far it gets. For example, an analysis of data with 569 observation rows took a Pentium 1.7 GHz computer about 36 hours to run. The take-home message is that it will (unfortunately) take a while to get complete diagnostics for large data sets – but don’t worry there is nothing wrong with the program!

VI. References

- Anderson, M.J. (2004). PERMANOVA_2factor: a FORTRAN computer program for permutational multivariate analysis of variance (for any two-factor ANOVA design) using permutation tests. Department of Statistics, University of Auckland, New Zealand.
- Anderson, M.J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology* 26: 32-46.
- Anderson, M.J. and Robinson. (2003). Generalised discriminant analysis based on distances. *Australian and New Zealand Journal of Statistics* 45(3): 301-318.
- Anderson, M.J. and Willis, T.J. (2003). Canonical analysis of principal coordinates: a useful method of constrained ordination for ecology. *Ecology* 84: 511-524.
- Gower, J.C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* 53: 325-338.
- Johnson, R.A. and Wichern, D.W. (1992). *Applied multivariate statistical analysis, 3rd edition*. Prentice-Hall, Englewood Cliffs, New Jersey.
- 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.
- Wilkinson, J.H. and Reinsch, C. (1971). Linear algebra. Pages 134-151 in *Handbook for automatic computation, Volume 2*. (F. L. Bauer, chief editor). Springer-Verlag, Berlin.
- Willis, T.J. & Denny, C.M. (2000). Effects of the Poor Knights Islands Marine Reserve on demersal fish populations. Report to the Department of Conservation, Science and Research Grant No. 2519, Leigh Marine Laboratory, University of Auckland, New Zealand.