

PCO

Principal Coordinate Analysis

A computer program
by Marti J. Anderson



Department of Statistics
University of Auckland
(2003)

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. For finding the eigenvalues of a real symmetric distance matrix, using a householder reduction, I used the canned FORTRAN routines: “tred2” and “tqli” from Numerical Recipes (Press et al. 1992).

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

Anderson, M.J. (2003). PCO: a FORTRAN computer program for principal coordinate analysis.
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. Description

PCO is a computer program that calculates a principal coordinate analysis of any symmetric distance matrix in the manner of Gower (1966). Namely, let $\mathbf{D} = (d_{ij})$ be an $(n \times n)$ distance matrix. Let

$\mathbf{A} = (a_{ij}) = (-\frac{1}{2}d_{ij}^2)$, then calculate Gower's 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 the identity matrix. Matrix \mathbf{G} is then decomposed into its component eigenvalues and eigenvectors. The eigenvectors, standardized by dividing by the square root of their corresponding eigenvalue, are output as the principal coordinate axes. This analysis is also called metric multi-dimensional scaling. It is useful for ordination of multivariate data on the basis of any distance function.

The user should be aware that the program does not correct for negative eigenvalues in any way. These are scaled by dividing by the square root of the *absolute value* of the corresponding eigenvalue and are simply given alongside the positive axes in the output.

II. Input file

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

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

PCO is written using dynamic allocation of memory and thus (theoretically) there are no limits to the size of the matrix (rows or columns) that is entered for analysis. Please contact the author if problems (other than the time required) are encountered in the analysis of large matrices. **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.

For the output, the user chooses how many axes he/she wishes to obtain. Usually, only 2 or 3 are of interest for ordination. The program outputs the values for each individual observation on each of the principal coordinate axes which can be imported into any graphics package (or Excel) for plotting.

III. 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).

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.

Program PCO

A program for calculating Principal Coordinates
(Metric Multi-dimensional Scaling)
by M.J. Anderson

Department of Statistics
University of Auckland (2003)

Type the name of the input file containing your data

Tim.txt

Nature of the data in the input file:

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

1

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.

Indicate the nature of the output file you wish:

- 1) Output=principal coordinates
- 2) Output=distance matrix D
- 3) Output=Gower's centred matrix G
- 4) Output=distance matrix based on centroids
calculated from principal coordinates
- 5) Output=centroids of raw data

1

Here, there are clearly several options. The most common of these is to wish to produce principal coordinates. However, there are situations where you may wish to produce simply the distance matrix or Gower's centred matrix. In other situations, you may wish to produce a distance matrix for further analysis which consists of the centroids from principal coordinates. This can be useful if you need to analyse centroids in, for example, Bray-Curtis space (or other non-Euclidean space), in which case the centroids are not the arithmetic averages of the original variables. You may also simply wish to output centroids of the raw data. Note that if you choose options 4 or 5, the program asks you the sample size from which centroids are to be calculated – there is unfortunately no catering here for unequal numbers of observations per cell.

Structure of the input file:

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

I

Choice of transformation:

- 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

I

Choice of standardization:

- 1) none
- 2) standardise by row (sample) sums
- 3) standardise by column (variable) sums
- 4) double standardize 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

I

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

12

Note: Although the Bray-Curtis distance measure is used by many ecologists for species abundance data, the Chi-squared distance measure, Orloci's chord distance, or other measure may also be appropriate in different situations. For example, the chi-square distance measure has been used extensively and is the basis of correspondence analysis. It will tend to emphasize compositional changes rather more than changes in abundance, compared to the Bray-Curtis measure. The CY distance measure is also useful; it is somewhat akin to using the Bray-Curtis measure on $\log(x+1)$ transformed data, but no log transformation is necessary. The Gower distance is somewhat akin to using the Canberra measure, but where each variable is standardized by its range. See Legendre and Legendre (1998) for an extensive review of various distance measures. Recall also that if you require a different measure, you may input a distance matrix directly. Note that in this case it needs to be a full ($N \times N$) matrix.

You have chosen the CY dissimilarity measure.
This measure modifies zero values by
adding a constant before taking logs.
Type the value you want for this constant
(e.g. 0.1, 0.9, etc.).
Note: The published value for this is 0.1.

0.1

How many observations or samples (rows) are there?

24

How many variables (columns) are there?

46

How many coordinates 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.

2

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

q

The program has completed the analysis and sent results to the named output file. In this case, the output file looks like this:

```
Program PCO
A program for calculating Principal Coordinates
(Metric Multi-dimensional Scaling)
```

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

```
Input file:Tim.txt
```

```
No transformation
No standardization
Analysis based on CY dissimilarities
Value for zero replacement was 0.100
```

```
--- Results ---
```

```
Percentage of variation explained by individual axes
```

		individual	cumulative
Axis	1	44.534%	44.534%
Axis	2	22.547%	67.081%

```
Principal Coordinates
```

Sample	Axes	
	1	2
1	0.272	-0.086
2	0.261	-0.052
3	0.303	-0.003
4	0.281	0.051
5	0.136	0.013
6	0.223	0.118
7	0.114	0.099
8	0.172	0.021
9	0.139	0.065
10	0.164	0.000
11	0.105	0.165
12	-0.060	0.371
13	-0.087	-0.133
14	-0.062	-0.185
15	-0.051	-0.312
16	-0.016	-0.329
17	-0.323	0.076
18	-0.259	0.071
19	-0.283	-0.035
20	-0.199	-0.016
21	-0.304	0.123
22	-0.219	0.043
23	-0.129	-0.055
24	-0.178	-0.010

The resulting two-dimensional principal coordinate plot obtained using the results of the program above is shown below. Each observation is labeled as belonging in one of three shading treatments and occurring either near to or far from the sea floor.

